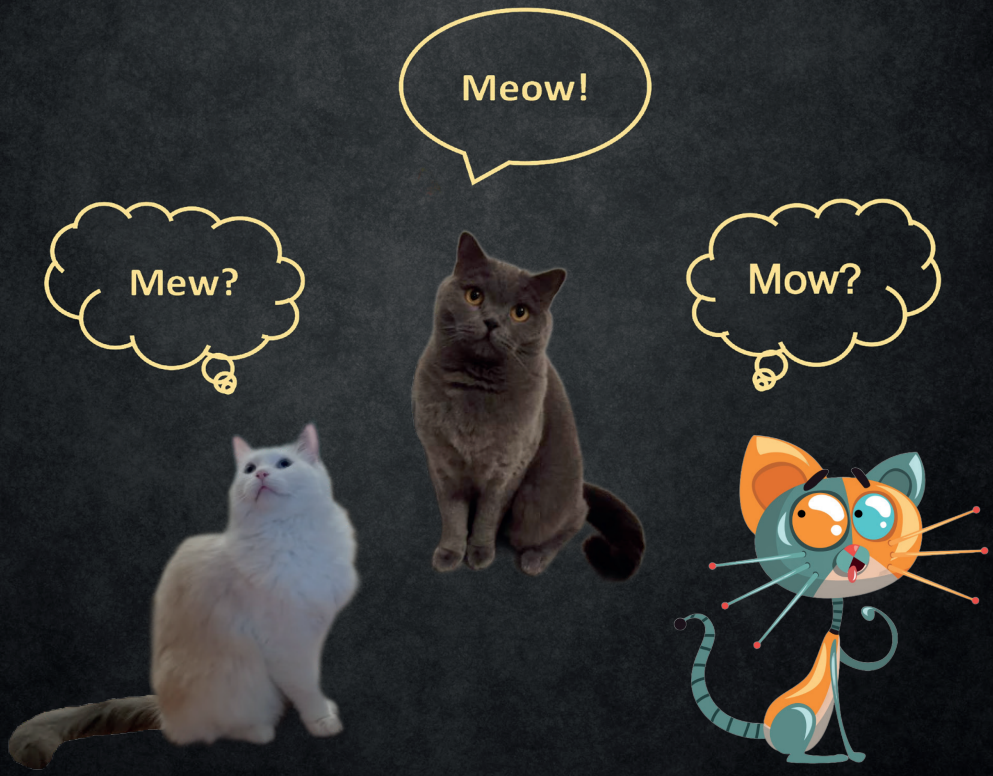
Measuring the intelligibility of pathological speech

through subjective and objective procedures

Wei Xue
薛 玮



MEASURING THE INTELLIGIBILITY OF PATHOLOGICAL SPEECH

through subjective and objective procedures

Wei Xue

薛玮

Funding body

This research was funded by EU's H2020 research and innovation programme under the MSCA GA 766287 (TAPAS Training Network on Automatic Processing of PAthological Speech, early stage researcher fellowship for Wei Xue) (www.tapas-etn-eu.org).

Radboud Universiteit

TAPAS

ISBN	978-94-6458-936-8
Cover	Wei Xue & Zhengyu Zhao
Lay-out	Publiss | www.publiss.nl
Print	Ridderprint | www.ridderprint.nl

Measuring the intelligibility of pathological speech through subjective and objective procedures

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

dinsdag 21 maart 2023
om 12.30 uur precies

door

Wei Xue
geboren op 19 maart 1992
te Shijiazhuang (China)

Promotoren:
Dr. W.A.J. Strik
Prof. dr. R.W.N.M. van Hout

Copromotor:
Dr. C. Cucchiarini

Manuscriptcommissie:
Prof. dr. F. Moscoso del Prado Martin
Prof. dr. H. Christensen (The University of Sheffield, Verenigd Koninkrijk)
Prof. dr. A. van Wieringen (KU Leuven, België)
Dr. K.P. Truong (Universiteit Twente)
Dr. E. Janse

Measuring the intelligibility of pathological speech through subjective and objective procedures

Dissertation to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.H.J.M. van Krieken,
according to the decision of the Doctorate Board
to be defended in public on

Tuesday, March 21, 2023
at 12.30 pm

by

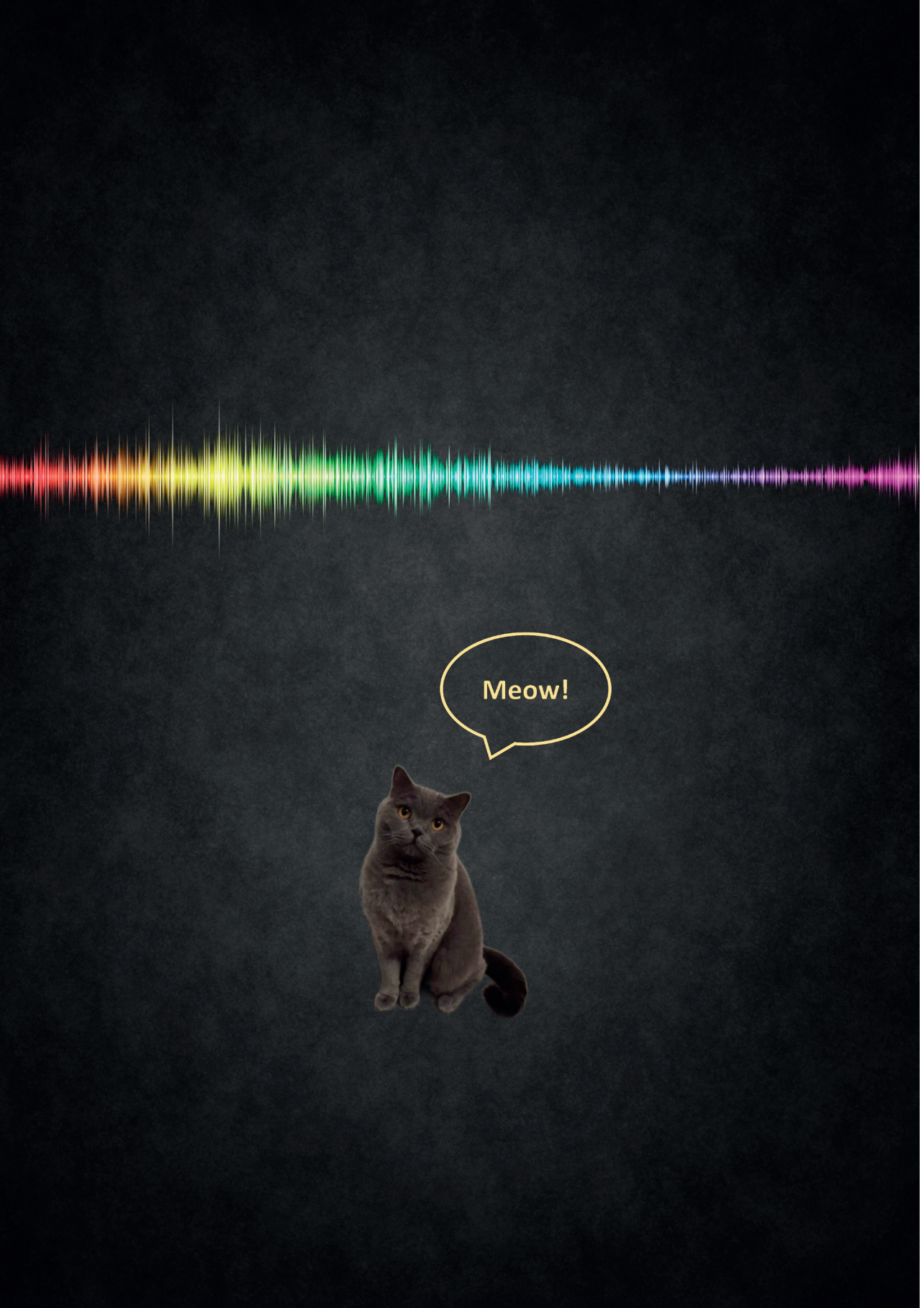
Wei Xue
born on March 19, 1992
in Shijiazhuang (China)

Supervisors:
Dr. W.A.J. Strik
Prof. dr. R.W.N.M. van Hout

Co-supervisor:
Dr. C. Cucchiarini

Manuscript Committee:
Prof. dr. F. Moscoso del Prado Martin
Prof. dr. H. Christensen (The University of Sheffield, United Kingdom)
Prof. dr. A. van Wieringen (KU Leuven, Belgium)
Dr. K.P. Truong (University of Twente)
Dr. E. Janse

Contents


Meow!

GENERAL INTRODUCTION

Living organisms on earth have the ability to communicate by using signs and signals. We, as human beings, can effectively and freely communicate in all kinds of daily communication scenarios by using a special tool – language, which can be conveyed by speech, writing, and sign. Having impairments in speech can affect the messages conveyed by the speech signal, affecting intelligibility and, consequently, communication.

People with dysarthria suffer from speech impairments due to neurological diseases (e.g., parkinsonism and amyotrophic lateral sclerosis) or injuries (e.g., traumatic brain injury and thrombotic/embolic stroke). Dysarthria comprises a set of neurological disorders that cause loss of control over the muscles used for speech, resulting in abnormalities in aspects of speech production, such as respiration, phonation, nasalization, articulation, and prosody (Duffy, 2013, p. 3). Such abnormalities can lead to, but are not limited to, disorders in strength, speed, range, steadiness, and tone (Duffy, 2013, p. 3), as well as decreases in speech intelligibility. As a consequence, people with dysarthria may lose contact with others and eventually become isolated from social life and society. These consequences severely affect their quality of life. To alleviate such speech disorders and their social repercussions, and to slow down the decrease in speech intelligibility, speech therapy has been shown to be useful. For measuring the effectiveness of therapeutical treatments and monitoring developments, such as through pre- and post-therapy evaluations, it is necessary to have a clear definition and a robust operationalization of speech intelligibility.

The concept of *intelligibility* is not only important in speech pathology, but also in other disciplines, such as second language learning (L2), telecommunication, and audiology. In each of these disciplines, the concept of *intelligibility* has been approached from a different perspective in connection with the classic *communication chain* which consists of a *speaker* to generate speech signal, a *transmission channel* to transfer the speech signal, and a *listener* to perceive the speech signal (Nicoras et al., 2022). For instance, in telecommunication, researchers define and measure intelligibility in relation to the losslessness of the transmission channel, while in audiology, researchers study the impact of listeners' hearing impairments on intelligibility assessments. Speech pathology and L2 research both study the effects of speaker-related properties on speech

intelligibility; however, in both disciplines, the reduction in intelligibility is caused by different deviations, such as having dysarthria and having L2 accents. This dissertation aims to investigate speech intelligibility with respect to speaker-related properties and, more specifically, to focus on the effect of deviations caused by dysarthria in speech pathology.

The concept of *intelligibility* is different from the concept of *comprehensibility*, which has also been frequently used in the clinical context and in research on speech pathology (Barefoot et al., 1993; Yorkston et al., 1996b; Hustad, 2008; Pommée et al., 2022). Barefoot et al. (1993) argued that "comprehensibility pertains to the domains of both speech and language, whereas intelligibility pertains principally to the domain of speech" (p. 32). They emphasized that "comprehensibility is intended to account for communication features of utterances that extend beyond the auditory–acoustic domain." (Barefoot et al., 1993). Similarly, Yorkston et al. (1996b) defined the two concepts as follows: "The term intelligibility refers to the degree to which the acoustic signal (the utterance produced by the dysarthric speaker) is understood by a listener. … The concepts of comprehensibility and intelligibility may be distinguished by the fact that comprehensibility incorporates signal-independent information such as syntax, semantics, and physical context" (p. 55). Hustad (2008) summarized and simplified the concept of *intelligibility* as "Intelligibility refers to how well a speaker's acoustic signal can be accurately recovered by a listener" (p. 562). According to the above definitions, the essential difference between the two concepts seems to be that comprehensibility is measured by taking contextual and signal-independent information into account (Yorkston et al., 1996b), while speech intelligibility is measured independently of context. A recent survey about the consensus on these two concepts (Pommée et al., 2022) further discussed this view. The authors recruited forty international professionals from different fields (such as clinicians, linguists, and computer scientists) to elaborate on their understanding of the two concepts and the approaches to assess them. They confirmed that the two concepts are related to each other and can contribute to functional human communication. However, the two concepts refer to different reconstruction levels of speech material. Intelligibility refers to the acoustic-phonetic decoding of speech while comprehensibility refers to the reconstruction of the message conveyed in the speech.

According to the abovementioned definitions, it is clear that *intelligibility* should be assessed with the participation of human listeners, and this procedure is therefore considered to be subjective. *Subjective procedures* for measuring intelligibility are commonly implemented through conducting listening experiments. Listeners participating in the experiments are asked to provide scalar judgments (Barreto & Ortiz, 2008; Miller, 2013; Yorkston & Beukelman, 1978; Finizia et al., 1998; Schiavetti, 1992), to make orthographic transcriptions (Hustad, 2006; Laures & Weismer, 1999), or to select the sentence they heard from a set of multiple sentences (Yorkston et al., 1996a; De Bodt et al., 2006). Measures of speech intelligibility are then obtained by averaging scores over multiple listeners, combined with an examination of their reliability. Many studies have shown that these procedures can produce reliable measures (e.g., Hirsch et al., 2022; Kent & Kim, 2011; Van Nuffelen et al., 2010; Ganzeboom et al., 2016) and that they have been widely used in research and clinical practice.

However, these studies are limited. In general, how different factors in subjective procedures influence measures of speech intelligibility has not been extensively analyzed. In particular, the comparison involving orthographic transcription between speech materials has been limited by the use of a typical form of transcription that allows only existing words. Furthermore, commonly used statistical analyses for reliability examination cannot handle all relevant factors in a procedure and different experimental designs. Also, the validity of speech intelligibility measures, a key question in research, has rarely been examined in the field of dysarthric speech.

In addition to subjective procedures, many studies have explored the possibility of using objective procedures to measure speech intelligibility, which do not require involving human listeners. This is because researchers have argued that while subjective procedures may be feasible in a research setting, they remain problematic in clinical practice where clinicians strongly prefer easy-to-use applications that provide interpretable results. Objective procedures, either studying acoustic features of dysarthric speech or employing sophisticated machine learning (ML) algorithms such as Automatic Speech Recognition (ASR) systems or End-to-End (E2E) deep learning models, have been shown to produce outputs that are highly correlated with intelligibility measures obtained through subjective

procedures (Schuster et al., 2006a; Middag et al., 2009c; Berisha et al., 2013; Pellegrini et al., 2015; Kim et al., 2015).

However, these studies have several limitations. Specifically, for acoustic feature-based procedures, it is still far from clear how acoustic features that are associated with deviations in dysarthric speech correlate to intelligibility measures. This limits the potential of using these features to develop easy-to-use tools for clinical practice. For ML-based procedures, the models' outputs are difficult for speech-language pathologists to interpret, not to mention being used for diagnosis. Moreover, dysarthric speech, in contrast to healthy speech, poses a *low-resource* problem since the resources of dysarthric speech are inadequate for training reliable models.

This dissertation aims to gain insights to establish valid procedures for measuring the intelligibility of pathological speech. To achieve this aim, both subjective and objective procedures are examined. As an introduction, Section 1.1 examines the factors in subjective procedures that have an impact on measures of intelligibility as well as the reliability and validity of human evaluations. Section 1.2 focuses on objective procedures, starting with discussing acoustic correlates of intelligibility, followed by discussing advances in sophisticated ML-based models and related issues such as the low-resource problem. Section 1.3 presents suggestions for further research and the outline of this dissertation.

1.1 Subjective procedures for measuring speech intelligibility

The common implementation of subjective procedures for measuring speech intelligibility is to conduct listening experiments in which groups of listeners with different *listener characteristics* are asked to assess the intelligibility of speakers with speech impairments. The assessment of intelligibility can be performed with different *measurement methods*, i.e., scalar judgments and item identifications, and with different *speech materials* presented in audio-only or audio-visual *presentation modes* (Hustad & Cahill, 2003). The speech under assessment can be in different *types of speech tasks*: reading aloud, repetition, and spontaneous speech (Barreto & Ortiz, 2008; Kempler & Van Lancker, 2002). Moreover, speech intelligibility can be evaluated at

different *granularity levels* with respect to the units to be studied, such as graphemes (letters), phonemes, syllables, words, and sentences.

All the above six factors involved in subjective procedures have been found to have an impact on measures of intelligibility (e.g., Borrie et al., 2012; Ganzeboom et al., 2016; Hustad, 2007; Hustad & Cahill, 2003; Kempler & Van Lancker, 2002; Yorkston & Beukelman, 1978). The following subsections first discuss the four affecting factors: speech materials and granularity levels (Section 1.1.1), measurement methods (Section 1.1.2), and listener characteristics (Section 1.1.3). Note that for *types of speech tasks*, we only focus on reading aloud speech because this has been found to result in similar intelligibility scores to repetition. Spontaneous speech, on the other hand, is more time-consuming and labor-intensive to analyze (Kempler & Van Lancker, 2002). For *presentation mode*, we choose the audio-only mode over the audio-visual mode following common practice. Then, reliability and validity in subjective procedures are also discussed (Section 1.1.4) since they are necessary for subjective procedures, which involve human listeners.

1.1.1 Speech materials and granularity levels

Commonly studied and used speech materials vary in degrees of morphosyntactic complexity and semantic predictability (Barreto & Ortiz, 2008). Some studies used lists of words (Kim et al., 2011a; Balzan et al., 2019) or pseudowords presented in isolation, such as the Dutch Intelligibility Assessment (DIA; De Bodt et al., 2006), or paired words to evaluate minimal contrasts (Kent et al., 1989; Levy et al., 2016). Some studies examined isolated sentences with semantic predictability (Abur et al., 2019; Barreto & Ortiz, 2008, 2016; Carvalho et al., 2021; Hodge & Gotzke, 2014; Hustad & Cahill, 2003; Hustad, 2006, 2007, 2008; Ishikawa et al., 2020; Liss et al., 2002; Middag, 2012; Miller, 2013; Stipancic et al., 2016; Tjaden & Liss, 1995a, 1995b; Tjaden et al., 2014a; Tjaden & Wilding, 2010; Sussman & Tjaden, 2012; Yorkston & Beukelman, 1978, 1981; Yorkston et al., 1996a) or without (Benoît et al., 1996; Beijer et al., 2012b; Ganzeboom et al., 2016). Other studies investigated intelligibility on phonetically-balanced texts (narratives or paragraphs), such as 'Papa en Marloes' ('Papa and Marloes' in English; Van Lierde et al., 1991; Middag, 2012) in the field of Dutch pathology, 'The North Wind and the Sun' (Maier et al.,

2007) and 'The Grandfather' passages (Laures-Gore et al., 2016; Rudzicz et al., 2012) in the field of English pathology. Although some researchers claimed that unpredictable speech material (e.g., syllables, pseudowords, minimal word pairs, and unpredictable sentences) should be used to assess speech intelligibility (Pommée et al., 2022) to avoid listeners 'guessing' at the content (Beijer et al., 2012b; Benoît et al., 1996), many others did not restrict themselves to such materials. They argued that speech materials with predictable information, like meaningful sentences, are closer to daily speech communication and may provide insights into perceptual properties that can influence the intelligibility of connected speech, such as stress, rhythm, and intonation (Miller, 2013). Thus, it remains to be seen how and to what extent contextual information affects the measure of intelligibility.

Speech intelligibility can be measured at different granularity levels, such as at the subword, word, and utterance levels. In this dissertation, measures at the subword level refer to those considering subword segments as basic units, namely phonemes, letters (graphemes), and syllables. Similarly, measures at the word level refer to those considering words as basic units, where words may occur either in isolation as in word lists or in sentences. Measures at the utterance level refer to those considering utterances as basic units, e.g., word lists, sentences, and narratives, which consist of multiple words. It is important to note that speech materials are categorized independently of the granularity levels in this dissertation, whereas in many studies, measures of intelligibility are calculated in conjunction with both factors (e.g., Hustad, 2007; Stipancic et al., 2016; Sussman & Tjaden, 2012; Yorkston & Beukelman, 1980, 1981). These studies used different names for the measures of the same unit in different speech materials. For example, a percentage of correctly transcribed words, considered as a word-level measure in this dissertation, was named as "word intelligibility" or "single-word intelligibility" (Stipancic et al., 2016; Sussman & Tjaden, 2012; Weismer, 2009; Yorkston & Beukelman, 1980, 1981) if words were uttered in isolation, whereas it was named as "sentence intelligibility" if words were presented and assessed in isolated sentences (Abur et al., 2019; Stipancic et al., 2016; Sussman & Tjaden, 2012).

Both factors, i.e., the speech material and the granularity level, have been found to affect the measure of intelligibility (Ganzeboom et al., 2016;

Hustad, 2007; Yorkston & Beukelman, 1978). Many studies have shown that intelligibility scores generally increase when the semantic predictability in speech materials increases. Furthermore, such an effect appeared to be related to the severity levels of dysarthria (speech disorders). The increased intelligibility due to richer semantic cues appears to be particularly true for speech with mild and moderate disorders (Yorkston & Beukelman, 1978). When speech disorders are more severe, the increase in intelligibility seems to be less evident or even absent. This effect is explained in detail below by using a word-level measure, i.e., the percentage of correctly transcribed words, as an example.

For the distinction between sentences and words, higher intelligibility scores have been found for sentences compared to those for isolated words when speech is mildly and moderately dysarthric (Hustad, 2007; Yorkston & Beukelman, 1978, 1981). One explanation is that listeners could rely on more contextual information to understand the message in sentences than in words. However, when speech is more severely disordered, divergent findings have been reported in the comparison of sentences and word lists, with intelligibility scores in sentences being higher than, equal to (Barreto & Ortiz, 2008; Dongilli, 1994; Middag et al., 2009a; Yorkston & Beukelman, 1978), or lower (Yorkston et al., 1981) than those in word lists. One explanation for these divergent findings in speakers with more severe dysarthria is that the speakers have too many difficulties in producing sentences so that listeners no longer benefit from the contextual cues occurring in sentences.

For the distinction between sentences and narratives, Hustad and Beukelman (2001) found an increase in intelligibility for mild, moderate and severe dysarthric speech, but not for the profound case (intelligibility scores < 20%). One explanation for this relates to the extent of degradation of the speech (Hustad, 2007). It could be that the sentences composing the narrative are related, invoking listeners to progressively build contextual knowledge for each sentence. In turn, the building of context would promote the listeners' ability to infer the words that they may not recognize directly from the speech by applying their top-down inherent linguistic knowledge (Hustad, 2007). However, when speech is too severely disordered, listeners are no longer able to benefit from this mechanism.

For the distinction between words and narratives, the increase in intelligibility holds for speakers with mild (Hustad, 2007; Beukelman & Yorkston, 1979), moderate, and severe dysarthria (Hustad, 2007), but again not for severe-to-profound dysarthric speech.

These findings indicate that it is necessary to study speech materials with varying degrees of semantic predictability and to investigate their effects on intelligibility measures at different granularity levels, which have been little studied so far. This may help to develop valid subjective procedures for measuring intelligibility. Moreover, other characteristics of speech materials, such as the length of speech sample, may be also of interest to study.

1.1.2 Measurement methods

The measurement methods to assess intelligibility can be divided into two categories: scalar judgments (Barreto & Ortiz, 2008; Schiavetti, 1992) and item identifications (Barreto & Ortiz, 2008; Kent et al., 1989). Scalar judgments include ratings on various scales, such as Visual Analogue Scales (VAS; Hirsch et al., 2022; Ganzeboom et al., 2016; Kent & Kim, 2011; Van Nuffelen et al., 2010) and Likert scales (Ganzeboom et al., 2016; Hashemi Hosseinabad et al., 2021; Kent & Kim, 2011; Yorkston & Beukelman, 1978), and estimates of intelligibility, such as direct magnitude estimates (Ellis & Fucci, 1991; Ellis et al., 1996; Weismer & Laures, 2002) and percentage estimates (Yorkston & Beukelman, 1978; Hustad, 2006). Item identifications, which are also known as *response types* (Beijer et al., 2012b; Barreto & Ortiz, 2008; Yorkston & Beukelman, 1978), refer to those identifying items or units, such as phonemes, syllables, words, and sentences, through providing orthographic transcriptions of speech sounds (Hashemi Hosseinabad et al., 2021; Hustad, 2006, 2007; Weismer, 2009; Yorkston & Beukelman, 1978), completing sentences (Yorkston & Beukelman, 1978), or choosing answers from multiple choices of sentences or words (Yorkston & Beukelman, 1978). Table 1.1 shows several examples demonstrating how intelligibility measures can be derived for different speech materials when using the typical form of orthographic transcription that allows only existing words. Note that procedures using orthographic transcriptions were considered to be ‘objective’ in some studies because intelligibility scores were ’objectively’

derived from the transcription (Miller, 2013; Yorkston & Beukelman, 1978). However, in this dissertation, they are considered to be subjective because they still require the participation of human listeners, in contrast to objective procedures that rely on ML-based models or acoustics features, as later discussed in Section 1.2. Furthermore, existing methods of item identification for assessing speech intelligibility show limitations in the possible granularity levels of intelligibility measures and in the comparison of intelligibility measures between speech materials. For example, Tables 1.1 and 1.2 clearly show that it is not possible to compare intelligibility measures at the word level between meaningful sentences and word lists with pseudowords when using orthographic transcriptions allowing only existing words.

Table 1.1. Examples demonstrating how intelligibility measures can be derived for different speech materials when using the typical form of orthographic transcription that allows only existing words.

	Meaningful sentences	**Semantically unpredictable sentences**	**Word list without pseudowords**	**Word list with pseudowords**
Prompt	*The boat is in the bay.*	*The table walked through the blue truth.*	*bad bat*	*veed veat*
Orthographic transcription	*The boat is in the bag.*	*The table walks XXX the blue tooth.*	*bad pat*	*feed fat*
Number of total words	6	7	2	2
Number of correct words	5	4	2	-
Number of total phonemes	15	24	6	6
Number of correct phonemes	13	19	5	3
Percentage of correct words	83.3	57.1	100	-
Percentage of correct phonemes	86.7	79.2	83.3	50.0

Measures of intelligibility have been found to differ not only between the two categories of methods but also between methods in the same

category. For instance, when comparing the two categories of measurement methods, many researchers have reported higher intelligibility scores derived from orthographic transcriptions than those from scalar judgments, such as 7-point Likert scales (Yorkston & Beukelman, 1978; Ganzeboom et al., 2016), VAS (Ganzeboom et al., 2016; Stipancic et al., 2016), and percentage estimations (Hustad, 2006). What makes the difference seems to be that in transcriptions, listeners focus only on segmental-level (articulation) information, whereas in rating scales, listeners' perception may be further influenced by the suprasegmental disorders, such as those in speed, stress, and intonation, thus leading to an underestimation of intelligibility. When comparing different methods of item identifications, Yorkston and Beukelman (1978) observed an increase in mean scores on intelligibility measures from orthographic transcription to completion (partial transcription), and to multiple-choice tasks, with the correlations varying from insignificant moderate to significant strong. When comparing different methods of scalar judgments, the relation between Likert scales and VAS or percentage estimate was found to be inconsistent (Ganzeboom et al., 2016; Yorkston & Beukelman, 1978).

Table 1.2. Possible granularity levels of intelligibility measures in relation to speech materials when using existing methods of item identification for assessing speech intelligibility.

	Sentences		**Word lists**	
	meaningful	**semantically unpredictable**	**without pseudowords**	**with pseudowords**
Possible granularity level of intelligibility measures	word, syllable, grapheme, phoneme			syllable, grapheme, phoneme

Beyond the effects of measurement methods, the two categories of methods have their own strengths and weaknesses. For instance, scalar judgments require relatively little time and so can be easily applied to longer, connected speech, such as read speech of narratives or spontaneous speech (Stipancic et al., 2016). On the other hand, they only indicate the overall intelligibility, and thus, they cannot indicate which specific speech deviations are the cause of decreases in intelligibility (Kent et al.,

1989). In contrast, orthographic transcriptions can present more detailed, segmental-level information that is related to speech degradation, but high requirements of human efforts in deriving intelligibility measures should be further addressed, for example, based on existing programs. Additionally, orthographic transcriptions have been considered to yield acceptable intrarater and interrater reliability (Miller, 2013), whereas the reliability of scalar judgements has been found to vary across methods. The reliability of Likert scales has been questioned (Miller, 2013; Schiavetti, 1992), while the VAS is considered to be as reliable as orthographic transcriptions (Tjaden et al., 2014a).

However, the abovementioned studies are limited to one specific speech material or granularity level, which makes it difficult to compare different procedures with different implementations of the factors. Thus, it is necessary to study how these two factors interact with measurement methods. This may help to provide not only important scientific insights but also practical implications for comparing different intelligibility measures. Also, using the typical form of orthographic transcription has limited previous research in comparing different subjective procedures because it only allows existing words. For instance, sentences consisting of existing words cannot be compared with a word list containing pseudowords, which can only be transcribed at the phoneme level but not at the word level. Thus, it is necessary to explore a new form of transcription that provides more freedom.

1.1.3 Listener characteristics

It is well known that listener characteristics have an impact on measures of speech intelligibility. The commonly studied factors related to listener characteristics are the *listener experience* with disordered speech (i.e., the *listener type*), the *listener familiarity* with speakers, and the *listener gender*.

Listener experience refers to listeners' knowledge of disordered speech, based on which listeners are mainly categorized into inexperienced (i.e., 'naïve' or 'lay') listeners, like college students, and experienced (i.e., 'expert') listeners, like speech-language pathologists. *Listener experience* can influence the assessment of intelligibility in different ways. For example,

Carvalho et al. (2021) reported a direct effect that significant differences in ratings of speech intelligibility for speakers with Parkinson's disease were found between naïve listeners and healthcare professionals (expert listeners). Some researchers have reported indirect effects that *listener experience* shows an effect on the level of listener effort, i.e., how much effort listeners need to hear the speech clearly (Maruthy & Raj, 2014; Smith et al., 2019). In addition, Mencke et al. (1983) found that measures collected from naïve listeners tend to show larger variations than those collected from well-trained expert listeners such as speech-language therapists. This is further supported by the fact that an acceptable reliability level of intelligibility measures can be achieved by fewer expert listeners compared to naïve listeners (Ganzeboom et al., 2016). However, the relation between *listener experience* and the other three factors still needs to be further explored.

Listener familiarity refers to the extent of listeners' prior exposure (familiarization) to a particular speaker (voice) or speech, and it is mainly categorized into *no familiarization*, *passive familiarization,* and *active (explicit) familiarization. Passive familiarization* is approached by exposure to audio signal only (Borrie et al., 2012; Kim & Nanney, 2014), and *active familiarization* is approached by exposure to both audio signal and written transcripts (Borrie et al., 2012; Kim & Nanney, 2014; Tjaden & Liss, 1995a, 1995b). In terms of the impact of *listener familiarity*, previous studies have shown that intelligibility scores are higher for familiarization compared to no familiarization (Borrie et al., 2012; DePaul & Kent, 2000; D'Innocenzo et al., 2006; Hustad & Cahill, 2003; Kim & Nanney, 2014; Liss et al., 2002; Spitzer et al., 2000; Tjaden & Liss, 1995a, 1995b), and are higher for active familiarization compared to passive familiarization (Borrie et al., 2012; Kim & Nanney, 2014), with active familiarization appearing to have a long-term effect (Kim & Nanney, 2014). Some researchers reasoned the impact of *listener familiarity* as enhanced segmental-level perception due to familiarization (Borrie et al., 2012; Liss et al., 2002; Spitzer et al., 2000) given that correctness patterns at segmental level (e.g., percent syllable resemblance) were significantly higher for familiarization compared to no familiarization (Borrie et al., 2012). In contrast, no difference in suprasegmental-level perception, as evaluated by lexical boundary errors,

was found between the listener groups with and without familiarization (Liss et al., 2002). The fact that the listener familiarity has an impact on intelligibility measures means that this factor needs to be controlled.

Listener gender has also been investigated. No significant differences were reported between male and female listeners tested either by direct magnitude estimation (Ellis et al., 1996) or transcriptions (Yoho et al., 2019). In contrast, an effect of listener gender was found on rating scales for crowd-sourced data. An interesting finding by Ellis et al. (1996) was that the listeners with a gender different from the speakers tended to better understand the speech than those with the same gender. Nonetheless, other measurement methods may exert other influences. Research on intelligibility assessment should consider and report gender information of speakers and listeners to see if similar findings are obtained.

1.1.4 Reliability and validity

In subjective procedures, human listeners decide themselves what responses to give when experiencing distortions in the speech signal. As different listeners may have different responses, it is necessary to study the reliability and validity of the subjective measures. The reliability of the subjective measure is generally evaluated in terms of intrarater and interrater reliability. Examining interrater reliability is essential since it can generalize to a measure representing a group of listeners, while intrarater reliability cannot and has the disadvantage of requiring repeated assessments, which may influence the result due to listener familiarity.

Interrater reliability has been evaluated by different statistical analyses. Some studies computed the percentage of inter-listener agreement, a measure that does not take chance agreement into account (e.g., Hustad, 2007; Van Nuffelen et al., 2008). Some calculated Pearson correlations between listeners, which take only *Listener* as a source of variance. To address this issue, some researchers used the Intraclass Correlation Coefficient (ICC; Fisher, 1992; Rietveld, 2020), which can handle two factors, i.e., *Speaker* and *Listener*, in a crossed experimental design. However, ICC cannot handle more than two factors or a nested experimental design, where different groups of listeners evaluate different groups of speakers. Thus, it is necessary to explore statistical analyses that can handle all relevant

factors in a procedure and different experimental designs. In fact, ICC for reliability has been expanded into an overarching type of analysis, called Generalizability Theory (G Theory; Brennan, 2001). Although G Theory is based on the ICC, it can take into account more than two sources of variance. G Theory can handle not only crossed designs but also nested designs. In addition, G Theory allows calculating the optimal number of listener and utterance samples required to obtain reliable measures by conducting a decision study.

In contrast to the reliability, the validity of speech intelligibility measures has rarely been examined (Ellis & Fucci, 1991; Hustad, 2007; Stipancic et al., 2016; Van Nuffelen et al., 2008) in the field of dysarthric speech. Validity indicates the extent to which the scores measure what they intend to measure, and it is a key question in research. Therefore, it is important to investigate the validity in subjective procedures for measuring the intelligibility of pathological speech (Barreto & Ortiz, 2016).

1.2 Objective procedures for measuring speech intelligibility

Researchers have explored different objective procedures to assess dysarthric speech. One procedure focuses on studying acoustic features of dysarthric speech. The other procedure employs more sophisticated ML models such as ASR systems and E2E deep neural networks. Combinations of these two procedures have also been investigated.

Regardless of what objective procedures are adopted to assess dysarthric speech, researchers generally evaluate their procedures in two different tasks, i.e., to assess (explain) speech intelligibility and to classify speech based on metrics of disordered speech and severity levels of dysarthria. First, for the task of explaining speech intelligibility, researchers have used regression models to explore the relation between the output of objective procedures and continuous, subjective measures of intelligibility, such as linear regression (LR; Middag et al., 2008, 2009a, 2010; Van Nuffelen et al., 2009a) and support vector regression (SVR; Haderlein et al., 2011; Kim & Kim, 2012; Laaridh et al., 2017; Middag et al., 2010, 2011; Riedhammer et al., 2007; Schuster et al., 2006a). Results of regression models are

evaluated by the commonly used primary performance criteria such as Pearson Correlation Coefficient (PCC; Middag et al., 2008) and the Root Mean Squared Error (RSME; Middag et al., 2009a). Second, for the task of classifying speech or speakers according to the speech type (disordered vs healthy) or the severity level of dysarthria, researchers have used ML-based classification models such as support vector machine (SVM; Arias-Vergara 2018a; Narendra & Alku, 2018; Vásquez-Correa et al., 2019a) and convolutional neural networks (CNN; Narendra & Alku, 2020). The results of classification models are evaluated by performance criteria, such as accuracy, precision, recall, the F1-measure, and the Matthew correlation coefficient.

1.2.1 Acoustic correlates of intelligibility

Regarding the procedures studying acoustic features, here we briefly discuss five relevant features that have been frequently studied in relation to subjective intelligibility measures and classification of speakers, namely *Vowel Space Area (VSA), second-format (F2) slope, temporal features, the fundamental frequency (Fo),* and *Sound Pressure Level (SPL).* The features discussed range from phonemes to prosody as dysarthria is characterized by abnormalities in speech production at both segmental and suprasegmental levels.

Vowel Space Area (VSA). Individuals with dysarthria may experience segmental-level (articulation) impairments in speech due to dysarthria-induced speech motor disorders, and VSA is a commonly studied feature related to articulation. VSA based on the corner vowels of /i/, /a/ and /u/, as well as features related to VSA, such as formant frequencies of the corner vowels, have been found to vary in relation to the type of speech (Allison et al., 2017; Arias-Vergara et al., 2017; Lansford & Liss, 2014; López-Pabón et al., 2020) and speech intelligibility (Mendoza Ramos et al., 2021a), although the strength of the relation varied across studies (Kim et al., 2011a; Liss et al., 2000; Liu et al., 2005; McRae et al., 2002; Tjaden & Wilding, 2004; Turner et al., 1995; Weismer et al., 2000, 2001).

Second-format (F2) slope. Many researchers have also examined the relation between F2 slope and speech intelligibility (Chiu et al., 2019;

Yunusova et al., 2012). F2 slope corresponds to the rate of vocal tract shape change for vowels, with shallower slopes representing slower tongue movement speeds (Yunusova et al., 2012). Shallower F2 slopes have been reported to be associated with poorer speech intelligibility across speech materials and neurological diseases, such as Parkinson's, stroke, and traumatic brain injury (Kim et al., 2009, 2011; Tjaden & Wilding, 2013; Tjaden et al., 2013). This suggested that the F2 slope could be used as an objective indicator of speech intelligibility.

Temporal features. Temporal features refer to the features related to speech timing (duration) and speech rate. Studies examining temporal features frequently use artificially modified speech rather than natural speech, and the relation between temporal features and speech intelligibility appears to be complex. Some studies reported higher speech intelligibility when the speech rate was slowed (Hammen et al., 1994; Yorkston et al., 1990), while other studies found a decrease in intelligibility (McRae et al., 2002; Van Nuffelen et al., 2009b, 2010) or little difference between habitual and slowed speech (Dagenais et al., 2006). When comparing speech rate with articulation rate, abnormalities in articulation rate were not found in some types of dysarthria, i.e., the type of flaccid and hypokinetic dysarthria (Nishio & Niimi, 2001). These results suggested that the relation between speaking rate and articulation rate may be associated with the type of dysarthria. The remaining unclear relation between temporal features and intelligibility measures suggests the need for further exploration. Also, using natural speech instead of artificially modified speech may lead to different findings.

Fundamental frequency (F0). Speech intelligibility has also been found to vary in relation to F0. Researchers have studied various descriptors and functionals that are related to F0 (Eyben et al., 2013; Mendoza Romas et al., 2020), such as mean and standard deviation of F0, F0 range, F0 contour, and F0 variation, at the sentence and syllable levels. Research on syllable level has shown a reduction in speech intelligibility when F0 variation was reduced in adjacent syllables. Liss et al. (1998, 2000) reasoned that reduced syllabic contrasts represented by acoustic presentations including F0 can hinder a listener's ability to locate lexical boundaries within an utterance, consequently leading to reduced intelligibility. In regard to sentence-level

research, features associated with F0 have been found to be linearly related to intelligibility. Decreased intelligibility has been observed for reduced F0 range for dysarthric (Bunton et al., 2001) and healthy speakers (e.g., Bradlow et al., 1996; Tjaden et al., 2014a; Watson & Schlauch, 2008). Furthermore, dysarthria may lead to relatively flat F0 contours. To study the relation between F0 contour and sentence-level intelligibility, some researchers have adopted resynthesis techniques to modify F0 contour and have found a reduction in intelligibility for flattened F0 contour (Laures & Weismer, 1999; Watson & Schlauch, 2008). Nevertheless, some researchers have commented that the contribution of F0 to intelligibility may vary with the type of dysarthria (Watson & Schlauch, 2008) and across speakers (Feenaughty et al., 2014).

Sound Pressure Level (SPL). SPL is one of the metrics that speech-language therapists often focus on (Cannito et al., 2012; Levy et al., 2020; Nakayama et al., 2020; Neel, 2009; Ramig et al., 1994, 1995; Sapir et al., 2011; Tjaden & Wilding 2004; Yuan et al., 2020). Many researchers have reported that increases in SPLs are a beneficial effect of speech-language treatments, especially the Lee Silverman Voice Treatment (LSVT; Muñoz-Vigueras et al., 2021; Sapir et al., 2011), and that intelligibility increased as the mean SPL increased with the treatments (Cannito et al., 2012; Levy et al., 2020; Nakayama et al., 2020; Ramig et al., 1994, 1995; Yuan et al., 2020).

In addition to the several features discussed above, more and more researchers have used opensource software, such as Praat (Boersma & Weenink, 2021) and the openSMILE toolkit (Eyben et al., 2013), to extract a large number of features at once (Narendra & Alku, 2018). A feature selection method such as a principal component analysis (PCA) can be applied before studying their relations with intelligibility measures or studying their ability to classify speech.

However, the abovementioned studies are limited to the relation between different acoustic features with only one specific intelligibility measure. Thus, it is worthwhile to extend previous research to different intelligibility measures since they may be influenced by different implementations of the factors. This more comprehensive exploration could also help to understand how such different measures can be used to develop easy-to-use tools in clinical practice.

1.2.2 Machine learning-based models

Beyond the study of acoustic features of dysarthric speech, many studies have employed sophisticated ML-based models such as ASR systems and E2E deep neural networks. These ML techniques normally require a feature extraction method at the front end to convert the speech signal from the time domain to the frequency domain. Typical features extracted for each frame of the speech signal are Mel-Frequency Cepstral Coefficients (MFCC; Gurugubelli, & Vuppala, 2020; Haderlein et al., 2011; Liu et al, 2005; Middag et al., 2008, 2009a, 2009c, 2010, 2011; Riedhammer et al., 2007; Van Nuffelen, 2009a) and Perceptual Linear Prediction (PLP; Gurugubelli, & Vuppala, 2020; Martínez et al., 2013, 2015).

Aside from feature extraction, a typical pipeline ASR system contains the other three main components: an acoustic model, a language model, and a pronunciation model (dictionary). Architectures of acoustic models are Gaussian Mixture Models (GMMs; Middag et al., 2008; Riedhammer et al., 2007) or neural networks (Yilmaz et al., 2016b, 2017), in combination with Hidden Markov Models (HMMs). The language model, either n-gram HMMs or Recurrent Neural Networks (RNNs), is normally trained on a large corpus (dictionary) to maintain high accuracy.

Early work used ASR systems in the role of 'objective' listeners to recognize items, i.e., phonemes, words, and sentences, and then evaluated the performance by correlating ASR outputs, such as phoneme error rate, word accuracy (WA), word error rate (WER; e.g., Parra-Gallego et al., 2018), to subjective intelligibility measures (Kim & Kim, 2012; Maier et al., 2007; Riedhammer et al., 2007; Schuster et al., 2006a; Feng et al., 2021; Ferrier et al., 1995; Hermann & Doss, 2020; Parra-Gallego et al., 2018; Yue et al., 2020b; Thomas-Stonell et al., 1998). With the help of larger amounts of training data, the aforementioned ASR systems can result in high correlations with subjective intelligibility measures across all severity levels of dysarthria. For instance, in a study by Maier et al. (2007), an ASR model trained with the speech of healthy children and adults showed a correlation of .90 between the results obtained on the speech of children with cleft lip and their subjective intelligibility via a 5-point Likert scale. Similar performance has also been reported in many other studies (Haderlein et al., 2011; Riedhammer et al., 2007; Schuster et al., 2006a, 2006b).

Later on, due to the success of deep learning models in many other areas (e.g., image processing and natural language processing), many studies have used ASRs in combination with deep learning models, leading to hybrid architectures. In this setting, ASR systems were used to generate model-based features, such as phonological features (e.g., Middag et al., 2008; Van Nuffelen et al., 2009a), phonemic features (e.g., Middag et al., 2008, 2009a; Van Nuffelen et al., 2009a) and log-likelihoods (e.g., Hosseini-Kivanani et al., 2019; Kim & Kim, 2012; Van Nuffelen et al., 2009a), or to provide alignments between speech signal and written transcriptions. These outputs of the ASR systems are then fed into deep learning models for assessing intelligibility or classifying dysarthric and healthy speakers. Examples of this approach can be found in the studies of Middag (Middag et al., 2008, 2009a, 2010, 2011; Van Nuffelen et al., 2009a). The results of this approach were comparable to those of typical ASR systems.

In recent years, E2E models have been widely studied to assess intelligibility and to classify speakers (Cummins et al., 2020; Fritsch & Magimai-Doss, 2021; Halpern et al., 2022; Pan et al., 2020; Tripathi et al., 2021). These models can generate speaker-level representations, such as supervectors (Bocklet et al., 2009, 2012; Martínez et al., 2013), *i*-vectors (Arsikere et al., 2019; Arias-Vergara et al., 2018b; Botelho et al., 2020; Gurugubelli & Vuppala, 2020; Laaridh et al., 2017; Martínez et al., 2015; Vásquez-Correa et al., 2018b), and *x*-vectors (Botelho et al., 2020; Quintas et al., 2020). Various architectures of E2E models have been extensively investigated, such as Artificial Neural Networks (ANNs; Bhat et al., 2017; Chandrashekar et al., 2019; Middag et al., 2010; Shahamiri & Salim, 2014), GMM-based Universal Background Models (GMM-UBMs; Arias-Vergara et al., 2018b; Bocklet et al., 2009; Gu et al., 2005; Martínez et al., 2013), RNNs (Bhat & Strik, 2020; Cummins et al., 2020; Vásquez-Correa et al., 2020), CNN (Arias-Vergara et al., 2018a; Chandrashekar et al., 2019; Cummins et al., 2020; Pan et al., 2020; Vásquez-Correa et al., 2018a, 2019a, 2019b) and models based on Connectionist Temporal Classification (CTC; Tripathi et al., 2021). The performance of E2E models for assessing intelligibility is comparable to that of ASR-based models (Laaridh et al., 2017; Martínez et al., 2013; Tripathi et al., 2021).

Although the sophisticated ML-based models, either ASR systems or E2E models, have been successful in assessing speech intelligibility and classifying speakers (Schuster et al., 2006a; Middag et al., 2009a; Berisha et al., 2013; Pellegrini et al., 2015; Kim et al., 2015), several problems remain to be addressed. First, neither the ASR outputs nor the speaker-level representations captured by E2E models are easy for speech-language pathologists to interpret, not to mention being used for diagnosis. A WA score only represents an overall measure of intelligibility and does not provide detailed information about where speakers exhibit articulatory defects. If ASR systems can automatically generate transcriptions that are comparable to transcriptions provided by human listeners, we can, on the one hand, automatically assess speech intelligibility, and, on the other, extract from these automated transcriptions detailed information about articulatory deficiencies that can be used by speech-language pathologists for diagnosis.

Second, since reliable ML-based models including ASR systems essentially require large amounts of labeled data for training, dysarthric speech presents a *low-resource* problem given that its speech resources are limited. Over the past decade, the *low-resource* problem has become an increasingly interesting topic in the field of speech processing, especially in the area of speech recognition. Previous studies have investigated how to build reliable ASR systems for low-resourced or even zero-resourced languages by training ASR models on high-resourced languages (e.g., Takashima et al., 2019b; Vásquez-Correa et al., 2019b). However, so far, the *low-resource* problem has not been explored in the context of pluricentric languages with respect to speech intelligibility. In this context, training on rich speech resources of dominant varieties may potentially benefit ASR performance on non-dominant varieties, which have relatively low speech resources. Since different varieties of a pluricentric language share common linguistic characteristics, a cross-variety task may be easier than a cross-language task. It is also worth investigating the possibility of using speech resources of different varieties in an ASR system for obtaining objective assessments of speech intelligibility.

1.3 Research questions and outline

This dissertation attempts to gain insights into the development of valid procedures for measuring the intelligibility of pathological speech. To this end, this dissertation evaluates both subjective and objective procedures by addressing three research questions:

1. For subjective procedures, how do different factors influence measures of speech intelligibility? (**Chapters 2, 3, and 4**)
2. For objective procedures based on acoustic features, how do different features correlate to speech intelligibility? (**Chapters 5 and 6**)
3. For objective procedures based on ASR models, how can the low-resource problem in assessing speech intelligibility of a pluricentric language be addressed? (**Chapter 7**)

To address the first research question, we conduct three listening experiments to comprehensively study the effects of four factors, i.e., *speech materials*, *measurement methods*, *granularity levels*, and *listener characteristics*, in **Chapters 2, 3 and 4**. Specifically, for speech materials, we consider word lists, meaningful sentences, and semantically unpredictable sentences, which are commonly used and vary in length, morphosyntactic complexity, and semantic predictability. For measurement methods, we consider both categories: scalar judgments and item identifications. They are examined through the commonly used VAS and orthographic transcriptions, respectively. Particularly, a novel form of transcription that allows pseudowords is applied in order to compare intelligibility measures across speech materials since the material of word lists contains pseudowords. As a result, listeners have more freedom in transcribing than in the typical form of transcription. For granularity levels, we consider subword (i.e., grapheme and phoneme), word, and utterance levels. For listener characteristics, we report the listener gender, constrain the listener familiarity to no familiarization, and constrain the listeners to be experienced. In addition, we explore the impact of listener experience by further comparing the scores with those in **Chapter 5**, where naïve listeners are involved.

Specifically, **Chapter 2** provides a comprehensive analysis of eight measures varied in granularity levels in the three experiments to study

whether these measures are reliable and valid. The interrater reliability of the measures is examined by the commonly applied ICC analysis. The validity of the measures is studied by examining the relation between these measures and another index of dysarthria, i.e., the severity level of dysarthria.

Chapter 3 further investigates two measures at the utterance and word levels in terms of a comprehensive analysis of reliability and expanded investigations of validity, and examines the usability of our novel pseudoword-allowing form of transcription. The reliability is studied by applying G Theory, which is relatively unknown in the field of pathological speech and language research, but which enables more comprehensive analyses than traditional methods such as ICC. Furthermore, G Theory is used to calculate the optimal number of listeners and utterance samples required to obtain reliable measures by conducting a decision study. By doing so, we can provide solid suggestions on the implementation of subjective procedures for measuring intelligibility in clinical trials. The validity analyses are expanded by studying the correlations of the measures within and across speech materials (construct validity) and by examining their relation to two external variables (concurrent validity).

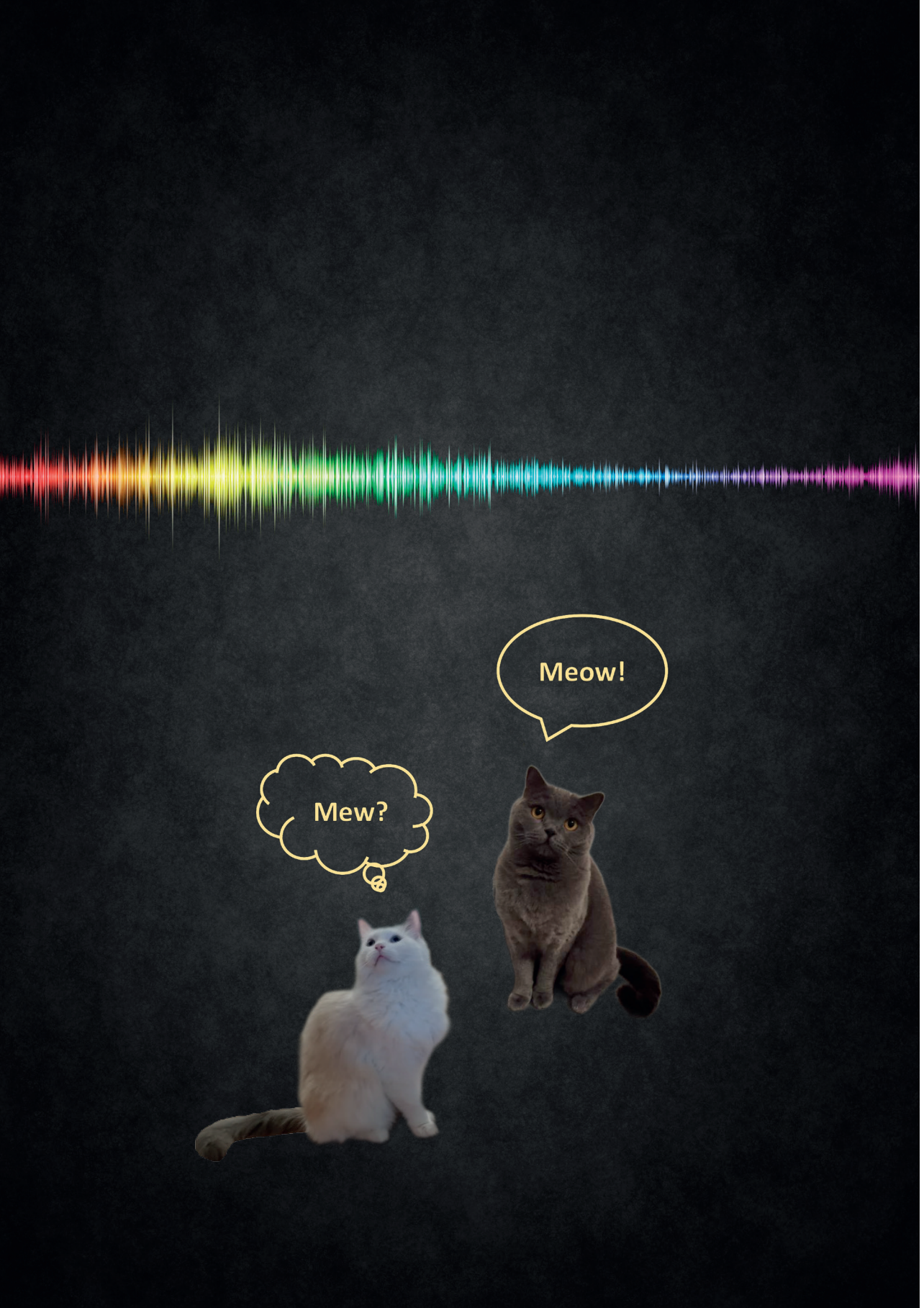
Following up on the preceding chapter's results, **Chapter 4** takes a closer look at two phoneme-level measures concerning the reliability, validity, and performance in classifying speakers according to speaker types and severity levels of dysarthria. The phoneme-level measures can be seen as representations of articulation, a key feature of dysarthria. As articulation has been found to be a stronger contributor to intelligibility compared to voice quality, prosody, and nasality (De Bodt et al., 2002), studying phoneme-level measures can help to better evaluate articulation imprecision in speakers with dysarthria and to monitor the effectiveness of therapy. Investigating phoneme-level measures' performance for classification may provide insights into building semi-objective procedures for assessing intelligibility.

To address the second research question, we extend previous research on acoustic correlates of speech intelligibility to different intelligibility measures by applying stepwise regression models in **Chapters 5 and 6**. Specifically, **Chapter 5** studies a small set of features that are related to

pitch, intensity, and formant frequencies. The features are extracted from both dysarthric and healthy speech, and a stepwise logistic regression model is applied to select relevant features to classify dysarthric and healthy speech. Based on the outcomes of the regression model, we calculate an acoustic-phonetic probability index (API) and study its relation with subjective measures of intelligibility at the utterance and word levels. **Chapter 6** studies a larger acoustic feature set – eGeMAPS, including features related to e.g., frequency, amplitude, and spectrum, and its relation with a phoneme-level measure, i.e., Phoneme Intelligibility (PI), in two types of speech materials. A set of temporal features is also considered to explore whether the relation between acoustic features and subjective intelligibility measures is material-dependent.

To address the last research question, we evaluate the contribution of resources from the dominant variety (Netherlandic Dutch) of a pluricentric language (Dutch) to improving the ASR models on the non-dominant variety (Flemish Dutch) in terms of predicting subjective measures of intelligibility and, for the first time, generating human-comparable transcriptions, in **Chapter 7**. The aim of studying the possibility of generating human-comparable transcriptions is to explore whether ASR models can, on the one hand, fully replace human listeners in the assessment of intelligibility and, on the other hand, retain the deviations of dysarthric speech so that therapists can further evaluate and use them for diagnosis.

Lastly, **Chapter 8** summarizes the findings in the preceding chapters and provides a general discussion. The discussion gives suggestions on the assessment of speech intelligibility with respect to the effects of the factors in subjective procedures, the acoustic correlates of intelligibility in objective procedures, and the possibility of solving the low-resource problem in ASR models in terms of intelligibility assessment and transcription generation. Together, these results and the discussion lead to guidelines for developing valid measurement procedures.


Meow!
Mew?

CHAPTER 2

TOWARDS A COMPREHENSIVE ASSESSMENT OF SPEECH INTELLIGIBILITY FOR PATHOLOGICAL SPEECH

ABSTRACT

Speech intelligibility is an essential though complex construct in speech pathology. It is affected by multiple contextual variables and is often measured in different ways. In this chapter, we evaluate various measures of speech intelligibility, with respect to their reliability and validity. For this study, the speech of three different speech materials was analyzed together with their respective perceptual ratings assigned by five experienced speech-language pathologists: a Visual Analogue Scale (VAS) and two types of orthographic transcriptions, one in terms of existing words and the other in terms of perceived segments, including pseudowords. Six subword measures concerning graphemes and phonemes were derived programmatically from these transcriptions. All measures exhibit high degrees of reliability. Correlations between the six subword measures and three independent measures (i.e., VAS, word accuracy, and severity level) reveal that the measures extracted programmatically from the orthographic transcriptions are valid predictors of speech intelligibility. We also observed differences in the three speech materials, suggesting that a comprehensive assessment of speech intelligibility requires different materials in combination with measures at different granularity levels (i.e., utterance, word, and subword). We discuss these results in relation to those of previous research and provide suggestions for possible avenues of future research.

This chapter is based on the following publication:

Xue, W., Mendoza Ramos, V., Harmsen, W., Cucchiarini, C., Van Hout, R. W. N. M., & Strik, H. (2020). Towards a comprehensive assessment of speech intelligibility for pathological speech. In Proceedings of Interspeech 2020, 3146-3150.

2.1 Introduction

Speech disorders in general, and dysarthria especially, lead to decreased speech intelligibility. This can have a severe impact on patients' quality of life because they can lose social contact and eventually become isolated from society. Most of the time, degraded speech intelligibility can be improved through speech therapy. However, the effects of intensive therapy are not always evident. For monitoring a possible evolution, pre- and post-therapy evaluations in which intelligibility scores play an important role are necessary. Thus, intelligibility requires a clear definition and a robust operationalization.

A clear definition has been proposed by Hustad (2008) "Intelligibility refers to how well a speaker's acoustic signal can be accurately recovered by a listener" (p. 562). In line with this definition, intelligibility can be measured in various ways. One of them is based on orthographic transcriptions of sentences, words, or phonemes (Kempler & Van Lancker, 2002; Yorkston & Beukelman 1978). The percentage of words or phonemes correctly identified is employed as a measure of intelligibility, and it is used in the Sentence Intelligibility Test (Yorkston et al., 1996a). Besides, intelligibility has also been measured by collecting scalar ratings from human listeners (Barreto & Ortiz, 2008; Miller 2013; Yorkston & Beukelman, 1978) through equal-appearing interval scales like the Likert scales (Yorkston & Beukelman, 1978), or by placing a point on a horizontal line like the visual analogue scale (VAS; Finizia et al., 1998).

All the methods described above rely on perceptual judgments. It is common practice to collect different measures from multiple listeners and check their reliability before the usage for future research purposes (Hustad, 2007; Stipancic et al., 2016; Ganzeboom et al., 2016) since these measures can be influenced by, for example, the listener experience. As pointed out by Mencke et al. (1983), measures collected from naïve listeners showed larger variances than those collected from well-trained expert listeners such as speech-language therapists (Mencke et al., 1983). In addition, the measures can also be influenced by the listeners' familiarity with the speech material. Specifically, Beukelman and Yorkston (1980) reported that the estimates of speech intelligibility increased as the listeners became familiar with the reading passage. Given that all these operations are time-

consuming, costly, and laborious in practice, there is a need for obtaining valid measures of speech intelligibility in a more objective way.

In line with this need, several studies have investigated the relationship between perceptual ratings of intelligibility and various automatically calculated measures. These automated measures are normally obtained using machine learning-based models such as ASR (Schuster et al., 2006a; Middag et al., 2008; Martinez et al., 2013) or neural networks (Van Nuffelen et al., 2009a, 2010; Middag et al., 2011, 2010; Yilmaz et al., 2017). Very high correlations have been reported. For example, the magnitude of correlations between ASR outputs (e.g., word accuracy and word error rate) and intelligibility ratings measured by a 5-point Likert scale reached 0.9 for patients with cancer of oral cavity (Schuster et al., 2006a) and 0.92 for children with cleft lip (Middag et al., 2008). Typical models can also be found in Van Nuffelen et al. (2010), Middag et al. (2011, 2009c, 2010), and Yilmaz et al. (2017), and their performance is comparable to the results presented above (e.g., Van Nuffelen et al., 2009a). In general, these measures are not detailed enough to be used by therapists to diagnose the problems that led to decreased intelligibility.

For this reason, a semi-automatic approach was proposed by Ganzeboom et al. (2016). In the study, a set of intelligibility ratings of disordered speech assigned by naïve listeners were investigated to obtain measures at three different levels of granularity: utterance, word, and subword level including grapheme and phoneme levels. Utterance-level evaluations were obtained using subjective rating scales (i.e., VAS and Likert scale), and word- and subword-level evaluations, i.e., distance scores, were obtained programmatically from human-generated orthographic transcriptions using automatic alignment and grapheme-phoneme conversion algorithms. The results indicated that the distance measure at the phoneme level was feasible and reliable, and it was a more sensitive measure to changes within patients, thus providing an informative measure of intelligibility.

In this chapter, we extend this semi-automatic approach based on orthographic transcriptions and its programmatically-derived metrics as intelligibility measures of pathological speech on a number of important points. First, we collect intelligibility measures for a larger number of speech samples, including both pathological and normal ones, covering different speech materials. Intelligibility measures are explored in relation to speaker

types and speech materials. Second, we collect ratings from experts as opposed to naïve listeners. Third, we ask the listeners to provide two types of transcriptions, one in terms of existing words and the other in terms of literal or perceived segments. More detailed intelligibility measures are calculated programmatically from them. Fourth, we evaluate the reliability and validity of intelligibility measures obtained from transcriptions in relation to other measures such as VAS and severity level of dysarthria.

In the remainder of the chapter, we first describe the experimental design and explain how measures of speech intelligibility were computed at different levels of granularity in Section 2.2. In Section 2.3, we present the results, and in Section 2.4, we discuss our findings and present ideas for future research.

2.2 Method

2.2.1 Experimental design

2.2.1.1 Speakers and speech materials

This study considers different speech materials that are often used in clinical practice to perceptually assess speech intelligibility. The speakers were selected to cover different types of dysarthria and severity levels (SL), at different ages and were gender-balanced.

The investigation comprised three experiments. Experiment 1 included 36 speakers (10 control and 26 speakers with dysarthria). For each speaker, the same four sentences were selected from the Dutch phonetically balanced text 'Papa en Marloes'. In Experiment 2, 18 speakers (4 control and 14 speakers with dysarthria) read the 50 existing and non-existing consonant-vowel-consonant words of the Dutch Intelligibility Assessment (DIA) task (De Bodt et al., 2006). The recordings used in the first two experiments were selected from the Dutch Corpus of Pathological and Normal Speech (COPAS; Middag, 2012). In Experiment 3, 23 speakers with dysarthria were involved, and for each speaker, six Semantically Unpredictable Sentences (SUS) with different lengths were selected from the Dutch Sentence Intelligibility Assessment (De Bodt et al., 2006; Martínez et al., 2011).

The severity levels of the speakers had been assigned by speech-language pathologists on a four-category scale (normal-mild-moderate-

severe), and they were already available in the speech corpora. Specifically, for Experiments 1 and 2, the 'normal' refers to the healthy control speakers, and for Experiment 3, the 'normal' category includes two speakers with mild dysarthria whose speech was classified as clear as healthy control speakers. The frequency plot of the severity level of speakers in all three experiments is shown in Figure 2.1.

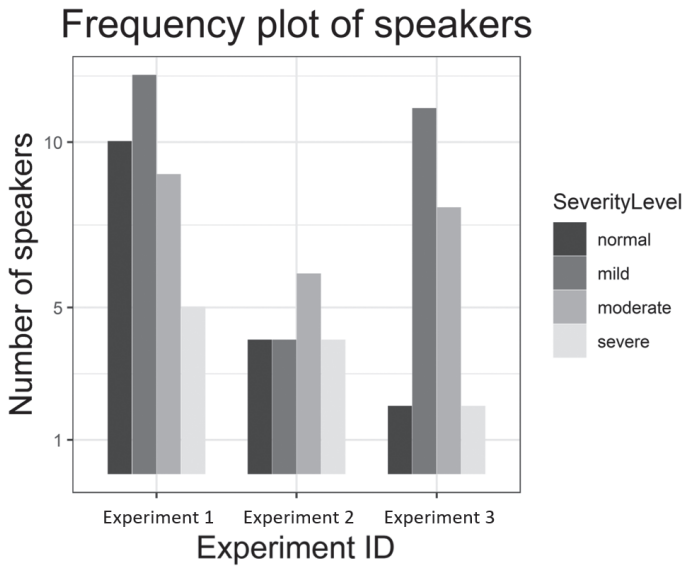


Figure 2.1. Frequency plot of the severity levels of speakers in all three experiments.

2.2.1.2 Procedure and listeners

Five expert listeners perceptually judged intelligibility. They were not familiar with the materials used in Experiments 2 and 3. However, they were familiar with the sentences in Experiment 1 since these sentences are part of a phonetically balanced text that is commonly used in their clinical practice.

For each recording, the experts perceptually rated intelligibility at the utterance level through a Visual Analogue Scale (VAS) and made two types of orthographic transcriptions which are Word – using only meaningful words, and Literal – also allowing pseudowords, to reflect the speech segments the experts perceived. Since Experiment 2 contains pseudowords, they only made Literal transcriptions. They were allowed to listen only once

to the sample and subsequently score it (i.e., assigning a VAS score and making a Literal transcription). In the other two experiments, they were allowed to listen twice to each sample to perform the task (i.e., assigning a VAS score and making two types of transcriptions). Limiting the number of listening times was done to reduce the impact of familiarity. The samples were also randomized to prevent any systematic order effect.

2.2.2 Intelligibility measures

Intelligibility measures at different levels of granularity were collected and calculated. The same eight intelligibility measures at the utterance, word and subword levels were obtained in all three experiments.

2.2.2.1 Intelligibility measures at the speaker and utterance levels

The severity levels of dysarthria were used as speaker-level measures. Utterance-level intelligibility ratings were obtained using VAS ranging from 0 (not intelligibility) to 100 (intelligible).

2.2.2.2 Intelligibility measures at the word level

The orthographic transcriptions were compared to the reference transcriptions after removing punctuation and symbols indicating missing words. Note that for this measure, pseudowords were treated as words. Based on these transcriptions, the accuracy (Acc) of words was computed as follows:

$$Acc = (N_{total} - N_d - N_s)/N_{total} \times 100$$

where N_{total} denotes the total number of words in the reference transcriptions, and N_d and N_s denote the number of deletions and substitutions in the corresponding orthographic transcriptions, respectively.

2.2.2.3 Intelligibility measures at the subword level

At the subword level (grapheme and phoneme), in addition to Acc, another two measures, the distance (Dist) between two transcriptions and the number of changes (Ch), were extracted as follows:

$$Dist = N_d \times C_d + N_i \times C_i + N_s \times C_s$$

$$Ch = N_d + N_i + N_s$$

where N_d, N_i and N_s denote the number of deletions, insertions, and substitutions, respectively. C_d, C_i and C_s denote their corresponding costs. For grapheme, C_d=1, C_i=1, and C_s=2. For phonemes, C_d=3, C_i=3, and C_s was calculated by employing matrices with articulatory features (Elffers et al., 2005; Cucchiarini, 1996). These measures were calculated by the software ADAPT. More details about ADAPT, the extracting procedures and the grapheme to phoneme conversion can be found in Ganzeboom et al. (2016) and Elffers et al. (2005).

2.3 Results

In this section, we present the experimental results of the different measures concerning reliability (Section 2.3.1), mean and standard deviations (Section 2.3.2), and correlations (Section 2.3.3).

2.3.1 Reliability of measures

The reliability of all eight measures in the three experiments was calculated using ICC (2, k) (items and listeners random with k = 5) in *psych* package in R (Revelle, 2019) since all the listeners assigned judgments to all the recordings from all the speakers.

The analyses reveal that for the eight measures, the reliability coefficients for the five listeners together are generally very high, above 0.90, except for a relatively lower reliability (0.82) for accuracy at the word level (Acc-W) in Experiment 1 in Word transcription.

2.3.2 Mean and standard deviation of measures

The mean and standard deviation (SD) of the intelligibility measures are shown in Table 2.1. For Experiment 2, we observed worse intelligibility, probably because of the isolated pseudowords. Experiments 1 and 3 show better scores. Moreover, the scores for Experiment 1 are higher compared to Experiment 3. Possible explanations are that Experiment 1 has a higher number of normal speakers than Experiment 3, and the latter used semantically unpredictable utterances. Besides, in both Experiments 1 and 3, better intelligibility scores are observed for Word than for Literal transcription.

Table 2.1. Mean and standard deviation of eight different measures in our three experiments. For the utterance level, we use Visual Analogue Scales (VAS). For the word (W) level, we use accuracy (Acc). For the grapheme (G) and phoneme (P) levels, we use Acc as well as Distance (Dist) and number of Changes (Ch). Specifically, for the last three levels, we have two types of transcription in Experiment 1 and Experiment 3: Literal (left part of the cell) and Word (right part of the cell)[1].

Levels	Measures	Experiment 1	Experiment 2	Experiment 3
Utterance	**VAS**	84.25 (25.52)	68.12 (25.29)	79.9 (24.48)
Word	**Acc-W**	79.99 (29.31) / 88.09 (24.50)	41.68 (24.52)	71.27 (29.01) / 76.74 (29.00)
Grapheme	**Acc-G**	92.29 (15.85) / 95.15 (13.03)	80.34 (12.65)	88.86 (15.53) / 89.02 (16.90)
	Dist-G	4.78 (10.13) / 3.24 (8.70)	30.28 (19.35)	6.89 (9.72) / 6.82 (10.39)
	Ch-G	3.89 (8.10) / 2.65 (6.87)	18.77 (12.08)	5.11 (7.31) / 4.98 (7.74)
Phoneme	**Acc-P**	91.13 (17.73) / 94.47 (14.68)	77.75 (13.92)	87.73 (17.16) / 88.14 (18.27)
	Dist-P	10.59 (21.58) / 7.17 (18.27)	37.64 (26.25)	12.28 (18.12) / 11.86 (19.12)
	Ch-P	3.81 (7.76) / 2.56 (6.62)	16.39 (10.58)	4.69 (6.82) / 4.53 (7.19)

[1] The mean values of each measure were compared between every two experiments through a *t* test and were found significantly different ($p < 0.05$), except for Dist-P derived from Literal Transcription between Experiment 1 and Experiment 3.

2.3.3 Intelligibility measures and correlations

The median values of the correlations between each pair of the subword-level measures turned out to be always higher than 0.9, with the lowest minimum correlation of 0.87.

Table 2.2 shows the correlations between SL at the speaker level, VAS at the utterance level and Acc at the word level, the three measures we want to use as criterion variables for the six subword-level measures. We can observe moderate to strong correlations between the three measures. Probably these three measures reflect different aspects of intelligibility, and thus, they constitute an interesting combination to investigate the validity of the subword-level measures. Besides, the SL correlations do not reflect intra-speaker variation with the consequence that these correlations are systematically lower than the other correlations.

Table 2.2. Correlations between VAS, Accuracy of words (Acc-W) and speakers' Severity Level of dysarthria (SL) in our three experiments. For Experiments 1 and 3, we have two types of transcription: Literal (left part of the cell) and Word (right part of the cell). The Multiple R was computed when SL was involved, treating SL as a nominal variable.

Correlation	VAS vs Acc-W	VAS vs SL	Acc-W vs SL
Experiment 1	0.85 / 0.73	0.70	0.60 / 0.70
Experiment 2	0.75	0.69	0.70
Experiment 3	0.71 / 0.74	0.51	0.46 / 0.48

Table 2.3 gives the correlations of the six subword-level measures and our three criterion variables mentioned above. Table 2.3 shows that the Acc measures (17 times the highest correlation) outperform the Dist and Ch measures (8 times the highest correlation). Within the Acc measures, Acc at the phoneme level (Acc-P) performs better (13 times) than Acc at the grapheme level (Acc-G) (4 times). We also looked at non-parametric Spearman correlations, but that did not change the overall strength of the correlations or the overall correlational pattern.

Table 2.3. Correlations between measures at the subword level and measures at the utterance, word and speaker levels. For Experiments 1 and 3, we have two types of transcription: Literal (left part of the cell) and Word (right part of the cell). The Multiple R was computed when SL was involved, treating SL as a nominal variable. For each row, the **bold** numbers are the highest ones in each type of transcription (Literal / Word).

	Correlation		Grapheme			Phoneme		
			Dist	**Ch**	**Acc**	**Dist**	**Ch**	**Acc**
Experiment 1	**Utterance**	**VAS**	0.83 / 0.75	0.81 / 0.73	0.88 / 0.78	0.81 / 0.72	0.83 / 0.74	**0.89 / 0.79**
	Word	**Acc-W**	0.77 / 0.79	0.77 / 0.79	0.88 / 0.89	0.78 / 0.79	0.78 / 0.80	**0.89 / 0.90**
	Speaker	**SL**	0.58 / 0.52	0.56 / 0.51	0.59 / 0.52	0.56 / 0.50	0.57 / 0.51	**0.60 / 0.53**
Experiment 2	**Utterance**	**VAS**	0.78	0.78	**0.79**	0.76	0.77	**0.79**
	Word	**Acc-W**	**0.90**	**0.90**	**0.90**	0.83	0.87	**0.90**
	Speaker	**SL**	0.69	0.68	0.70	0.67	0.69	**0.71**
Experiment 3	**Utterance**	**VAS**	0.72 / 0.72	0.72 / 0.72	**0.73 / 0.75**	0.71 / 0.71	0.72 / 0.72	**0.73 / 0.75**
	Word	**Acc-W**	0.81 / 0.86	0.80 / 0.85	0.88 / 0.91	0.79 / 0.84	0.80 / 0.85	**0.89 / 0.92**
	Speaker	**SL**	0.55 / **0.55**	**0.56 / 0.55**	0.43 / 0.46	0.55 / **0.55**	**0.56 / 0.55**	0.45 / 0.47

2.4 Discussion and conclusions

In this chapter, we have conducted an extended evaluation of a semi-automatic approach to measuring the intelligibility of pathological speech in which we have investigated the reliability and validity of several descriptors of intelligibility for normal and pathological speech. The analyses reveal that for the eight measures we acquired, the reliability coefficients were very high in the different experiments. This supports the usability of these measures, in particular, because it means that a limited number of listeners might be sufficient to obtain highly reliable ratings, which is of course very important in a clinical setting. The measures used show differences between the two types of transcription, word and literal, with the latter displaying more variability, as would be expected in a non-lexically driven context. As a consequence, all measures indicating accuracy at the various granularity levels are lower in the literal condition than in the word condition. All the measures appear to be sensitive to the different severity levels of the speakers with dysarthria.

We were interested in whether we could derive more detailed information from the expert transcriptions than merely computing the percent accuracy at the word level, as is often done (see e.g., Ganzeboom et al., 2016). We computed six additional measures at the subword level, three based on graphemes and three based on phonemes. In both cases, we computed Accuracy (Acc), distance (Dist) and the number of changes (Ch).

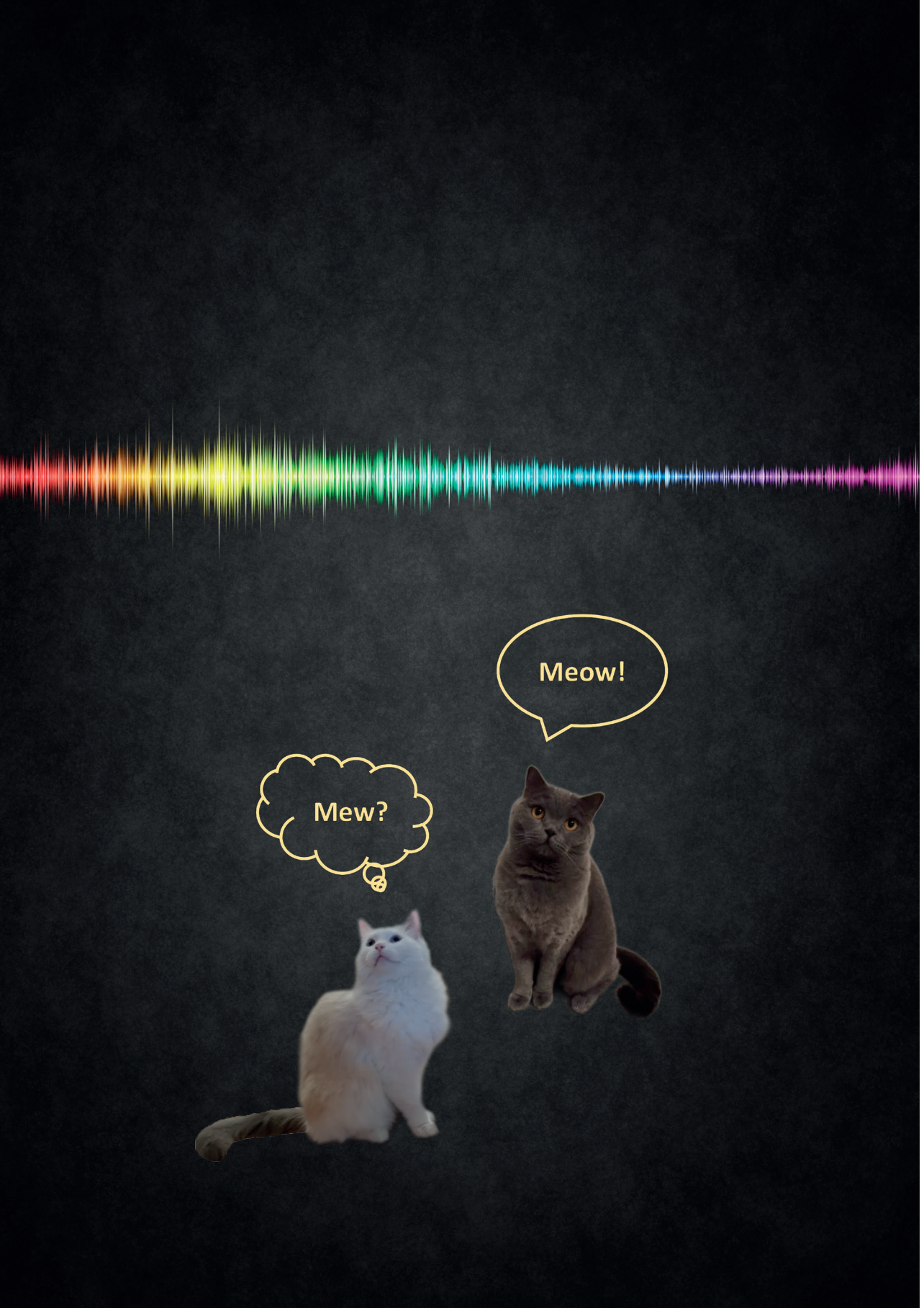
At the subword level, the mean values of Acc were very similar to Ch, with the phoneme-level results always slightly lower than those at the grapheme level. This is understandable because a phoneme may be associated with more than one grapheme, and then its overall correctness requires correctness in its associated graphemes. The six programmatically calculated subword-level measures are strongly correlated with each other, which could be explained by the fact that they are all based on the same orthographic transcriptions. However, it is worth noting that they are strongly related and that it does not make much difference using one or the other. For instance, in terms of use in clinical practice, a grapheme-level measure may be easier to apply than a phoneme-level one, but both will yield accurate results.

The results also showed that these orthography-based measures are strongly correlated with an independent measure, the VAS ratings, which are based on the listeners' perceptual judgments. To test the external validity of the subword-level measures, we included the accuracy at the word level and the dysarthria severity level at the speaker level as additional evaluation criteria. The correlations between the above three measures are moderate to strong (see Table 2.2), presumably showing that evaluative components involved in estimating intelligibility are different.

Correlations between these three measures and the six subword-level measures (see Table 2.3) indicate that the phoneme measures outperformed the grapheme measures and that the best phoneme measure seems to be accuracy. This suggests that the orthography-based subword-level measures investigated are not only reliable indicators of speech intelligibility, but that they can also be considered as valid descriptors of speech intelligibility in pathological speech.

These results show the possibility of using orthographic transcriptions and the programmatically derived phoneme measures to determine which mispronounced phonemes cause decreased speech intelligibility. In other words, these measures have potentially additional diagnostic value and can, therefore, be applied in speech therapy. Compared with other automatic measures (Schuster et al., 2006a; Middag et al., 2008), these measures can provide more detailed information. In addition, they can be easily obtained from orthographic transcriptions without time-consuming human annotations at the phoneme level.

Future work will explore the possibility to fully automate intelligibility evaluation without any human-generated orthographic transcriptions. This could be achieved with the help of ASR technology and the increasing availability of dysarthric speech data (Yilmaz et al., 2016b; Le et al., 2016; Martinez et al., 2013, 2015). Another option for prompted speech would be to use ASR in forced alignment mode, which is one of the methods we intend to investigate in future research.


Meow!
Mew?

CHAPTER 3

ASSESSING SPEECH INTELLIGIBILITY OF PATHOLOGICAL SPEECH: TEST TYPES, RATINGS, AND TRANSCRIPTION MEASURES

ABSTRACT

Speech intelligibility is an essential though complex construct in speech pathology. In this chapter, we investigated the interrater reliability and validity of two types of intelligibility measures: a rating-based measure, through Visual Analogue Scales (VAS), and a transcription-based measure called Accuracy of Words (AcW), through two forms of orthographic transcriptions, one containing only existing words (EWTrans) and one allowing all sorts of words, including both existing words and pseudowords (AWTrans). Both VAS and AcW scores were collected from five expert listeners. We selected speakers with various severity levels of dysarthria (SevL) and employed two types of speech materials, i.e., meaningful sentences and word lists. To measure reliability, we applied Generalizability Theory, which is relatively unknown in the field of pathological speech and language research but enables more comprehensive analyses than traditional methods, e.g., the intraclass correlation coefficient. The results convincingly indicate that five expert listeners were sufficient to provide reliable rating-based (VAS) and transcription-based (AcW) measures, and that reliability increased as the number of listeners or utterances increased. Generalizability Theory has proved effective in systematically dealing with reliability issues in our experimental design. We also investigated construct and concurrent validity. Construct validity was addressed by exploring the correlations between VAS and AcW within and across speech materials. Concurrent validity was addressed by exploring the correlations between our measures, i.e., VAS and AcW, and two external measures, i.e., Phoneme intelligibility and SevL. The correlations corroborate the validity of VAS and AcW to assess speech intelligibility, both in sentences and word lists.

This chapter is based on the following publication:

Xue, W., van Hout, R., Cucchiarini, C., & Strik, H. (2021). Assessing speech intelligibility of pathological speech: test types, ratings and transcription measures. *Clinical Linguistics & Phonetics*. Advance online publication. DOI: 10.1080/02699206.2021.2009918

3.1 Introduction

Dysarthria is a motor speech disorder caused by neurological injury, e.g., Parkinson's disease and stroke. It can result in losing control of tongue, larynx, vocal folds and surrounding muscles, thus leading to reduced speech intelligibility and possibly to consequent loss of social participation (Hustad, 2008). Patients with dysarthria may receive intensive speech therapy, e.g., Lee Silverman Voice Treatment (LSVT), to improve their intelligibility (Cannito et al., 2012; Levy et al., 2020; Mendoza Romas et al., 2021b; Nakayama et al., 2020; Yuan et al., 2020). For diagnosis and therapy effectiveness, it is necessary to have a clear definition of speech intelligibility. Over the years different definitions and related measurement methods of speech intelligibility have been advanced (for an overview, see Miller, 2013; Barreto & Ortiz, 2008). In our research, we have adopted the definition proposed by Hustad (2008): "how well a speaker's acoustic signal can be accurately recovered by a listener" (p. 562).

According to this definition, an acceptable measurement method of speech intelligibility could be an orthographic transcription that represents the number of correctly perceived words by a listener. In collecting these transcriptions, listeners are normally instructed to use only existing words. Accordingly, word lists consisting of only existing, meaningful words are commonly used, as well as sentences, which are semantically either predictable (meaningful) or unpredictable (Abur et al., 2019; Barreto & Ortiz, 2008, 2016; Beijer et al., 2012b; Carvalho et al., 2021; Ganzeboom et al., 2016; Hodge & Gotzke, 2014; Hustad & Cahill, 2003; Hustad, 2006, 2007, 2008; Ishikawa et al., 2020; Liss et al., 2002; Middag, 2012; Miller, 2013; Stipancic et al., 2016; Tjaden & Liss, 1995a, 1995b; Tjaden et al., 2014a; Tjaden & Wilding, 2010; Sussman & Tjaden, 2012; Xue et al., 2020, 2021b; Yorkston & Beukelman, 1978, 1981).

However, such transcriptions force listeners to align a sequence of phonemes with an existing word, and thus, more specific information on speech deviations is likely to be omitted. Therefore, less restricted instructions about the transcriptions may be needed to capture this information. For example, Chapter 5 found a very low interrater reliability (0.47) for a word-level measure of intelligibility, namely word accuracy, obtained from orthographic transcriptions based on existing words only. This reliability was much lower than that (0.93) for the measure of intelligibility obtained

with Visual Analogue Scales (VAS). The low reliability of word accuracy scores was caused by the poor variability of scores since many of the utterances received scores of 100, indicating perfect intelligibility. On the other hand, these utterances did receive VAS scores lower than 100, indicating imperfect and lower intelligibility. Also, the VAS scores had a wider range of variability. These results suggested that although listeners may have perceived speech deviations, e.g., a distortion, deletion, or substitution of a phoneme in a word, as indicated by the imperfect intelligibility and the wide range of variability in the VAS scores, they still had to transcribe the same existing, meaningful words to follow the instructions requiring only existing words. Alternatively, instructions that allow also pseudowords would seem to provide listeners with more flexibility in transcribing words and so help report speech deviations at the segmental level.

Because orthographic transcriptions are laborious and time-consuming, researchers have adopted more efficient methods that are based on listeners' perception or estimates of speech intelligibility (Abur et al., 2019; Ganzeboom et al., 2016; Ishikawa et al., 2020; Sussman & Tjaden, 2012; Xue et al., 2020; Yorkston & Beukelman, 1978, 1981; Yorkston et al., 1996a). In such methods, listeners are asked to indicate how intelligible a speech utterance is by assigning a numeric value on a scale such as a VAS (Abur et al., 2019; Ganzeboom et al., 2016; Stipancic et al., 2016; Sussman & Tjaden, 2012; Xue et al., 2020; Yorkston & Beukelman, 1978), or an x-point scale with equal appearing intervals, as in Likert scales (EAI; Ganzeboom et al., 2016; Miller, 2013), or by estimating the percentage of understandable words, which is called percentage estimates (Yorkston & Beukelman, 1978). Studies using these methods have found different results. Some researchers reported that transcription-based measures showed higher intelligibility scores for the same speech samples than rating-based ones using EAI and VAS (Ganzeboom et al., 2016), or percentage estimates (Hustad, 2006). Other studies found transcription-based measures were comparable to percentage estimates (Yorkston & Beukelman, 1978) or were highly correlated to VAS scores (Abur et al., 2019; Ishikawa et al., 2020; Schiavetti, 1992; Stipancic et al., 2016).

A broad consensus is that orthographic transcriptions yield reliable and valid measures (Bunton et al., 2001; Miller, 2013; Tjaden & Wilding, 2010) since they rely on the amount of information listeners accurately perceived. In contrast, scale ratings have been questioned since they rely on the listeners'

impression of intelligibility. Scale ratings have been found to yield measures with reliability values too low for research purposes (Miller, 2013; Schiavetti, 1992). Nevertheless, rating-based measures through VAS have shown promise for measuring intelligibility (Kent & Kim, 2011; Van Nuffelen et al., 2010), with interrater reliability comparable to that of orthographic transcriptions.

A point of concern here is that the interrater reliability of these measures has been evaluated by different statistical analyses. Some studies (e.g., Hustad, 2007; Van Nuffelen et al., 2008) reported percentage agreement between listeners without dealing with chance agreement. Others (e.g., Hustad, 2006, 2008) used Pearson correlations taking only listeners as a source of variance. The Intraclass Correlation Coefficient (ICC; Fisher, 1992) has become the standard measure of interrater reliability in pathological speech research (Rietveld, 2020). However, ICC can only handle two factors, i.e., speaker and listener, in a crossed design where all listeners assess all speakers. This approach to reliability has been expanded into an overarching type of analysis, called Generalizability Theory (G Theory; Brennan, 2001), which is based on the ICC but can take more than two sources of variance (utterances, speakers, and listeners in our case) into account. G Theory can handle not only crossed designs but also nested designs, in which different listeners assess different utterance samples of one or more speakers. In addition, G Theory allows calculating the optimal number of listeners and utterance samples required to obtain reliable measures by conducting a decision study. In fact, a growing number of studies have conducted reliability analyses through G Theory and showed effectiveness in defining optimal measurement procedures in different disciplines. For example, Ford and Johnson (2021) explored a multidimensional understanding of reliability through G Theory and gained insights into optimal measurement procedures for examining the language of preschool educators interacting with children with an autism spectrum disorder. O'Brian et al. (2003) applied G Theory for assessing the reliability of ratings from 15 listeners through the 9-point speech naturalness scale for adults' speech collected before and after treatment for stuttering. They successfully distinguished various sources of measurement error and used these to estimate the minimum number of listeners and ratings per listener for a reliable result. Hollo et al. (2020) applied G Theory to optimize the analysis of spontaneous teacher talk in elementary classrooms with teacher and sample duration as two factors. They assessed the minimum number and duration of samples needed for a reliable result. They

found that a large proportion of variance was attributable to individuals rather than the sampling duration. To the best of our knowledge, G Theory has not been used for investigating experimental designs of intelligibility assessment of pathological speech. Rietveld (2020) explains the relevance of the G Theory approach in speech and language pathological research when multiple sources of variance are involved, including listeners.

Another point of concern is that, up to now, relatively few studies have addressed the validity of speech intelligibility measures (Ellis & Fucci, 1991; Hustad, 2007; Stipancic et al., 2016; Van Nuffelen et al., 2008). Validity indicates the extent to which the scores measure what they intend to measure. This is a key question in research, and it is therefore important to investigate validity. Studies addressing the validity of intelligibility measures were normally conducted with hearing-impaired subjects or children. Barreto and Ortiz (2016) investigated the criterion validity of a transcription-based measure using two types of materials, i.e., sentences and word lists. Validity was studied in relation to speaker types, i.e., control and dysarthric speakers. Word lists appeared to have significantly greater discriminatory power than sentences. Hodge and Gotzke (2014) evaluated the construct-related validity of the Test of Children's Speech (TOCS), which uses transcription-based measures for children with and without a speech disorder. Results supported the usage of TOCS as a valid tool for measuring the intelligibility of children.

Many factors such as speech materials, severity levels of dysarthria, and listeners' experience and familiarity have been shown to affect intelligibility measures (Miller, 2013). First, intelligibility measures perform differently on different lengths of speech materials, but this difference does not seem to be consistent across different speakers' severity levels. Specifically, intelligibility scores for sentences have been found higher than those for words due to additional contextual cues when speech is mildly and moderately dysarthric (Hustad, 2007; Yorkston & Beukelman, 1978, 1981). However, when speech is more severely dysarthric, the intelligibility measures collected from sentences could be higher than, equal to (Barreto & Ortiz, 2008; Dongilli, 1994; Middag et al., 2009a; Yorkston & Beukelman, 1978), or lower (Yorkston & Beukelman, 1981) than those from isolated words. One possible reason may be that speakers with more severe dysarthria might have so many difficulties in producing sentences that listeners are no longer able to benefit from the contextual cues present in sentences. Second, listeners' experience

could also influence intelligibility assessment. For instance, Carvalho et al. (2021) reported significant differences in speech intelligibility ratings of speakers with Parkinson's disease assigned by healthcare professionals, referring to expert listeners, and naïve listeners. Similarly, Monsen (1983) found that expert and naïve listeners significantly differed in the evaluation of intelligibility in adolescents with hearing impairment. In contrast, other researchers reported no differences (Ellis & Fucci, 1991; Maruthy & Raj, 2014). For instance, Maruthy and Raj (2014) investigated the performance of 10 naïve and 10 expert listeners in evaluating speakers with hypokinetic dysarthria. They found no effect of listener experience on speech intelligibility computed as the percentage of correctly transcribed words, although they did find a significant effect on listener effort ratings. Nevertheless, as pointed out by Mencke et al. (1983), measures collected from naïve listeners tend to show larger variation than those collected from well-trained expert listeners such as speech-language therapists. Therefore, expert listeners such as speech-language therapists were preferred over naïve listeners in the current study. In addition, listeners' familiarity with either speakers or speech materials has been reported to increase intelligibility scores (Hustad & Cahill 2003; Liss et al., 2002; Tjaden & Liss, 1995a, 1995b), and thus, listening times of utterances to be assessed should be limited to reduce the impact of familiarity.

In order to better understand the performance of different speech intelligibility measures, as well as their interrater reliability and validity, we conducted a study that addressed (a) two types of intelligibility measures, i.e., one rating-based measure through VAS and one transcription-based measure through orthographic transcriptions, (b) two types of speech materials, i.e., meaningful sentences selected from a phonetically-balanced narrative and word lists consisting of unconnected, monosyllable pseudowords and existing words, and (c) speakers with different severity levels of dysarthria. Moreover, for transcription-based measures, we adopted two forms of transcription. One, called Existing-Word Transcription (EWTrans), allows only existing, meaningful words and has been commonly applied in previous studies. The other one, called All-Word Transcription (AWTrans), allows all sorts of words, including pseudowords. One of the reasons for applying AWTrans was that AWTrans was the only reasonable choice for one of our speech materials (word lists) due to the pseudowords it contained. In this way, we were able to compare intelligibility measures between the two types of speech materials. Another

reason was that as AWTrans has not been investigated on meaningful sentences, we applied it together with EWTrans to investigate (1) whether AWTrans can generate reliable measures for meaningful sentences, and (2) whether the newly-proposed form (AWTrans) differs from the commonly-applied form (EWTrans). In addition, for assessing interrater reliability, we applied G Theory to (1) analyse the effects of utterances, listeners, and speakers in one overall analysis, and (2) evaluate the number of listeners and utterances needed to obtain reliable measures. By doing so, we focused on providing interesting insights and guidance for reliability analyses in research of speech and language pathology. Our study addressed the following three research questions:

- RQ1: To what extent are intelligibility measures reliable?
- RQ2: How many listeners and utterance samples per speaker are needed to obtain reliable intelligibility measures?
- RQ3: To what extent are intelligibility measures valid?

3.2 Method

In this study, we investigated two types of speech intelligibility measures, a rating-based measure through VAS and a transcription-based measure through two forms of orthographic transcriptions. Two separate listening experiments, the Sentence Experiment and the Word Experiment, were designed with each involving one specific type of speech material, meaningful sentences selected from a narrative and word lists containing existing words and pseudowords. The speech materials and the speakers were selected from the Corpus of Pathological and Normal Speech (COPAS) database[2] (Middag, 2012). This database contains recordings from a large number of speakers of Belgian Dutch (the variety of Dutch spoken in Flanders, the northern part of Belgium) with and without speech disorders, with reading materials (isolated words, isolated sentences, and short passages) and spontaneous speech. The two listening experiments were conducted within the research project Developing valid measurement procedure of pathological speech intelligibility (application 2019–3197) that has been approved by the Ethics Assessment Committee Humanities of the Faculty of Arts and the Faculty of

[2] More information and the manual of COPAS can be found on https://taalmaterialen.ivdnt.org/download/tstc-corpus-pathologische-en-normale-spraak-copas/.

Philosophy, Theology and Religious Studies at the Radboud University with reference number Let/MvB19U.514400.

3.2.1 Speech material

For the Sentence Experiment, we selected four meaningful sentences from the Dutch commonly-used phonetically balanced narrative 'Papa en Marloes' (PM, 'Papa and Marloes' in English; Van Lierde et al., 1991):

1. 'Papa en Marloes staan op het station. ' (PM1, in English 'Papa and Marloes are at the station.'),
2. 'Marloes kijkt naar links.' (PM2, in English 'Marloes looks to the left.'),
3. 'In de verte ziet ze de trein al aankomen. ' (PM3, in English 'In the distance she can see the train coming.'),
4. 'Het is al vijf over drie dus het duurt nog vier minuten.' (PM4, in English 'It is already five past three so it will take another four minutes.').

These sentences vary in length and contain the corner vowels, i.e., /a:/, /u/, and /i/, which may be relevant for future acoustical analyses. Accordingly, for each speaker, four recordings were made, each of which being a reading of one of the sentences.

For the Word Experiment, we selected word lists from those constructed in the Dutch Intelligibility Assessment (DIA) task (De Bodt et al., 2006), which was designed to assess intelligibility at the phoneme level called Phoneme Intelligibility (PhonI). Unlike the Sentence Experiment, in which speakers read the same four sentences, speakers in the Word Experiment received three word lists, each of which was a variant of those for each of three subsets, i.e., A, B, and C, constructed in the DIA task. These three subsets are designed to assess initial consonants, final consonants, and medial vowels of Consonant-Vowel-Consonant (CVC) words, including both existing words and pseudowords, respectively. Accordingly, for each speaker, three recordings were made, each of which being a reading of a word list (a variant of a subset).

3.2.2 Speakers

The COPAS database contains recordings from 197 dysarthric speakers and 122 healthy speakers. The recordings covered different speech materials. However,

since we were interested in the four meaningful sentences and word lists of the DIA task, as described above, we focused on speakers (49 dysarthric and 83 healthy speakers) whose recordings of these two speech materials were available. In order to ensure the diversity of speaker data, we carefully selected 26 dysarthric speakers based on their identical proportions among 49 dysarthric speakers in terms of dysarthria type, severity levels of dysarthria (mild–moderate–severe), PhonI scores obtained through the original DIA task, age, and gender. Based on the same selection principle, we selected 10 healthy speakers out of 83 healthy speakers as a non-dysarthric group. The number of non-dysarthric speakers was smaller than that of dysarthric speakers because we focused on dysarthric speakers, but we also maintained the possibility of comparing dysarthric and non-dysarthric speakers. In total, we selected 36 speakers for the Sentence Experiment, and half of them for the Word Experiment (the reason is described in Section 3.2.4). Table 3.1 presents the information about the selected 36 speakers regarding dysarthria type, etiology, the severity level of dysarthria, PhonI scores, which were extracted from the COPAS dataset, and whether a speaker was involved in the Word Experiment. We set the severity levels of dysarthria (SevL) at four levels (non–mild–moderate–severe). Figure 3.1 shows the distribution of the speakers over the four different levels of SevL.

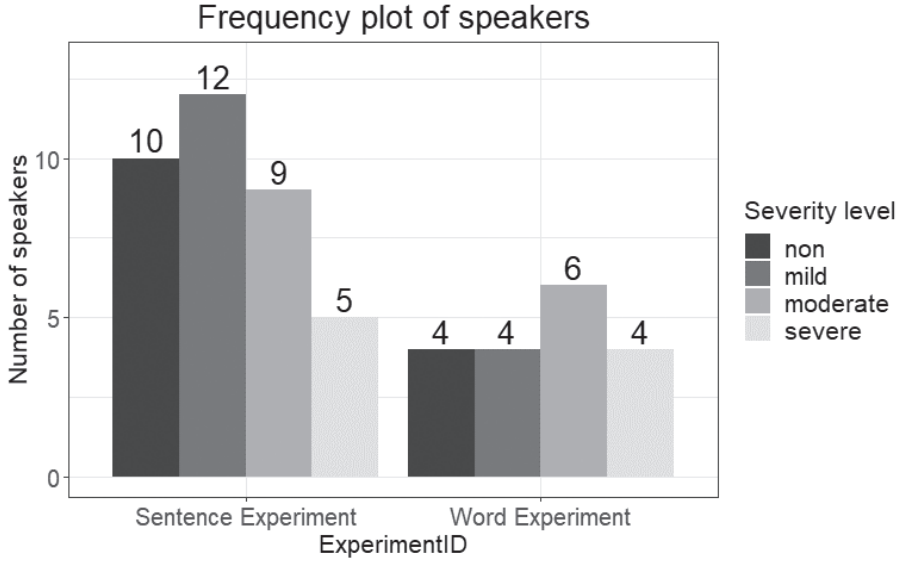


Figure 3.1. Distribution of speakers over four severity levels of dysarthria (SevL) in our two experiments.

Both the SevL and PhonI had been assigned by experienced speech-language pathologists at the time the COPAS database was compiled. PhonI, calculated as the percentage of correctly transcribed target phonemes over all three word lists (each is a variant of a subset) for each speaker, is a highly reliable measure, with an inter-rater correlation of 0.91 and an intra-rater correlation of 0.93 using ICC (De Bodt et al., 2006; Middag et al., 2009a; Van Nuffelen et al., 2008). SevL has also been used in many international publications (e.g., Middag, 2012; Yilmaz et al., 2016b; Van Nuffelen et al., 2009a) as a basis for selecting speech recordings for experiments. Therefore, both of them can be used for selecting speakers and for evaluating our measures in the current study (see details in Section 3.2.6).

Table 3.1. Detailed information for 36 speakers including Speaker ID, gender, age, etiology, dysarthria type, PhonI score (%), the severity level of dysarthria (SevL; non-mild-moderate-severe), and whether the speakers were selected for the Word Experiment.

Speaker ID	Gender	Age	Etiology	Dysarthria	PhonI	SevL	Word Experiment
D3	M	66	PD	hypokinetic	94	mild	No
D9	M	19	stroke	Mixed	100	mild	No
D12	M	23	myotonic dystrophy	Flaccid	80	mild	No
D13	F	38	mitochondrial disorder	-	92	moderate	Yes
D14	M	85	PD	hypokinetic	76	mild	No
D15	F	76	stroke	Flaccid	72	moderate	No
D17	M	51	PD	hypokinetic	76	severe	No
D19	M	80	cerebellar syndrome	-	84	mild	No
D20	F	51	stroke	UUMND	90	mild	No
D21	M	18	congenital	Flaccid	50	moderate	Yes
D22	F	8	congenital	hyperkinetic	78	moderate	No
D24	M	12	congenital	Mixed	76	severe	Yes
D25	M	16	congenital	Spastic	88	mild	Yes
D26	F	12	congenital	hypokinetic	36	severe	Yes
D27	F	16	congenital	Spastic	80	moderate	Yes

Table 3.1. (Continued)

Speaker ID	Gender	Age	Etiology	Dysarthria	PhonI	SevL	Word Experiment
D29	F	15	congenital	hypokinetic	70	severe	Yes
D30	F	81	peripheral nerve damage	Flaccid	84	mild	No
D46	M	53	MS	Flaccid	76	moderate	Yes
D49	F	43	stroke	–	68	moderate	Yes
D60	M	59	stroke	hyperkinetic	66	mild	Yes
D67	M	51	MS	–	90	mild	No
D68	M	33	MS	–	76	moderate	Yes
D71	M	48	PD	hypokinetic	100	severe	Yes
D79	M	81	PD	hypokinetic	52	mild	Yes
D81	M	33	MS	–	88	moderate	No
D83	M	51	MS	–	86	mild	Yes
N23	F	21	–	–	88	non	No
N28	M	50	–	–	84	non	No
N30	F	50	–	–	84	non	No
N56	M	23	–	–	90	non	Yes
N62	M	41	–	–	98	non	No
N71	F	77	–	–	82	non	Yes
N72	F	67	–	–	96	non	No
N114	F	35	–	–	96	non	Yes
N117	M	58	–	–	100	non	Yes
N130	M	23	–	–	94	non	No

Note. **PD**: Parkinson's disease. **MS**: multiple sclerosis. **UUMND**: Unilateral upper motor neuron disorder. **M** = male; **F** = female; '**–**' for dysarthric speakers: information not included in the COPAS.

3.2.3 Expert listeners

As mentioned earlier, measures collected from naïve listeners tend to show larger variation than those collected from well-trained expert listeners (Mencke et al., 1983). We selected speech-language therapists as expert listeners rather than naïve listeners. A previous study (Van Nuffelen et al.,

2010) reported reliable measures with three experts. To ensure reliability, we recruited five Belgian Dutch-speaking speech-language pathologists (one male and four females) from the University Antwerp Hospital. They were working at the ear, nose, and throat (ENT) revalidation center for communication disorders, and were all familiar with evaluating and testing dysarthric patients through listening experiments.

3.2.4 Experimental procedure

All recordings included in our listening experiments were made in a quiet clinical setting without a sound-attenuated box, as described in the COPAS manual. Originally, two microphones were used, one on the table with a mouth-microphone distance of about 30 cm, and one headset. Recordings of the selected speakers were made through the microphone on the table, with the exception of one speaker (in the Sentence Experiment only), for whom the used microphone was not known. We evaluated this speaker's recordings and found the sound quality to be similar to those of the other recordings.

Both experiments were set up through the online survey tool Qualtrics and were conducted on the same day, the Sentence Experiment in the morning and the Word Experiment in the afternoon, with two resting hours in between. Before starting, the listeners received consent forms and descriptions of both experiments on the Qualtrics website. They gave their explicit consent by clicking on the 'agree' button. In each of the two experiments, they first received instructions in Belgian Dutch. Following this, they received three and two practice examples to familiarize themselves with the procedure in the Sentence Experiment and the Word Experiment, respectively. For each experiment, the same two anchor items selected from healthy and severely dysarthric speakers in the COPAS dataset were repeated after every ten utterances in a pop-up format to remind the listeners of what high and low intelligibility sound like. We ensured the recordings used as examples and as anchor items were not from the speakers involved in the two actual experiments. Moreover, the utterance samples (recordings) were randomized to prevent any systematic order effect. We prevented every six consecutive samples from being selected from the same speakers and every two consecutive samples from being about the same sentences or subsets. All the listeners received the same randomized order of samples.

Specifically, for the Sentence Experiment, the listeners assessed each of the 144 utterances (recordings), consisting of the same set of four sentences read by 36 speakers, by performing two kinds of tasks, i.e., making orthographic transcriptions and filling in a VAS scale ranging from 0 (not intelligible) to 100 (intelligible). In detail, the VAS contained tick marks with numbers for every ten scores shown (e.g., 10, 20, 30, etc.), no scale endpoints' labels, and was oriented horizontally with written instructions '*Wat voor score zou u toekennen aan de spraakverstaanbaarheid?*' (in English 'what score would you assign for speech intelligibility?'). Regarding the orthographic transcription task, two forms of transcriptions, i.e., EWTrans and AWTrans, were made by the listeners. The new form of transcription, AWTrans, can provide the listeners with more flexibility in their transcriptions, thus helping report speech deviations at the segmental level. In addition, each utterance together with both tasks was presented on the same page in the order of AWTrans, VAS, and EWTrans, as illustrated in Figure 3.2. To prevent the listeners from adapting to the speakers and the speech materials, the listening time for each utterance was limited. They were allowed to listen to each utterance twice since they had to complete two forms of transcriptions. Further, it was up to the listeners to decide which form of transcription to complete first and when (after or between completing the two forms of transcriptions) to assign a score on VAS. The total time required for completing the Sentence Experiment was around one hour, and the listeners were encouraged to take a break after half an hour to prevent them from being fatigued.

For the Word Experiment, the listeners assessed each of 54 utterances (recordings), consisting of three word lists (three variants of the three DIA subsets) read by 18 speakers, by making AWTrans and filling in a VAS scale. The three word lists (variants) of the subsets for each speaker in the experiment were presented in three recordings, in each of which three seconds of silence had been inserted manually between successive words to ensure that the listeners had enough time to transcribe each word. Unlike the original DIA procedure (Middag et al., 2009b), in which the listeners were asked to transcribe the missing target phonemes while the remaining phonemes of a word were presented (e.g., transcribing target phoneme 'n' in 'nit', which was presented as '.it'), the listeners in our experiment had to transcribe the whole words in the word lists without any phonemes being presented.

English translation for instructions:

Utterance 2:

Listen 1x to the following speech

1)Write down literally what you hear. You may use not only existing words but also nonsense words. [This is AWTrans]

2) What score would you assign for speech intelligibility?

Listen 1x to the following speech

3) Write down what you hear. You may only use existing words. [This is EWTrans]

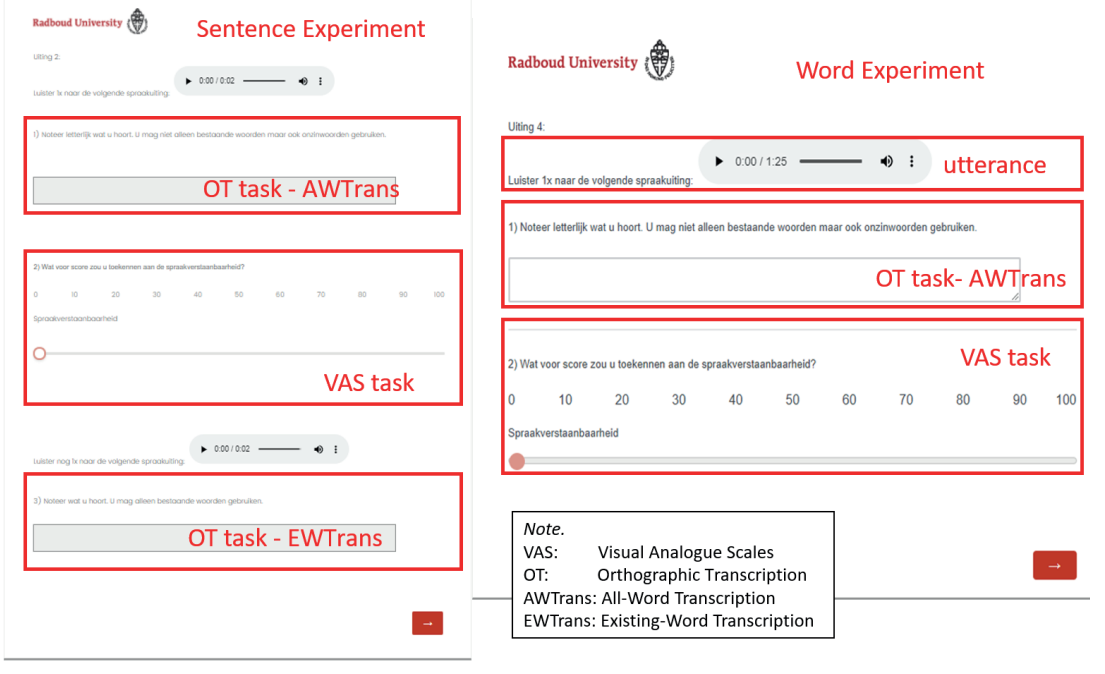


Figure 3.2. An illustration of our online listening experiments. **Best viewed in color.**

The three utterances were assessed separately for each speaker. Moreover, only one form of transcription, i.e., AWTrans, was applied. This was because pseudowords were contained in the word lists, so it was not reasonable to transcribe using only existing words, as in EWTrans. Accordingly, all the listeners were allowed to listen only once to each utterance to complete the AWTrans first and then assign a score on VAS. Also, considering that the required time for completing the two tasks for each utterance was much longer than that of the Sentence Experiment, and to ensure that this experiment could also be completed within one hour, only half of the speakers of the Sentence Experiment were involved in the Word Experiment.

3.2.5 Intelligibility measures

For each utterance, we obtained scores from the VAS and the orthographic transcription tasks. For the orthographic transcriptions, we calculated the Accuracy of Words (AcW) as follows[3]:

$$AcW = \frac{N_{match} - N_{insertion}}{N_{total}} \times 100$$

where N_{total} denotes the total number of words in the reference transcriptions, N_{match} denotes the number of matched words between the orthographic and the reference transcriptions, and $N_{insertion}$ denotes the number of insertions in the orthographic transcriptions. Note that we removed punctuation and symbols indicating missing words to obtain cleaned transcriptions for the calculation. We also removed pseudowords in EWTrans before calculating AcW. In addition, no errors, such as misspellings, homophones or incorrect tense markers, were permitted. This was because the listeners recruited in the present study were experienced, well-trained experts, and thus, we believed that they transcribed what they thought they had heard.

3.2.6 Data analysis

To address the first two research questions regarding the interrater reliability of the VAS and AcW scores, we applied G Theory by using the *gtheory* package

[3] We used the asr-evaluation python module provided in https://github.com/belambert/asr-evaluation. The calculation was done fully automatically.

(Moore, 2016) in RStudio (Rstudio Team, 2020) with R version 4.0.1 (R Core Team, 2020). The advantage of applying G Theory is that all sources of variance relevant in the experiments can be dealt with simultaneously, e.g., listeners, speakers, and utterances. In G Theory, the reliability coefficient is defined as the proportion of score variance attributable to the different sources in relation to the total variance. This model of analysis produces two reliability coefficients, i.e., the Generalizability coefficient (G-coefficient) and the Phi coefficient (D-coefficient). A G-coefficient should be calculated when one is interested in making decisions about an individual's performance relative to that of his or her peers. The more demanding or strict D-coefficient should be calculated when one is interested in an individual's performance irrespective of that of his or her peers and is therefore most likely to be used when making criterion-referenced screening or progress monitoring decisions. We chose the D-coefficient as reliability coefficient and to evaluate the number of listeners and utterances needed to achieve an acceptable reliability level. The model designs for the two experiments were different because of the utterance samples. In the Sentence Experiment, the model was fully crossed since all five listeners assessed all four utterances from all 36 speakers: *Listener*×*Utterance*×*Speaker*. However, in the Word Experiment, each speaker received a random variant of each subset A, B, and C, resulting in three utterances per speaker. We considered the subsets to be replications of sets of CVC words. The actual word list (*Utterance*) was nested under *Speaker*. Such a design can be summarized as: (*Utterance*:*Speaker*)×*Listener*[4], meaning that all combinations of subsets, i.e., utterances and speakers were rated by all listeners. In addition, *Utterance*, *Listener*, and *Speaker* were defined as random factors in both experiments since they are all potential samples from their universe. Also, to gain more insight into the reliability of AcW and VAS in the Sentence Experiment, we computed the reliability of the measures on the four utterances separately. Since each utterance was rated by all listeners, giving two sources of variance, ICC could be used as a reliability measure. We applied the ICC by using the R *psych* package (Revelle, 2019).

[4] We also explored the model including the subsets of DIA as a fourth factor which led to the model design being (*Utterance*:(*Speaker*×*Subset*))×*Listener*. However, we did not find a main effect of Subset on the reliability of the measures. Therefore, we simplified the model design as it is now in the current study, which is (*Utterance*:*Speaker*)×*Listener*.

Moreover, G Theory allows calculating the consequences of modifying the size of a factor, such as the number of listeners, in measuring reliability. Consequently, it is possible to calculate the minimum size of a factor required to obtain reliable data (Li et al., 2015; Shavelson & Webb, 2006). By taking this strength of G Theory, we were able to address the second research question regarding the optimal numbers of listeners and utterances per speaker. Specifically, following the common practice (Brennan, 2001; Briesch et al., 2014; Hollo et al., 2020; O'Brian et al., 2003; Webb et al., 2006), we first calculated sources of measurement error based on the collected data and then used these to estimate the reliability (D-coefficient) for different numbers of listeners and utterances per speaker. According to the estimations, we can infer the minimum number of listeners and the minimum number of utterances per speaker required for a reliable measure. A detailed explanation of G Theory can be found in Brennan (2001); Briesch et al. (2014) is a practical guide in implementing the analyses.

To address the third research question regarding validity, we investigated construct validity and concurrent validity for VAS and AcW, averaging the scores of each speaker. Construct validity investigates whether the test measures the concept it intends to measure and was analysed through Pearson correlations between VAS and AcW within each experiment and between experiments for each measure (VAS/AcW). Concurrent validity is a type of criterion validity and measures how well a test compares to other criteria. We correlated our measures to the two external measures that were available: SevL and PhonI. We applied multinomial regression to investigate the correlations of SevL, as this variable defines four severity groups, without the claim of being a continuous variable. Then based on the predicted labels (severity levels of dysarthria) generated by multinomial regression analysis, we calculated the percentage of speakers correctly classified in these four groups. To interpret the validity results, we used the guidelines from Evans (1996, p. 146): a correlation between 0.80 and 1.0 is 'very strong', between 0.60 and 0.79 is 'strong', between 0.40 and 0.59 is 'moderate', between 0.20 and 0.39 is 'weak', and that even lower is 'very weak'. We used the *stats* (R Core Team, 2020), *nnet* (Venables & Ripley, 2002), and *DescTools* (Signorell, et al., 2021) packages in R version 4.0.1 for the implementation and the *ggplot2* package (Wickham, 2016) for making plots.

3.3 Results

3.3.1 General results of the intelligibility measures

The means and standard deviations of the VAS and AcW scores in both experiments are shown in Table 3.2. For the Sentence Experiment, higher mean values were observed for EWTrans than for AWTrans. Compared with the Sentence Experiment, the Word Experiment showed lower scores, especially on AcW. This might be due to the absence of contextual cues in the word lists compared to the meaningful sentences.

Table 3.2. Means (standard deviations) for the two types of intelligibility measures, VAS and AcW, in our two experiments. The results for AcW in the Sentence Experiment are denoted by All-Word Transcription (AWTrans) on the left and Existing-Word Transcription (EWTrans) on the right.

Measures	Sentence Experiment	Word Experiment
VAS	84.25 (25.52)	68.12 (25.29)
AcW (%)	79.63 (29.86) / 87.86 (25.19)	41.07 (25.08)

Note. **VAS**: Visual Analogue Scale. **AcW**: Accuracy of Words.

3.3.2 Interrater reliability of the intelligibility measures

We computed the interrater reliability of the VAS and AcW scores based on the D-coefficient. Table 3.3 shows that the reliability values were high (above 0.90) for VAS in both experiments and for AcW in the Word Experiment. The values for AcW were slightly lower in the Sentence Experiment with the one of EWTrans being the lowest.

Table 3.3. Interrater reliability of the intelligibility measures in our two experiments based on the D-coefficient. The results for AcW in the Sentence Experiment are denoted by All-Word Transcription (AWTrans) on the left and Existing-Word Transcription (EWTrans) on the right.

Measures	Sentence Experiment	Word Experiment
VAS	0.93	0.91
AcW	0.89 / 0.83	0.95

Note. **VAS**: Visual Analogue Scale. **AcW**: Accuracy of Words.

3.3.3 Interrater reliability of the intelligibility measures per utterance in the Sentence Experiment

As shown in Table 3.4, low reliability values were observed for the PM2 sentence, which is 'Marloes kijkt naar links', especially for EWTrans. After analysing the transcriptions, we noticed four problems with this sentence. First, the first word in this sentence is not a very common proper name. This name could be modified in various ways which caused the increase in the number of incorrectly transcribed words. Second, the second and third words are 'kijkt' and 'naar'. Dutch native speakers realize only one release burst in the cluster 'kt' and even no release at all when the nasal 'n' follows. This reduces the distinction with 'keek naar' (past tense), taking into account the regional variation in pronouncing diphthongs and tensed vowels in Dutch (cf. Adank et al., 2007). Third, the third word is 'naar'. Many Dutch native speakers do not distinguish spatial 'naar' (i.e., 'to' as in English) and temporal 'na' (i.e., 'after' as in English) in their spontaneous speech. Finally, the fourth word is 'links', which is often pronounced without a plosive burst related to the 'k' ('lings' in AWTrans). The cumulation of four pronunciation variants made the listeners' transcriptions of this sentence less reliable than the other three, particularly in EWTrans. This should be avoided when constructing sentences to calculate the accuracy of transcribed words for intelligibility measures.

Table 3.4. Interrater reliability of the intelligibility measures in the Sentence Experiment based on the Intraclass Correlation Coefficient for the four sentences, i.e., PM1, PM2, PM3, and PM4. The results for AcW in the Sentence Experiment are denoted by All-Word Transcription (AWTrans) on the left and Existing-Word Transcription (EWTrans) on the right.

Measures	PM1	PM2	PM3	PM4
VAS	0.89	0.92	0.95	0.97
AcW	0.91 / 0.90	0.84 / 0.65	0.96 / 0.93	0.96 / 0.94

Note. **VAS**: Visual Analogue Scale. **AcW**: Accuracy of Words.

3.3.4 Reliability as a function of the number of listeners and the number of utterance samples

Figure 3.3 shows that in both experiments, the D-coefficient increased when the number of listeners or the number of utterance samples increased. When the number of utterance samples was fixed, the reliability increased rapidly at

first and then this increase began to plateau. As suggested by Wells and Wollack (2003), professionally developed high-stake tests should have a reliability of at least 0.90. We can observe that VAS reached this reliability level with three listeners and three samples in the Sentence Experiment, and with four listeners and three samples in the Word Experiment. The results for VAS were comparable, but those for AcW were not. Specifically, for AcW in the Sentence Experiment, the scores using AWTrans reached the reliability level with seven listeners and four utterances, but those using EWTrans were the lowest and remained below this level for all cases of listeners and utterances. This might also be due to the problematic sentence PM2. For AcW in the Word Experiment, the reliability level can be reached with only two listeners and two utterances.

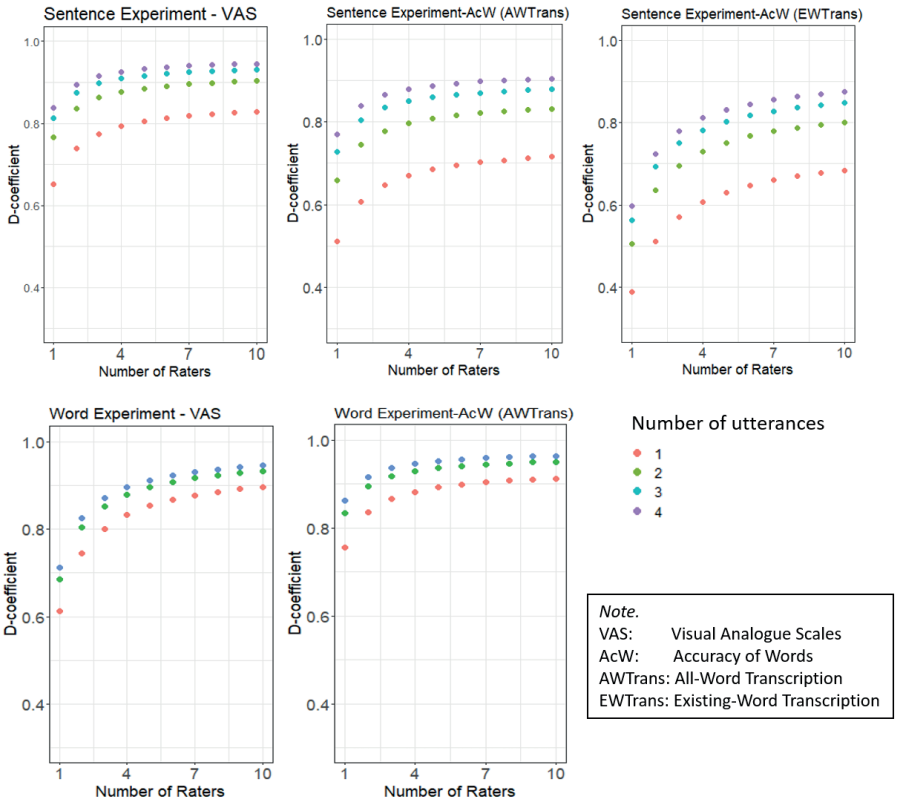


Figure 3.3. D-coefficients for a different number of utterances and listeners of scores of VAS and AcW (with two forms of transcriptions). **Best viewed in color.**

3.3.5 Construct validity

Table 3.5 shows that all correlations between VAS and AcW in the same experiment were very strong (above 0.94). In contrast, the correlations of VAS and AcW between the two experiments were slightly weaker, but were still strong, with 0.88 for VAS and 0.81 for AcW.

Table 3.5. Pearson correlations between VAS and AcW in the two experiments, with two forms of transcriptions in the Sentence Experiment.

Correlations	Sentence Experiment (N=36)		Word Experiment (N=18)	
	AcW AWTrans	AcW EWTrans		AcW AWTrans
VAS	.98	.95	**VAS**	.94
AcW AWTrans	--	.95		

Note. **VAS**: Visual Analogue Scale. **AcW**: Accuracy of Words. **AWTrans**: All-Word Transcription. **EWTrans**: Existing-Word Transcription.

3.3.6 Concurrent validity

We investigated the concurrent validity of our measures with two external measures, i.e., SevL and PhonI. We first computed the correlations between these two external measures. Specifically, the multinomial regression with SevL as criterion and PhonI as predictor gave a Nagelkerke R^2 (as the correlation) of 0.296 in the Sentence Experiment and 0.299 in the Word Experiment. The percentages of correctly classified speakers in the four levels of SevL were 44.4% in the Sentence Experiment and 50.0% in the Word Experiment. These outcomes suggest that SevL and PhonI reflect different constructs of intelligibility.

Table 3.6. Pearson correlations between the external measure PhonI and our two intelligibility measures, i.e., VAS and AcW. The results for AcW in the Sentence Experiment are denoted by All-Word Transcription (AWTrans) on the left and Existing-Word Transcription (EWTrans) on the right.

Correlations	Sentence Experiment (N=36)	Word Experiment (N=18)
VAS	0.65	0.61
AcW	0.67 / 0.60	0.70

Note. **VAS**: Visual Analogue Scale. **AcW**: Accuracy of Words.

The correlations between PhonI and VAS/AcW were strong, as shown in Table 3.6. The correlations of AcW using AWTrans were slightly stronger than those of VAS for both experiments. Figure 3.4 shows five scattergrams, including the regression lines and their confidence intervals (95%) with distinguished SevL levels of the speakers. In the Sentence Experiment, the points were concentrated at the top-right corner, but in the Word Experiment, they were scattered across the scale. This might be due to the differences in speech material. These scattergrams also show that intelligibility measured at the phoneme level, i.e., PhonI, was different from intelligibility at higher levels, i.e., VAS at the utterance level and AcW at the word level.

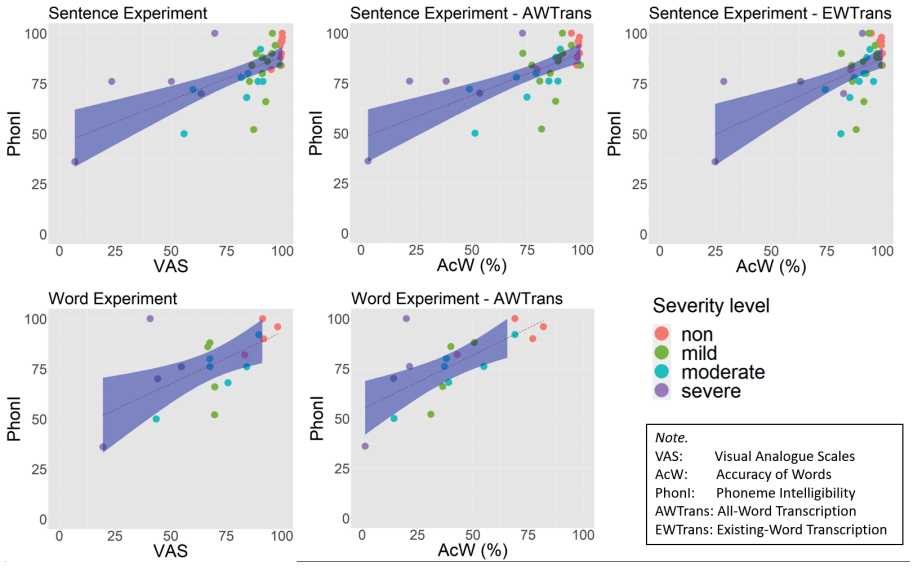


Figure 3.4. Scattergrams, including the regression lines and their confidence intervals (95%) with distinguished SevL of the speakers, between the external measure PhonI and our two intelligibility measures, VAS and AcW (with two forms of transcriptions). **Best viewed in color.**

Table 3.7 shows the Nagelkerke R^2s and the percentage of correctly classified speakers by using multinomial regression with SevL as criterion and VAS/AcW as predictor. The results showed that most of the time VAS was better than AcW. Compared with EWTrans, AWTrans provided better results for AcW.

Table 3.7. Correlation (Nagelkerke R^2s) and SevL classification (percentage of correctly classified speakers) for the external measure SevL as the criterion and one of our two intelligibility measures, VAS and AcW, as a predictor. The results for AcW in the Sentence Experiment are denoted by All-Word Transcription (AWTrans) on the left and Existing-Word Transcription (EWTrans) on the right.

Correlation (SevL classification)	Sentence Experiment (N=36)	Word Experiment (N=18)
VAS	0.84 (63.9 %)	0.76 (55.6 %)
AcW	0.70 (61.1%) / 0.59 (52.8%)	0.70 (61.1%)

Note. **VAS**: Visual Analogue Scale. **AcW**: Accuracy of Words. **SevL**: severity level of dysarthria.

Figure 3.5 shows that VAS and AcW overlapped between levels of SevL, but the tendency was that the more severe levels corresponded to lower VAS/AcW scores, whereas the less severe levels corresponded to higher VAS/AcW scores. We can also see, in line with the results in Table 3.7, that VAS better discriminated the levels of SevL than AcW does, with AWTrans being better than EWTrans.

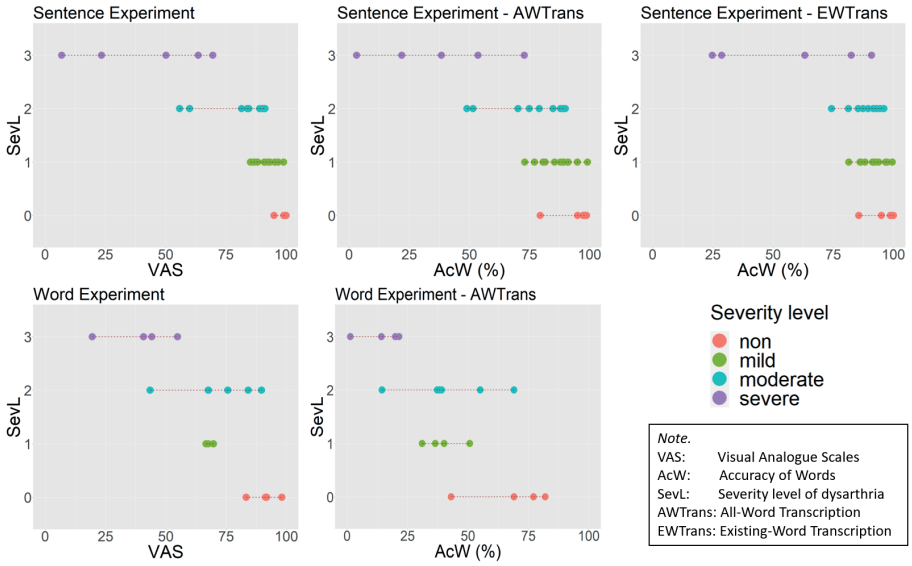


Figure 3.5. Scattergrams between the external measure SevL and our two intelligibility measures, VAS and AcW (with two forms of transcriptions). **Best viewed in color.**

3.4 Discussion

In this study, we investigated the interrater reliability and validity of two types of speech intelligibility measures, one rating-based measure through Visual Analogue Scales (VAS) and one transcription-based measure, i.e., Accuracy of Words (AcW), through orthographic transcriptions, by conducting two listening experiments, one targeting meaningful sentences (Sentence Experiment) and one targeting word lists (Word Experiment). For AcW in the Sentence Experiment, we studied two forms of transcriptions, i.e., Existing-Word Transcription (EWTrans) allowing meaningful, existing words, and All-Word Transcription (AWTrans) allowing all sorts of words, including both existing words and pseudowords. The mean values of VAS and AcW were generally higher for the meaningful sentences than for the word lists with comparable standard deviations. This is in line with previous findings (Barreto & Ortiz, 2008; Hustad, 2007; Miller, 2013; Yorkston & Beukelman, 1978, 1981) that intelligibility measures result in higher values for meaningful sentences than for word lists due to the presence of contextual information. However, this difference was not found for severe dysarthria in our VAS scores. In addition, we observed higher mean values for VAS than for AcW in AWTrans. This seems to be in conflict with the finding in Stipancic et al. (2016), which showed that the percent correct scores obtained from orthographic transcriptions were higher than the VAS scores. However, the same result can actually be observed in our study when AcW was derived from EWTrans, as was done in Stipancic et al. (2016). This suggests that EWTrans and AWTrans can provide different results, and thus, listeners should be clearly instructed when using one or the other form. In the remainder of this section, we first discuss the results of the present study in relation to the specific research questions we addressed. Following this, we describe the limitations of the present study. After that, we stress our recommendations for designing listening experiments and the focus of future work.

3.4.1 RQ1: To what extent are intelligibility measures reliable?

For VAS, very high interrater reliability values (above 0.90) were observed for both speech materials. However, for AcW, the reliability was higher

for the word lists (0.95) than for the meaningful sentences (below 0.90), especially when using EWTrans (0.83). The relatively lower reliability for EWTrans may be explained by the fact that one of the specific sentences we used (PM2) contained an uncommon name and an unintended cumulation of pronunciation variability. These problems should be avoided when selecting sentences to calculate the accuracy of words for intelligibility measures through orthographic transcriptions. In general, the results suggest that VAS is more reliable than AcW. This finding is consistent with the results of previous studies (Ganzeboom et al., 2016; Stipancic et al., 2016; Tjaden et al., 2014a), in which high interrater reliability values (above 0.90) were also reported for VAS. Reliability was also high for AcW, which is in line with the broad consensus that transcription yields good interrater reliability (Bunton et al., 2001; Miller, 2013; Tjaden et al., 2014a; Tjaden & Wilding, 2010).

Note that we did not measure intrarater reliability due to several reasons. First, intrarater reliability has the disadvantage of requiring repeated measurements. Second, it does not generalize to a measure representing a group of listeners, e.g., experts or human listeners overall. Moreover, high intrarater reliability does not imply high inter-rater reliability.

3.4.2 RQ2: How many listeners and utterance samples per speaker are needed to obtain reliable intelligibility measures?

The interrater reliability analyses showed that the number of listeners and utterance samples per speaker was positively related to the reliability of VAS and AcW. For VAS, regardless of speech materials, at least three samples per speaker in combination with four listeners were needed to obtain reliable results, i.e., passing the criterion of 0.90 for professionally developed high-stake tests (Wells & Wollack, 2003). However, for AcW, different materials yielded different results. Specifically, for the word lists, at least two samples per speaker in combination with two listeners were needed, while for the meaningful sentences, many more listeners with all the samples involved were needed. In detail, at least seven listeners were needed when using AWTrans, while more than ten listeners were needed for EWTrans. In this case, if all individual utterances meet the criteria of a good test item, four listeners in combination with four utterances may be also sufficient. Note that our study used expert listeners. Recruiting naïve listeners as listeners

may lead to different results and the number of listeners needed for high reliability may be much larger than four (Ganzeboom et al., 2016).

3.4.3 RQ3: To what extent are the intelligibility measures valid?

Regarding construct validity, very strong correlations were found between VAS and AcW within the same speech material. This is in line with the finding by Abur et al. (2019), who also found strong, positive relationships (0.886; $p<0.001$) between scores derived from the orthographic transcription and VAS tasks. This suggests that for the same material, VAS and AcW are related to the same construct of intelligibility, even when different transcription forms are used for AcW. These results indicate that both VAS and AcW are valid to a great extent when they are collected for the same material. The correlations of the same measure, i.e., VAS or AcW, between different speech materials were below 0.90, indicating perhaps that different constructs of intelligibility are measured in different speech materials. Specifically, the correlations were 0.88 for VAS and 0.81 for AcW. This suggests that VAS and AcW remain valid to a substantial extent across materials and that VAS might be a more stable or robust indicator of intelligibility.

Regarding concurrent validity, the weak correlations between the two external measures, i.e., the severity level of dysarthria (SevL) and Phoneme Intelligibility (PhonI), indicate that they reflect different constructs of intelligibility. VAS was much more strongly correlated to SevL than to PhonI and showed better discriminations of the levels of SevL than AcW. Differently, the correlations between AcW and the two external measures (SevL/PhonI) were comparably strong. These results seem to suggest that AcW is related to a construct of intelligibility that is similar to those of SevL/PhonI. For both VAS and AcW, we found that more than half of the speakers were classified in the correct levels of SevL. These results applied to both sentences and word lists.

3.4.4 Limitations

The present study investigated the interrater reliability and validity of VAS and AcW on two types of speech materials, i.e., meaningful sentences

from a narrative and word lists containing pseudowords. The meaningful sentences were employed to obtain more ecologically valid intelligibility scores since these sentences are closer to those used in daily conversation in comparison to the word lists. Previous studies criticized these measures arguing that listeners could rely on more contextual information to understand the message in sentences than in words (Dongilli, 1994; Hustad, 2007; Yorkston & Beukelman, 1978, 1981). Thus, to limit the contextual information in sentences, some studies (e.g., Ellis & Fucci, 1991; Ganzeboom et al., 2016) have used Semantically Unpredictable Sentences (SUS) in an attempt to avoid listeners 'guessing' the content. Therefore, the absence of SUS in the present study could be seen as one of the limitations. However, using contextual information to understand a message is what listeners actually do in normal life. Thus, our employment of speech material with contextual information may be very useful to understand how patients would fare under more realistic conditions. Another limitation might be that only expert listeners were involved in our study, and thus, we cannot compare their performance to that of naïve listeners as reported in many studies (Abur et al., 2019; Ganzeboom et al., 2016; Ishikawa et al., 2020; Stipancic et al., 2016; Sussman & Tjaden, 2012). Moreover, we did not permit any errors in the transcriptions, including misspellings and homophones, based on the assumption that the listeners transcribed what they thought they had heard as they were well-trained. However, such an assumption may be too ideal and the subsequent processing of transcriptions may be rigorous because in practice people can make such errors in transcriptions. Therefore, although the measures showed very high reliability values, more research is necessary to refine and further elaborate our novel findings.

In addition, some settings of the experiments in our study were different from those used in cited studies and, thus, might influence the results. For example, the VAS in the current study was presented with an anchor every 10 points, which differs from the typical presentations, with only a beginning (0) and end (100 or 1) anchor (Abur et al., 2019; Ganzeboom et al., 2016; Stipancic et al., 2016). Another example is the different diagnoses and severity levels of dysarthria of speakers involved in the intelligibility assessment. Also, no errors were permitted in calculating AcW since the listeners in the present study were experienced, well-trained

experts. This is stricter than studies in which naïve listeners were recruited (Ishikawa et al., 2018; Stipancic et al., 2016).

Furthermore, as we used the existing dataset to design the experiments, our setting options were limited by the restrictions of this dataset. For instance, the number of listeners for SevL and PhonI measures in the dataset was not clear although the reliability of these measures seems to be sufficient for selecting speakers and evaluating our intelligibility measures. Also, for some of the selected speakers, the etiology of dysarthria was indicated, but not the dysarthria type. Another example is that to collect the measures in meaningful sentences, we were only able to use the exact same four sentences for all the speakers. Although we randomized the meaningful sentences and speakers to avoid listeners' adaptation to them, repeating the sentences may still impact the results. Thus, future studies may use comparable but different meaningful sentences for different speakers rather than the same sentences to further elaborate our findings.

3.4.5 Recommendations

The results presented in this study show that VAS is as reliable and valid as AcW. This indicates that VAS could be a good alternative for research and clinical practice, as also suggested by Ishikawa et al. (2020), especially if we consider that VAS also appears to be more robust due to the small difference in reliability between different speech materials. However, the rating task might not provide enough information for in-depth diagnosis and detailed analysis in research and clinical practice. In turn, this suggests that deciding which measure to apply depends directly on the specific goals of research and clinical practice.

The analyses of the two forms of transcriptions suggest that AWTrans seems to provide more reliable and valid AcW scores than EWTrans, anyway in the case that the listeners are familiar with the speech materials. This shows that the contextual information in meaningful sentences might be limited to a certain extent by using AWTrans. It is of note that listeners should be clearly instructed when using AWTrans.

The analyses of each utterance in the Sentence Experiment indicate, as we have already discussed above, that an uncommon name and an unintended cumulation of pronunciation variability should be avoided when

selecting sentences to calculate AcW for intelligibility through orthographic transcriptions.

Last but not least, it is important that we were able to systematically handle the issue of reliability in terms of listeners and utterances by using Generalizability Theory. In this way, the optimal numbers of utterance samples per speaker and listeners can be determined. This is very helpful for researchers and clinicians in designing listening experiments on speech intelligibility. Having four expert listeners in combination with three samples per speaker is sufficient for obtaining reliable VAS scores regardless of speech materials. Having two expert listeners in combination with two samples per speaker is sufficient for obtaining reliable AcW scores on the word lists, while many more listeners and samples are required for meaningful sentences.

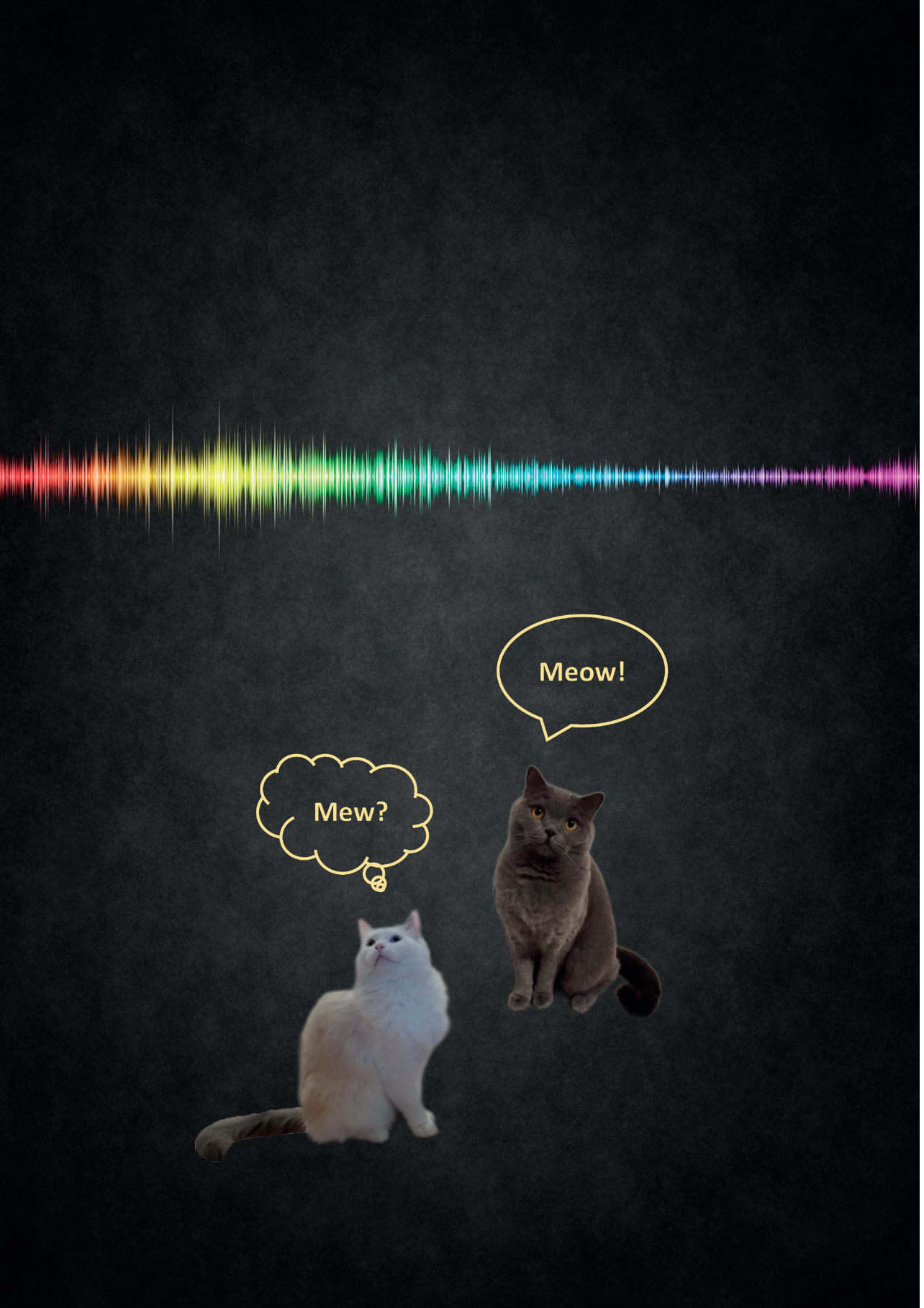
3.4.6 Future work

The finding that EWTrans and AWTrans, as the two forms of transcriptions, led to different reliability measures is a good reason for further investigation at a more fine-grained granularity level, i.e., subword level, also because many studies (Kent et al., 1989; Hustad, 2006; De Bodt et al., 2006; Xue et al., 2020) focusing on the subword level have shown its potential value for both research and clinical practice. In addition, since it was required to prevent the listeners from being fatigued, we evaluated fewer speakers in one experiment than in the other. Future work can address such restrictions by using more complex designs. For example, speakers can be split into multiple groups, and each of the groups can be evaluated by a different group of listeners (Hubers et al., 2019). This takes advantage of one of the strengths of G Theory to handle diverse designs including both crossed and nested factors.

3.5 Conclusions

The present study investigated the interrater reliability and validity of two types of speech intelligibility measures, one rating-based measure, VAS, and one transcription-based measure, AcW, for two different speech materials. With five expert listeners, VAS is as reliable and valid as the commonly used AcW regardless of speech materials and forms of transcriptions.

Our reliability analysis of intelligibility measures by five expert listeners on speech from speakers with different severity levels of dysarthria with respect to two different speech materials leads us to recommend that future studies on intelligibility measures use the D-coefficient, which is part of Generalizability Theory, as a measure of reliability. The D-coefficient can be used for all kinds of experimental designs, and it is allowed to be generalized across listeners and/or samples. This metric also allows assessing the minimum numbers of listeners and samples per speaker that are required to obtain reliable data.


Meow!
Mew?

CHAPTER 4

ASSESSING SPEECH INTELLIGIBILITY OF PATHOLOGICAL SPEECH IN SENTENCES AND WORD LISTS: THE CONTRIBUTION OF PHONEME-LEVEL MEASURES

ABSTRACT

Speech intelligibility is an important indicator of the degree of speech impairment for pathological speech. Articulation, as a key feature of dysarthria, has been found to be a stronger contributor to the intelligibility of dysarthric speech compared to voice quality, nasality, and prosody. In fact, therapy addressing articulation is often used by speech-language pathologists. Since phoneme-level measures are more directly related to articulation, they may contribute to better evaluating articulation imprecision in speakers with dysarthria and to monitoring the effectiveness of therapy. In this chapter, we collected two types of phoneme-level measures: a) Accuracy of Phonemes and b) Phonetic Distance, from orthographic transcriptions obtained from expert listeners in two types of speech materials (i.e., meaningful sentences and word lists). We first examined the measures' interrater reliability using Generalizability Theory. Then we studied the validity of the measures by correlating them to three criterion variables. Following this, we explored their ability in distinguishing speakers in two classification tasks according to speakers' types (i.e., healthy vs dysarthric) and their severity levels of dysarthria, respectively. The results showed that both types of phoneme-level measures are highly reliable and valid in two different speech materials. They also showed acceptable results for both classification tasks in different speech materials, with word lists performing better than meaningful sentences. The differences between the two speech materials may be largely caused by differences in word structures and contextual cues in the materials. These measures perform better in word lists than in meaningful sentences, suggesting an advantage for using word lists in clinical practice and research. On the other hand, meaningful sentences can be used for classifying healthy and dysarthric speakers. Our results suggest that using different speech materials gives a better overview of the speakers' intelligibility at the segmental level and the implications of their articulation impairments.

This chapter is based on:

Xue, W., van Hout, R., Cucchiarini, C., & Strik, H. (2023). Assessing speech intelligibility of pathological speech in sentences and word lists: the contribution of phoneme-level measures. *Journal of Communication Disorders, 102*, Article 106301.

4.1 Introduction

Speech intelligibility is an important indicator of the degree of speech impairment in pathological speech and is often used by speech-language pathologists (SLPs) to diagnose individuals and evaluate the effectiveness of therapy (Van Riper & Emerick, 1984; Ishikawa et al., 2020). Individuals with dysarthria, a motor speech disorder caused by neurological injury (e.g., Parkinson's disease and stroke) experience difficulties in speech production that can lead to reduced speech intelligibility. To limit this loss in speech intelligibility and to even achieve some improvements, intensive speech therapy can be provided (Cannito et al., 2012; Levy et al., 2020; Mendoza Romas et al., 2021b; Nakayama et al., 2020; Yuan et al., 2020). To monitor therapy effectiveness, a clear definition of speech intelligibility is necessary. From many of the definitions presented in the literature, we adopted the one proposed by Hustad (2008): "Intelligibility refers to how well a speaker's acoustic signal can be accurately recovered by a listener" (p. 562). This definition clearly implies that speech intelligibility cannot be assessed without the participation of human listeners, either experts like speech-language pathologists or naïve listeners like college students. Common procedures of measuring intelligibility ask listeners to evaluate the intelligibility of different speech materials (e.g., a word list, a sentence or a passage) through different measurement methods (e.g., a visual analogue scale (VAS), an orthographic transcription, or a multiple-choice task).

Through these procedures, measures of intelligibility can be derived at different granularity levels, namely subword, word, sentence, and narrative. In this study, measures at the subword level refer to subword segments as basic units, namely phoneme, letter (grapheme), and syllable. Similarly, measures at the word level refer to the word as the basic unit, where words may occur either in isolation as in word lists or in sentences. Likewise, such definition can be generalized to sentence and narrative levels. It is important to note that some other studies used different names for the measures of the same unit in different speech materials. For example, a percentage of correctly transcribed words, considered as a word-level measure in this study, was named as "word intelligibility" or "single-word intelligibility" (Stipancic et al., 2016; Sussman & Tjaden, 2012; Yorkston & Beukelman, 1980, 1981) if the words were uttered in isolation, whereas it

was named as "sentence intelligibility" if the words were presented and assessed for individual sentences (e.g., Abur et al., 2019; Stipancic et al., 2016; Sussman & Tjaden, 2012).

Measures of intelligibility perform differently in speech materials with varying lengths (Hustad, 2007; Yorkston & Beukelman, 1978; Miller, 2013; Xue et al., 2021a, 2021b). For example, Hustad (2007) studied a word-level measure (i.e., percentage of correctly transcribed words) in three kinds of speech materials (i.e., word lists, sentences, and paragraphs). The author found higher intelligibility scores in paragraphs than in the other two types of materials. Chapter 3 studied the correlations of two measures (i.e., percentage of correctly transcribed words and ratings through VAS) between sentences and word lists. The correlations below 0.90 suggested that these two types of speech materials can result in different constructs of intelligibility being measured.

Further, the different performance in speech materials with varying lengths does not seem to be consistent across severity levels of dysarthria of speakers. First, when speech is mildly to moderately degraded or dysarthric, higher scores for word-level measures have been found on words in sentences than on words in isolation (Hustad, 2007; Yorkston & Beukelman, 1978, 1981). This may be due to the additional contextual cues provided by sentences. Second, when speech is more severely dysarthric, this difference between words in isolation and sentences seems to become less clear (Dongilli, 1994; Hustad, 2007; Miller, 2013; Yorkston & Beukelman, 1978, 1981). For instance, regarding intelligibility scores of profound dysarthric speech, some studies found higher scores on sentences than on words; some found equal scores (Barreto & Ortiz, 2008; Dongilli, 1994; Yorkston & Beukelman, 1978), and some others found lower scores (Yorkston & Beukelman, 1981). These inconsistencies between sentences and word lists may arise because speakers with more severe dysarthria have so many difficulties with pronunciation that listeners are no longer able to benefit from the contextual cues present in sentences.

Although measures can be derived at different granularity levels, intelligibility measures at the subword level may provide more detailed insights into misarticulation detections and their relationships with intelligibility, particularly at the segmental level. In fact, articulation is a key

feature of dysarthria (Tjaden, 2007), and it has been found to be a stronger contributor to intelligibility compared to other speech characteristics (i.e., voice quality, nasality, and prosody) (De Bodt et al., 2002). Also, therapy focusing on articulation has been preferred by many SLPs regardless of severity levels of dysarthria (Miller & Bloch, 2017). Therefore, studying subword-level measures can provide information on articulation impairment to help SLPs in diagnosing individuals and monitoring therapy.

For detecting misarticulation, measures at the phoneme level may be better than those at the grapheme and syllable levels. This is because languages may differ in their degree of orthographic depth (Katz & Frost, 1992), which is the consistency in relationships between graphemes and phonemes. English is often cited as an example of a deep orthography with inconsistent relationships, whereas Italian has more consistent relationships and is usually categorized as a shallow orthography. Dutch, the language under study, is somewhere in between, and in Dutch, a cluster of two graphemes in some cases may represent a single phoneme (Landerl & Reitsma, 2005). In addition, the same grapheme may represent different phonemes depending on the languages under study or the grapheme's position in a word. An example of the difference between languages is that vowel duration spelling ('aa' vs 'a') in Dutch is phonologically consistent in terms of a closed vs an open syllable but lexically inconsistent (e.g., 'paar'–'pa-ren' for singular - plural), whereas this is the opposite in German (e.g., 'Paar'–'Paa-re'). An example of the difference between different positions in a word is that in English, the letter A (lower case a) is usually pronounced /æ/ in closed syllables, like 'cat', and /eɪ/ in open syllables, like 'ta' in 'take'. Hence, for assessing articulation impairment and providing diagnostic information, phoneme-level measures may be more efficient than grapheme- and syllable-level measures. Therefore, the current study focuses on phoneme-level measures.

Until now, research on phoneme-level measures had several limitations. First of all, phoneme-level measures of intelligibility have been limited to word lists rather than sentences. In detail, these measures have been generally computed on word lists consisting of paired words with phoneme contrasts (Ansel & Kent, 1992; Kent et al., 1989; Kim et al., 2011a; Levy et al., 2016). Considering that speech materials of different lengths can impact higher-level

measures in different ways, it is necessary to study phoneme-level measures in speech materials varying in length. In this way, phoneme-level measures may help explore the above-mentioned different performance of higher-level measures in sentences and word lists, as suggested in Chapter 3. In addition, existing studies of phoneme-level measures have considered only the target unit (Platt et al., 1980; Furia et al., 2001), such as target consonants and vowels in the Dutch Intelligibility Assessment (Middag, 2012), rather than all the units composing a word. This may lead to loss of articulation information and may thus impact the evaluation of the effectiveness of articulation therapy. Moreover, studying phoneme-level measures in sentences may help investigate speakers' articulation imprecision in this more natural condition, in which words occur within a meaningful context, compared to when they are uttered in isolation as in word lists. Therefore, it is worthy and necessary to consider differences in speech materials when investigating the impact of measures at the phoneme level.

In addition to the limitations of current research on phoneme-level measures, it is important to note that the reliability of intelligibility measures used can be affected by transcription instructions. Currently, two types of instructions have been studied for orthographic transcriptions. In one of them, denoted as Existing-Word Transcription (EWTrans) in Chapter 3, listeners are instructed to transcribe only existing, meaningful words that have the closest pronunciation to the words they heard. The other one instructs listeners to transcribe words consisting of phonemes that have the same pronunciations as what they heard, regardless of whether the transcribed words are meaningful or not. This type of transcription is denoted as Acceptable-Word Transcription (AWTrans) and has been explored in Chapter 3.

Although EWTrans has been widely used (Abur et al., 2019; Barreto & Ortiz, 2008, 2016; Carvalho et al., 2021; Ganzeboom et al., 2016; Hodge & Gotzke, 2014; Hustad, 2006, 2007, 2008; Ishikawa et al., 2020; Liss et al., 2002; Middag, 2012; Miller, 2013; Stipancic et al., 2016; Tjaden & Liss, 1995a, 1995b, 2014, 2010; Sussman & Tjaden, 2012; Yorkston & Beukelman, 1978, 1981), it may generate intelligibility measures with lower reliability and poorer variability (i.e., scores were clustered rather than spread) than those obtained from AWTrans. For example, Chapter 5 found very low reliability (0.47) for a

word-level measure of intelligibility, namely word accuracy (AcW), obtained through EWTrans, which was much lower than that for the measure of intelligibility obtained from VAS (0.93). Similarly, Chapter 3 examined AcW scores obtained from both EWTrans and AWTrans and found lower reliability of AcW in EWTrans (0.83) compared to that in AWTrans (0.89). Both reliability scores were lower than the reliability of VAS (0.93). In both studies mentioned above, the low reliability values of AcW obtained in EWTrans were caused by a ceiling effect. Many of the sentences received scores of 100, indicating perfect intelligibility, which led to poor variability of AcW scores in EWTrans. A possible explanation of the presence of perfect intelligibility in EWTrans is that listeners, following the instructions, may transcribe the same meaningful words even if they perceived distortions, substitutions, deletions or insertions of a phoneme in a word. In contrast, AWTrans provides listeners with more flexibility by allowing them to use pseudowords in transcriptions. The rich variability in AWTrans would be useful for providing diagnostic information since it may reflect smaller but relevant differences between the speakers. Therefore, to examine phoneme-level measures of intelligibility derived from transcriptions, AWTrans is preferred over EWTrans.

To fill the knowledge gaps presented above, this study, as a follow-up to Chapter 3, aims to examine how phoneme-level intelligibility measures can complement higher-level intelligibility measures. Specifically, we explored what insights the phoneme-level intelligibility measures can provide in relation to two types of speech materials and two classification tasks. For the two types of speech materials, we considered meaningful sentences from a narrative and word lists consisting of words and pseudowords with a Consonant-Vowel-Consonant (CVC) structure. For the two classification tasks, we considered a two-way classification of speaker types (healthy vs dysarthric speakers) and a four-way classification of severity levels of dysarthria (SevL; healthy, mild, moderate, and severe) by spliting dysarthric speakers into three fine-grained levels. Two types of phoneme-level intelligibility measures programmatically extracted from AWTrans were studied: Accuracy of Phonemes (AcP), which is the percentage of correctly transcribed phonemes, and Phonetic Distance (PhonD), which is the degree of difference between transcribed phonemes and the target reference phonemes. We address three research questions:

- RQ1: To what extent are phoneme-level intelligibility measures reliable across different types of speech materials?
- RQ2: To what extent are phoneme-level intelligibility measures valid across different types of speech materials?
- RQ3: To what extent are phoneme-level intelligibility measures able to classify speakers in speaker type and severity levels of dysarthria across different types of speech materials?

In detail, to address RQ1, we examine the interrater reliability of the phoneme-level intelligibility measures by using Generalizability Theory. To address RQ2, we study the validity by correlating the phoneme-level intelligibility measures with three criterion variables. To address RQ3, we first calculate intelligibility measures for different categories of phonemes (i.e., consonants and vowels) and then use different combinations of intelligibility measures to classify speakers. Studying different categories of phonemes is important to explore whether speakers show different tendencies in different categories.

4.2 Method

This study evaluates two types of phoneme-level measures of intelligibility (i.e., AcP and PhonD) in different speech materials regarding their reliability, validity, and performance in two classification tasks. The tasks included a two-way classification of speaker types and a four-way classification of severity levels of dysarthria. Both types of phoneme-level measures were derived from orthographic transcriptions in AWTrans form collected through two listening experiments described in Chapter 3. Each of the two listening experiments targeted only one type of speech material, namely the Sentence Experiment for meaningful sentences and the Word Experiment for word lists consisting of CVC words. All the speech materials, speakers and their recordings to be assessed in these two experiments were selected from the Corpus of Pathological and Normal Speech (COPAS) database[5] (Middag, 2012), which consists of a large number of recordings of various reading materials, such as isolated words, isolated sentences, short passages, and spontaneous

[5] More information and the manual of COPAS can be found on https://taalmaterialen.ivdnt.org/download/tstc-corpus-pathologische-en-normale-spraak-copas/.

speech, by speakers of Belgian-Dutch with and without speech disorders. Both experiments received ethical approval as shown in Chapter 3.

4.2.1 Speech material

Two types of speech materials were evaluated in two separate listening experiments, as mentioned above. The Sentence Experiment covered four meaningful sentences that were selected from the Dutch commonly-used phonetically-balanced narrative 'Papa en Marloes' ('Papa and Marloes' in English; Van de Weijer & Slis et al., 1991): (1) 'Papa en Marloes staan op het station.' (in English 'Papa and Marloes are at the station.'), (2) 'Marloes kijkt naar links.' (in English 'Marloes looks to the left.'), (3) 'In de verte ziet ze de trein al aankomen.' (in English 'In the distance she can see the train coming.'), and (4) 'Het is al vijf over drie dus het duurt nog vier minuten.' (in English 'It is already five past three so it will take another four minutes.').

The Word Experiment covered word lists that were selected from those in the three subsets A, B, and C constructed in the Dutch Intelligibility Assessment (DIA) task (De Bodt et al. 2006), which is also contained in the COPAS dataset. This DIA task was originally designed to assess intelligibility at the phoneme level, resulting in a measure called Phoneme Intelligibility (PhonI). The three subsets are constructed to assess initial consonants, final consonants, and medial vowels in CVC words, respectively. Note that different speakers received a variant of each subset, leading to different combinations of three word lists, whereas all speakers received the same four sentences in the Sentence Experiment.

4.2.2 Speakers

In total, 26 speakers with dysarthria and 10 healthy speakers were selected from the COPAS database, as described previously in Chapter 3. These speakers covered different dysarthria types, severity levels of dysarthria, PhonI scores obtained through the original DIA task, age, and gender to ensure the diversity of speaker data. All of the speakers were involved in the Sentence Experiment, and only half of them were involved in the Word Experiment. Involving fewer speakers in the Word Experiment was to ensure that listeners would be able to finish both experiments within one hour based on the fact that more words were contained in the word lists

than in the sentences. Figure 4.1 shows the distribution of the speakers over the four different levels (healthy, mild, moderate, and severe from low to high) of SevL. The number of healthy speakers is comparable to the numbers of speakers in the different levels of dysarthria. We included healthy speakers to allow comparisons with dysarthric speakers.

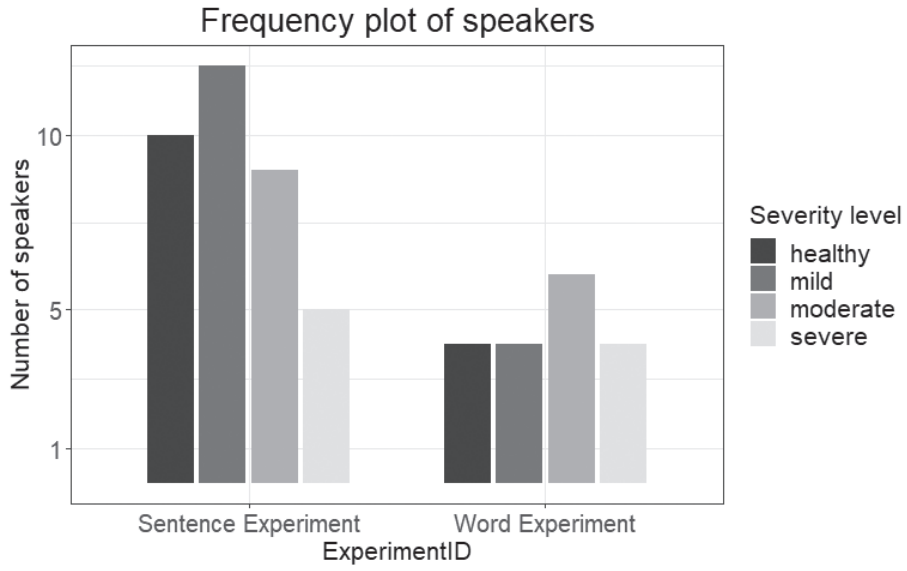


Figure 4.1. Distribution over four severity levels of dysarthria (SevL) of speakers in our two experiments.

4.2.3 Expert listeners

Listeners for assessing speech intelligibility in both experiments were the same five Belgian Dutch-speaking speech-language pathologists (one male and four females) recruited from the University Antwerp Hospital, as described in Chapter 3. They all worked with individuals with communication disorders and were all familiar with evaluating and testing dysarthric individuals through the intelligibility tasks used in the two experiments.

4.2.4 Experimental procedure

In the listening experiments, all recordings covering both speech materials were made in a quiet clinical setting without a sound-attenuated booth,

which was described in the COPAS manual, and had the same sound quality. Both listening experiments were conducted online through the survey tool Qualtrics on the same day, the Sentence Experiment first and the Word Experiment second, with two resting hours in between. The listeners gave their digital consent before they participated. All the instructions and explanations were presented in Belgian Dutch, the target language. Before the actual experiments, the listeners familiarized themselves with the procedure of assessment through practice examples. Both experiments employed anchor items to ensure reliability. The recordings for these items were selected from healthy speakers and speakers with severe dysarthria from the COPAS dataset. They were presented at the beginning of the actual experiments and repeated after every ten utterances. The recordings used for the practice examples and anchor items were not from speakers involved in the two actual experiments. To avoid any systematic order effect on the assessment, the utterances (recordings) were randomized. Specifically, the utterances (recordings) were randomized in a way that every six consecutive samples were not from the same speakers, and every two consecutive samples were not about the same sentences or subsets. The randomized order of samples was kept the same for all the listeners.

In detail, the Sentence Experiment contained 144 utterances (recordings), consisting of the same set of four sentences read by 36 speakers. The Word Experiment involved 54 utterances, consisting of three word lists (three variants of three DIA subsets) read by 18 speakers. Each utterance was presented once to prevent the listeners from adapting to speakers and speech materials. As described in Chapter 3, for each utterance, the listeners received an orthographic transcription task (in AWTrans form) and a VAS task on the same page at the same time, with AWTrans presented above VAS. The listeners completed both tasks in the order they preferred. The AWTrans form of transcription allows not only existing words but also pseudowords. The VAS task was a horizontal scale ranging from 0 (not intelligible) to 100 (intelligible) with written instructions ‘Wat voor score zou u toekennen aan de spraakverstaanbaarheid?’ (in English ‘what score would you assign for speech intelligibility?’). The scale contained tick marks with numbers for every ten scores shown (e.g., 10, 20, 30, etc.) but no scale endpoints’ labels. Moreover, the listeners in the Word experiment had to

transcribe the whole words in the word lists without any phonemes being presented. This was different from the original DIA procedure, in which the listeners were asked to transcribe the missing target phonemes while the remaining phonemes of a word were presented.

4.2.5 Intelligibility measures

To obtain phoneme-level measures, we first cleaned the collected orthographic transcriptions by removing punctuation and symbols indicating missing words. Following this, we applied a Grapheme-to-Phoneme (G2P)[6] conversion for all the unique words in both the prompts and the orthographic transcriptions of all utterances. Then, we manually checked and corrected the transformed phonetic transcription of each word. Finally, we concatenated the phonetic transcriptions of words into phonetic transcriptions of sentences, word lists, and prompts.

Based on the phonetic transcriptions, two types of phoneme-level measures were computed: Accuracy of Phonemes (AcP) and Phonetic Distance (PhonD). The AcP was calculated as follows:

$$AcP = \frac{N_{match}}{N_{total}} \times 100$$

where N_{total} denotes the total number of phonemes in the reference transcriptions, and N_{match} denotes the number of matched phonemes between the reference and the phonetic transcriptions. The PhonD was calculated as follows:

$$PhonD = \begin{cases} \frac{Dist_{phoneme}}{N_{total}} \times 100, & if \ \frac{Dist_{phoneme}}{N_{total}} \leq 1 \\ 100, & if \ \frac{Dist_{phoneme}}{N_{total}} > 1 \end{cases}$$

where $Dist_{phoneme}$ denotes the sum of the phonetic distance between the reference and phonetic transcriptions calculated by ADAPT (Elffers et al., 2005), which is a dynamic programming algorithm that computes the optimal alignment

---

[6] The applied G2P were conducted through the "Grapheme to Phoneme Converter" provided in https://webservices.cls.ru.nl/g2pservice.

between two strings of phonemes. The computed optimal alignment has the smallest phonetic distance between two strings of phonemes. For each pair of phonemes (one in reference and one in phonetic transcription), the phonetic distance is calculated as the total difference between the feature vectors of the pair of phonemes based on the feature-based distance metrics developed by Cucchiarini (1993, 1996). These metrics present the feature vectors of all phonemes. In detail, as described and established by Cucchiarini (1993, 1996), the feature vectors for vowels consist of five features (i.e., length, front/back, tongue, roundedness, and diphthong). The feature vectors for consonants consist of eleven features. These eleven features are place of articulation (plc), voice (voc), nasal (nas), stop (stp), glide (gld), lateral (lat), fricative (frc), trill (trl), aspirant (asp), dental (dnt), and strength (str). Table 4.1 presents an example of how the phonetic distance between two phonemes is calculated in ADAPT (Elffers et al., 2005).

Table 4.1. The feature vectors of phoneme /t/ and /s/ and the phonetic distance between them used in ADAPT (Elffers et al., 2005).

Feature vector	plc	voc	nas	stp	gld	lat	frc	trl	asp	dnt	str	Phonetic distance
/t/	4.0	1.0	0.0	0.5	0.0	0.0	0.0	0.0	1.0	1.0	1.0	
/s/	4.0	1.0	0.0	0.0	0.0	0.0	0.5	0.0	1.0	1.0	1.0	
Difference	0.0	0.0	0.0	**0.5**	0.0	0.0	**0.5**	0.0	0.0	0.0	0.0	**1.0**

Note. **plc**: place of articulation; **voc**: voicing; **nas**: nasal; **stp**: stop; **gld**: glide; **lat**: lateral; **frc**: fricative; **trl**: trill; **asp**: aspirant; **dnt**: dental; **str**: strength.

4.2.6 Data analysis

All the following statistical analyses were conducted in RStudio (RStudio Team, 2020) with R version 4.0.2 (R Core Team, 2020). We first studied the descriptive results (mean and standard deviation) of the two types of phoneme-level intelligibility measures. For a detailed study of the AcP measure, we calculated AcP by phoneme categories. We distinguished three global AcP scores, namely for all phonemes (AcP_all), for consonants only (AcP_consonants), and for vowels only (AcP_vowels). We applied Levene's test to study whether the variances of the four measures between the two experiments are significant or not. We also applied Welch's *t*-test to study whether the mean values are significant or not. We further studied whether

the mean values of AcP between the two categories (i.e., consonants and vowels) are significant or not through paired *t*-test. The statistical analyses were implemented by using the *car* package (Fox, 2019).

To address the first research question regarding the reliability of AcP and PhonD, we applied Generalizability Theory to compute the D coefficient as interrater reliability by using the *gtheory* package (Moore, 2016). The reason for applying Generalizability Theory was that it can simultaneously deal with all sources of variance relevant in the same experiments (e.g., listeners, speakers, and utterances). The model designs for the two experiments were different due to the speech materials. The model design for the Sentence Experiment was a fully-crossed design as *Listener*×*Utterance*×*Speaker* because all four sentences (the same utterance sample) from all 36 speakers were assessed by all five listeners. The model design for the Word Experiment was a nested design as (*Utterance*:*Speaker*)×*Listener* because word lists (different utterance samples) were nested under speaker, the combination of which was assessed by the listeners. More information and explanations about these model designs and their analyses can be found in Chapter 3.

To address the second research question regarding the validity of AcP and PhonD, we calculated Pearson correlations between our phoneme-level measures and three criterion measures, namely scores obtained from VAS, the accuracy of words (AcW) obtained from orthographic transcriptions in AWTrans form, and PhonI obtained from the original DIA task. The former two higher-level measures (i.e., VAS and AcW) and the phoneme-level measures in this chapter were collected from the same two listening experiments as described in Chapter 3. The PhonI obtained from the original DIA task was calculated as the percentage of correctly transcribed target phonemes over all three word lists. In more detail, these target phonemes were transcribed in the presence of the remaining phonemes of a word (e.g., transcribing target phoneme 'n' in 'nit' which was presented as '.it'). All three criterion measures were shown to be reliable and valid (Van Nuffelen et al., 2008; Xue et al., 2021b). Note that the correlations were computed when the scores were aggregated for speakers.

To address the third research question regarding the performance of measures in classifying speakers into two speaker types and four severity levels, we first aggregated the scores for speakers. For a detailed study of

the AcP measure, we calculated AcP by phoneme categories. Aside from the three global AcP scores (i.e., AcP_all, AcP_consonants, and AcP_vowels), we computed detailed scores for place and manner of articulation for consonants and frontness and openness for vowels. The classes of consonants according to the place of articulation were labial (/p, b, f, v, m, ʋ/), alveolar (/t, d, s, z, n, l/), post-alveolar (/ʃ, ʒ/), dorsal (/k, g, x, ɣ, ŋ, j/) and glottal (/ɦ/). The manner classes were plosive (/p, b, t, d, k, ɡ/), fricative (/f, v, s, z, ʃ, ʒ, x, ɣ, ɦ/), nasal (/m, n, ŋ/), trill (/r/), and approximant (/ʋ, l, j/). The frontness vowel classes were front (/ɪ, i, ɛ, eː, ɛi/), central (/Y, y, ə, œy, øː/), and back (/u, ɔ, oː, ɑ, aː, ʌu/). The openness classes were open (/ɑ, aː/), middle (/ɛ, eː, ɔ, oː, œy, øː, ʌu, ɛi, ə/), and closed (/ɪ, i, Y, y, u/). In total, we calculated 20 phoneme-level intelligibility measures. The list of all intelligibility measures collected through the two experiments is shown in Table 4.2.

Table 4.2. The list of intelligibility measures collected through the two experiments.

Level of measures	Type of measures	Unit	Phoneme category	Names of measures	Number of measures
higher-level	VAS	utterance	-	VAS	2
	AcW	word	-	AcW	
phoneme-level	PhonD	phoneme	all phonemes	PhonD	20
	AcP	phoneme	all phonemes	AcP_all	
		consonant	all consonants	AcP_consonants	
			place of articulation	AcP_labial, AcP_alveolar, AcP_post-alveolar, AcP_dorsal, AcP_glottal	
			manner of articulation	AcP_plosive, AcP_fricative, AcP_nasal, AcP_trill, AcP_approximant	
		vowel	all vowels	AcP_vowels	
			frontness of vowels	AcP_front, AcP_central, AcP_back	
			openness of vowels	AcP_open, AcP_middle, AcP_closed	

We visualized all 20 phoneme-level measures per speaker through boxplots over four SevL levels and calculated the Pearson correlations between those scores in the three sets of experiments. The three sets of experiments are the whole set of the Sentence Experiment (N=36), a subset of the Sentence Experiment (N=18) in which the same 18 speakers in the Word Experiment were involved, and the whole set of the Word Experiment (N=18). We studied whether the variances and the mean values of all 20 phoneme-level measures between every two levels of severity are significant by applying Levene's test and Welch's *t*-test, respectively.

Finally, we used the scores of the different phoneme categories to conduct two classification tasks (i.e., speaker type and severity level) through Support Vector Machine (SVM) with Radial (Gaussian) kernel using 5-folds cross-validation. We applied three groups of variables as predictors in SVM: (1) all 20 phoneme-level intelligibility measures, (2) two higher-level measures, and (3) combining these former groups, resulting in all 22 intelligibility measures. The higher-level measures were intended to explore whether these measures supplement or may be different from detailed phoneme-level measures. Moreover, our phoneme-level measures generally had very strong correlations (see Section 4.3.4). As this may impact the robustness of the SVM results, we also explored the classification power of these measures using Random Forest (RF). Furthermore, we applied a multinominal regression to study the relation between each of the 20 phoneme-level measures with SevL. Detailed information and the corresponding results about RF and the multinominal regression can be found in Appendix C and Appendix D, respectively.

The correlation results were interpreted based on the guidelines from Evans (1996, p. 146). That is, a correlation between 0.80 and 1.0 is 'very strong', between 0.60 and 0.79 is 'strong', between 0.40 and 0.59 is 'moderate', between 0.20 and 0.39 is 'weak', and even lower is 'very weak'. The implementation of the analyses for the descriptive results and addressing the second and third research questions was conducted by using the *stats* (R Core Team, 2020), *caret* (Max, 2022), *e1071* (Meyer et al., 2021), and *ggplot2* (Wickham, 2016) packages.

4.3 Results

4.3.1 General results of phoneme-level measures of intelligibility

The mean and standard deviation (SD) of the two types of phoneme-level intelligibility measures (i.e., PhonD and AcP) in both experiments are shown in Table 4.3. In particular, three global AcP scores were studied. In general, significantly higher mean AcP and lower mean PhonD values were found for the Sentence Experiment than for the Word Experiment (Welch's *t*-test, $p < .001$). The variances between the two experiments were significantly different (Levene's test, $p < .001$) for all four measures. Also, when comparing AcP between the two categories (i.e., consonants and vowels), the AcP of consonants showed significantly lower means than the AcP of vowels in both experiments. The paired *t*-test reported t(719)=-7.96 ($p < .001$) for the Sentence Experiment and t(269)=-2.6363 ($p < .01$) for the Word Experiment.

Table 4.3. Mean (Standard Deviation) of the measures in the two experiments. Higher AcP scores are associated with higher intelligibility, while PhonD scores have the opposite direction.

Mean (Standard Deviation)		Sentence Experiment	Word Experiment
PhonD	**all phonemes**	21.27 (31.53)	51.98 (30.83)
AcP	**all phonemes (AcP_all)**	90.01 (19.96)	71.38 (18.79)
	consonants (AcP_consonants)	89.08 (21.23)	70.51 (20.73)
	vowels (AcP_vowels)	91.90 (18.89)	73.01 (19.10)

4.3.2 Interrater reliability of phoneme-level measures of intelligibility

Table 4.4 shows the results of the interrater reliability (D coefficients) for PhonD and AcP of different categories in both experiments. In general, all four phoneme-level measures showed relatively high reliability values (> 0.90), except PhonD in the Sentence Experiment and AcP_vowels in the Word Experiment. Interestingly, even though both AcP and PhonD were extracted from the same phonetic transcriptions, AcP appeared to be more

reliable than PhonD in the Sentence Experiment. Surprisingly, the reliability for the AcP of vowels in the Word Experiment was lower than all the other results. This may be due to the presence of pseudowords and the lack of contextual information.

Table 4.4. Interrater reliability (D coefficients) of our two types of phoneme-level measures (i.e., PhonD and AcP) for both experiments. AcP scores were calculated for all phonemes (AcP_all), for consonants only (AcP_consonants), and for vowels only (AcP_vowels).

Interrater reliability		Sentence Experiment	Word Experiment
PhonD	**all phonemes**	0.89	0.95
AcP	**all phonemes (AcP_all)**	0.93	0.97
	consonants (AcP_consonants)	0.93	0.97
	vowels (AcP_vowels)	0.90	0.83

4.3.3 Validity of phoneme-level measures of intelligibility

To investigate the validity of our phoneme-level measures, we correlated each with three criterion measures (i.e., VAS, AcW, and PhonI through the original DIA task) for both experiments. As shown in Table 4.5, both types of phoneme-level measures (i.e., AcP and PhonD) showed very strong correlations (magnitude > 0.90) with both VAS and AcW regardless of speech materials, whereas they showed weaker correlations (magnitude between 0.60 and 0.75) with PhonI.

Table 4.5. Correlations between criterion measures (VAS, AcW, and PhonI) and our phoneme-level measures (i.e., PhonD and AcP for all phonemes, consonants only, and vowels only) for both experiments.

Correlations		Sentence Experiment (N = 36)			Word Experiment (N = 18)		
		VAS	**AcW**	**PhonI**	**VAS**	**AcW**	**PhonI**
PhonD	**all phonemes**	-0.96	-0.98	-0.65	-0.94	-0.98	-0.65
AcP	**all phonemes (AcP_all)**	0.97	0.95	0.67	0.97	0.95	0.74
	consonants (AcP_consonants)	0.97	0.96	0.68	0.97	0.93	0.72
	vowels (AcP_vowels)	0.96	0.93	0.63	0.90	0.95	0.74

4.3.4 Phoneme-level measures of intelligibility in classifying speakers

Figure 4.2 illustrates the distributions of 20 phoneme-level measures for the four levels of SevL through boxplots in the whole set of the Sentence Experiment (N=36), the subset of the Sentence Experiment (N=18), and the Word Experiment (N=18). In detail, these measures were PhonD, three global AcP scores for three different categories, namely one for all phonemes (AcP_all), one for only consonants (AcP_consonants), and one for only vowels (AcP_vowels), as well as AcP of the place and the manner of articulation for consonants and AcP of frontness and openness for vowels. We also applied *t*-test to explore whether the mean values of the 20 phoneme-level measures between every two levels of severity are significant or not. We summarized the significant results ($p < .05$) in Table 4.6.

Figure 4.2 generally shows that in the whole set of the Sentence Experiment, the median of the AcP scores decreased and the median of PhonD increased with increased severity levels of dysarthria. This tendency was further supported by the statistical results shown in Table 4.6. That is, significantly higher mean values of AcP and lower mean values of PhonD were frequently observed for lower severity levels of dysarthria. This tendency did not persist in the Word Experiment. In detail, although the median values for healthy speakers were consistently the highest and those for severe dysarthric speakers were the lowest, the median values for the middle two SevL levels were similar, but with sometimes one being higher than the other and sometimes the opposite. This was also supported by the statistical results that differences between mild and moderate levels were not significant for most of the 20 measures. Further, AcP_glottal showed different trends in the median values for the four SevL levels. We did not observe any significant results in the mean values of AcP_glottal except between mild and severe levels. Moreover, the subset of the Sentence Experiment showed different results. Although the boxplots show decreases in the median values of AcP with increased severity levels, differences between most pairs of AcP scores were not statistically significant. In contrast, PhonD more frequently showed significant results.

Table 4.6. The results of *t*-test for comparing the mean values of phoneme-level measures between every two levels of severity. Only significant results ($p < .05$) are reported.

***t*-test results**	**Sentence Experiment (N=36)**						**Sentence Experiment (N=18)**						**Word Experiment (N=18)**					
	01	**02**	**03**	**12**	**13**	**23**	**01**	**02**	**03**	**12**	**13**	**23**	**01**	**02**	**03**	**12**	**13**	**23**
PhonD	–	–	–		–	–		–	–		–	–	–		–		–	–
all	+	+	+	+	+	+		+					+	+	+		+	+
consonants	+	+	+	+	+	+							+	+	+		+	+
vowels	+	+	+	+	+	+		+					+		+		+	+
labial	+	+											+	+	+		+	+
alveolar		+	+	+	+	+							+	+	+		+	+
post-alveolar	+	+	+	+	+			+	+				+	+	+		+	+
dorsal	+	+	+	+	+								+		+			
glottal		+	+		+	+			+		+						–	
plosive	+	+	+	+	+	+							+	+	+		+	+
fricative		+	+	+	+								+		+			+
nasal	+	+	+	+	+										+			
trill	+	+	+		+			+	+				+	+	+		+	+
approximant	+	+	+		+	+							+	+	+		+	
front		+	+	+	+	+		+		+			+		+			+
central		+		+											+	–		+
back	+	+	+		+			+					+	+	+		+	+
open	+	+	+		+	+		+					+		+		+	+
middle		+	+	+	+	+		+		+			+		+		+	+
closed		+	+	+	+			+					+		+			

Note. Except PhonD, the others are all AcP scores; **01**: healthy vs mild; **02**: healthy vs moderate; **03**: healthy vs severe; **12**: mild vs moderate; **13**: mild vs severe; **23**: moderate vs severe; **+**: greater with p <.05; **–**: less with p <.05; an example: '02' with '+' means the mean values for healthy level are higher than those for moderate level.

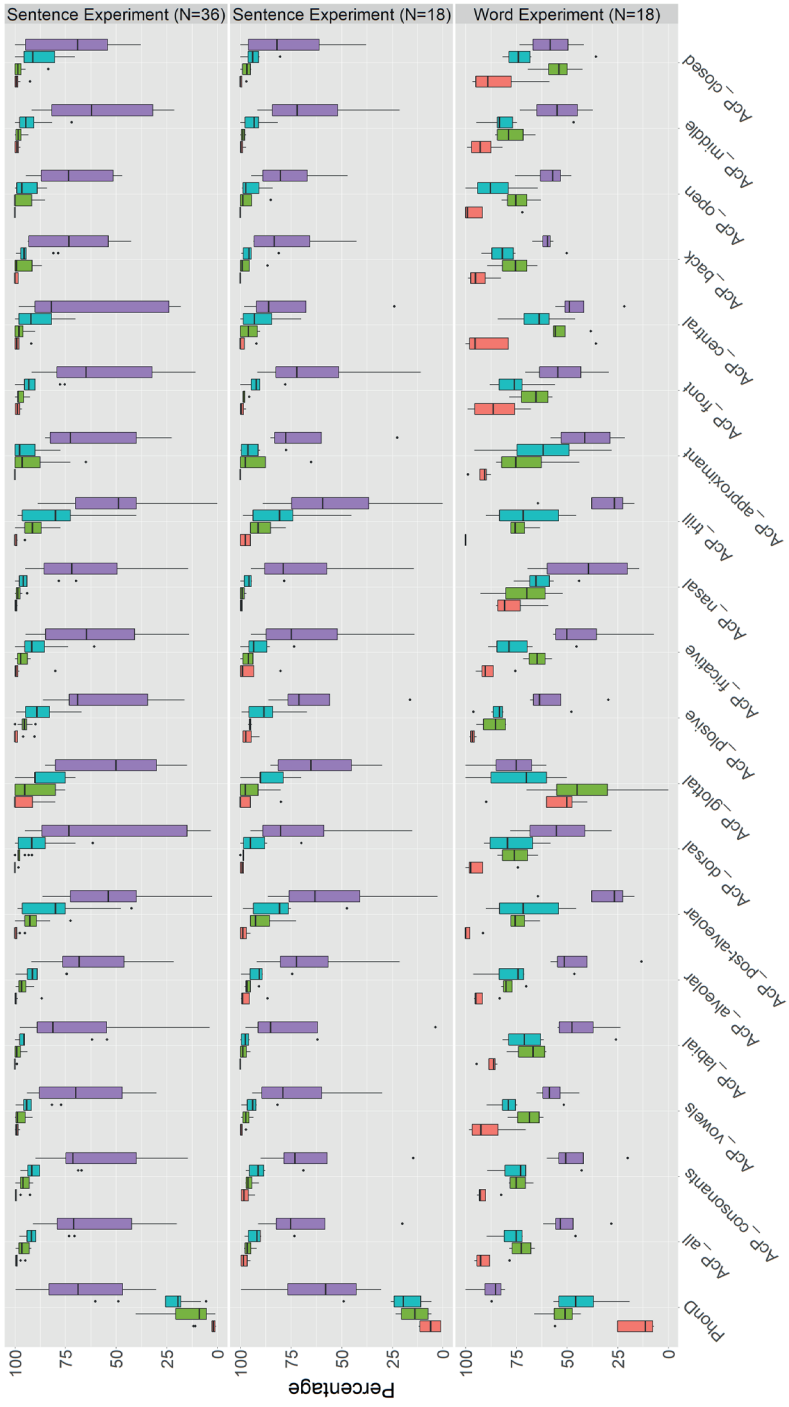


Figure 4.2. Boxplots of PhonD, Accuracy of phonemes (AcP) in three categories, namely all phonemes (AcP_all), consonants only (AcP_consonants) and vowels only (AcP_vowels), AcP for place and manner of articulation for consonants and frontness and openness for vowels in the whole set of the Sentence Experiment (N=36), the subset of the Sentence Experiment subset (N=18) and the Word Experiment (N=18) from top to bottom. The classes of consonants according to the place of articulation were labial (/p, b, f, v, m, ʋ/), alveolar (/t, d, s, z, n, l/), post-alveolar (/ʃ, ʒ/), dorsal (/k, g, x, ɣ, ŋ, j/) and glottal (/ɦ/). The manner classes were plosive (/p, b, t, d, k, g/), fricative (/f, v, s, z, ʃ, ʒ, x, ɣ, ɦ/), nasal (/m, n, ŋ/), trill (/r/), and approximant (/ʋ, l, j/). The frontness vowel classes were front (/i, i, ɛ, eː, ɛi/), central (/Y, y, ə, œy, ø/), and back (/u, ɔ, oː, ɑ, aː, ʌu/). The openness classes were open (/ɑ, aː/), middle (/ɛ, eː, ɔ, oː, œy, øː, ʌu, ɛi, ə/), and closed (/i, i, Y, y, u /). **Best viewed in color.**

Aside from the visualization of the phoneme-level measures, we further explored the correlations between them before applying any classification algorithms. The correlations between each pair of phoneme-level measures were generally very high in both sets of the Sentence Experiment (whole set: Mean=0.914, SD=0.052; subset: Mean=0.925, SD=0.056). This is in line with the boxplot results that all phoneme-level measures always showed similar trends. Differently, the correlations in the Word Experiment were slightly lower (Mean=0.724, SD=0.243) and those for AcP_glottal (Mean=0.112, SD=0.095) were very low. These results were also in line with trends represented in the boxplots. More detailed results about correlations can be found in Appendix B.

Following the correlation calculation, we further explored the performance of our phoneme-level measures together with the two higher-level measures in three groups for classifying speakers into two types of speakers and four severity levels by using SVM. The results in Table 4.7 show higher accuracy values for the two-way speaker type classification than for the four-way severity level classification. This is the consequence of having more classes. In addition, the accuracy values in the Word Experiment were generally higher than those in the two sets of the Sentence Experiment.

Table 4.7. Results for speaker type (healthy vs dysarthric) and severity level (healthy, mild, moderate, and severe) classification tasks using three groups of measures as independent variables in SVM implementation for the three sets of the experiments. Train = training speaker set; Valid = Validation speaker set.

Classification task	Independent variables (input)	Sentence Experiment (N=36)	Sentence Experiment (N=18)	Word Experiment (N=18)
		Accuracy (Train/Valid)	Accuracy (Train/Valid)	Accuracy (Train/Valid)
speaker type	**phoneme**	0.88/0.83	0.96/0.80	1.00/0.90
	high	0.92/0.92	0.78/0.80	0.88/0.90
	all	0.88/0.84	0.99/0.85	1.00/0.90
severity level	**phoneme**	0.81/0.64	0.86/0.33	1.00/0.64
	high	0.59/0.56	0.64/0.55	0.58/0.54
	all	0.85/0.70	0.88/0.38	1.00/0.64

Note. "**phoneme**" – 20 phoneme-level intelligibility measures; "**high**" – 2 higher-level intelligibility measures; "**all**" – all 22 measures.

The differences between the Sentence Experiment and the Word Experiment could be due to the different characteristics of the two speech materials. Nevertheless, when combining all phoneme-level intelligibility measures, around half of the speakers were classified correctly in both classification tasks regardless of speech materials. Such results were also supported when using different algorithms such as RF (see Appendix C) and multinominal regression (see Appendix D).

Furthermore, for speaker type classification, the accuracy values of the higher-level measures outperformed the phoneme-level measures in the whole set of the Sentence Experiment, whereas, in the other two sets, the phoneme-level measures outperformed the higher-level measures. For severity level classification, the phoneme-level measures outperformed the higher-level measures in all three experiments.

Combining the higher-level measures (i.e., VAS and AcW) with the phoneme-level measures seems to increase the accuracy values to a limited extent. This might be because the higher-level measures were highly correlated with the phoneme-level measures, as shown in Section 4.3.3, and consequently had a limited impact on the classification results when the number of predictors was large enough.

4.4 Discussion

In the present study, two types of phoneme-level measures were investigated (i.e., AcP and PhonD). AcP indicated whether a transcribed phoneme was correct or not, and PhonD showed how different a transcribed phoneme was from its reference phoneme. The phoneme-level measures were calculated based on the more permissive form of transcription (AWTrans), which allows all kinds of acceptable words, instead of the more common EWTrans, which allows only existing words. We considered two types of speech materials (i.e., meaningful sentences extracted from a narrative and word lists consisting of CVC existing and pseudowords). In particular, we studied these phoneme-level measures in terms of reliability, validity, and their performance in classifying speakers according to their speaker types and severity levels of dysarthria.

Our findings show that both types of measures showed largely similar reliability and validity in both speech materials, indicating that they can

be used in future studies and clinical trials. Regarding their performance in classifying speakers, the type of speech material seems to be the most critical factor. That is, both types of measures can classify speakers into two speaker types and four severity levels of dysarthria to a certain extent when using meaningful sentences. In contrast, they are more successful in classifying speakers when using word lists.

Before answering the research questions, we will first discuss the general results of both types of measures in both speech materials. The result that AcP indicated higher intelligibility in the Sentence Experiment than in the Word Experiment may be explained in two ways. First, this is perhaps due to the contextual information involved in sentences compared to word lists. This seems to be in line with previous studies of word-level measures (Hustad, 2007; Miller, 2013; Xue et al., 2020, 2021b; Yorkston & Beukelman, 1978) that higher scores were found for sentences than for word lists. Another possible explanation is the differences in listeners' familiarity with the words. The listeners were more familiar with the words involved in the sentences than those in the word lists because these sentences were extracted from a narrative that was frequently used by them in clinical trials. Also, the words contained in the sentences were existing, commonly used words, whereas the words in the word lists contained pseudowords, which were more difficult to transcribe than existing words.

Furthermore, when comparing AcP scores between the two categories (i.e., consonants and vowels), we found significantly lower means and higher standard deviations for the consonants than for the vowels, regardless of speech materials. This largely supports the finding by Miyakoda (2003) that consonants are more error-prone than vowels in both healthy and pathological speech.

4.4.1 RQ1: Interrater reliability of phoneme-level measures of intelligibility

Our first research question was to what extent these phoneme-level measures are reliable in the two types of speech materials. The generally very high reliability values (around 0.90) indicate that both types of phoneme-level measures are reliable regardless of speech materials. These findings are comparable to those in previous studies of phoneme-level measures.

For instance, De Bodt et al. (2006) reported very high reliability values of PhonI, which was used to examine the validity of our phoneme-level measures, with an interrater reliability of 0.93 and intrarater reliability of 0.91. Further, the relatively higher reliability values of AcP compared to PhonD seem to suggest that listeners performed similarly in transcribing a phoneme correctly. In contrast, they probably transcribed different symbols for the same target phoneme, which resulted in different distance scores and thus caused slightly lower reliability values of PhonD.

Moreover, our AcP for both consonants and vowels categories performed surprisingly well in terms of reliability, unlike the results reported by Van Haaften et al. (2019). In fact, the authors did report high point-to-point interrater agreement (larger than 95%) for both consonants and vowels obtained from two listeners for Dutch children speech of two tasks that were comparable to ours (i.e., picture naming and non-word imitation). However, they reported only sufficient reliability values of ICC for consonants and vowels. In detail, the reliability values for the percentage of consonants correct in syllable-initial position in both tasks were sufficient, with 0.80 in picture naming and 0.77 in non-word imitation. In contrast, the reliability values for the percentage of vowels correct were insufficient, with 0.59 in picture naming and 0.62 in non-word imitation. One possible explanation for the different results of reliability is the larger number of listeners (5) in our current study compared to that (2) in the study of Van Haaften et al (2019).

4.4.2 RQ2: Validity of phoneme-level measures of intelligibility

Our second research question was to what extent these phoneme-level measures are valid in the two types of speech materials. We can conclude that both measures are valid to a large extent according to the very strong correlations (magnitude > 0.90) to two of the criterion measures (i.e., VAS and AcW) and the strong correlations (magnitude > 0.60) to PhonI regardless of speech materials.

The very strong correlations with VAS and AcW are easy to understand because VAS and the orthographic transcriptions for generating AcW, AcP, and PhonD were all collected in the same experiments. The very strong

correlations with VAS seem to suggest that the listeners' global perception of intelligibility, as reflected by VAS, is highly correlated to articulation accuracy, as reflected by phoneme-level measures (AcP and PhonD). This substantially supports the previous findings of De Bodt et al. (2002) that articulation contributes the most to the overall intelligibility through a linear combination although these authors used rating scales for assessing articulation and the overall intelligibility.

In addition, unlike the very strong correlations of phoneme-level measures with VAS and AcW, both types of phoneme-level measures showed weaker correlations with PhonI. This may be because PhonI was obtained differently. PhonI was collected from a different speech material compared to the measures in the Sentence Experiment. For the Word Experiment, although PhonI and our two types of phoneme-level measures were collected from the same speech material, different procedures were used. In detail, PhonI was calculated as the percentage of correctly transcribed target phonemes with the rest phonemes of CVC words being presented (De Bodt et al., 2006; Middag et al., 2009b), whereas our phoneme-level measures were collected without any phoneme being presented. In other words, our approach examined more phonemes and, thus, leads to more room for error and more information on articulation imprecision. Also, transcribing a target phoneme with the context of the rest phonemes of a word requires less effort than transcribing the whole word.

4.4.3 RQ3: Phoneme-level measures of intelligibility in classifying speakers to different speaker types and severity levels of dysarthria

Our third research question was how well these phoneme-level measures can classify speakers into different types and different severity levels of dysarthria. We can conclude that our phoneme-level measures were similarly distributed over severity levels of dysarthria within each type of speech material, but showed different trends in different speech materials. By combining all phoneme-level intelligibility measures, we were able to classify more than half of the speakers correctly in both classification tasks regardless of speech materials, but the performance was much better when using word lists.

In general, the boxplots and the statistical results showed that intelligibility decreased for speakers with more severe dysarthria regardless of phoneme-level measures and speech materials. This seems to be in line with previous findings showing that higher-level measures of intelligibility were lower for speakers with more severe dysarthria (Hustad, 2007, 2008). However, meaningful sentences showed higher intelligibility than word lists, as reflected by the significantly higher mean AcP and lower mean PhonD values. This may be due to the contextual information in the sentences.

Furthermore, mixed trends were observed for different classes in the categories of consonants and vowels in different speech materials in terms of their median values and the sizes of boxes in the boxplots. For instance, moderate dysarthric speakers showed higher median and significantly higher mean values for the AcP of central vowels than mild dysarthric speakers in word lists, which was opposite to the results in meaningful sentences. Such mixed results in different speech materials, especially for dysarthric speakers, as well as the lower intelligibility in word lists compared to meaningful sentences, seem to broadly support the findings by Barreto and Ortiz (2010). They reported significant differences in a subword-level measure (i.e., percentage of correctly transcribed syllables) between sentences, lists of words, and lists of pseudowords among intelligible speakers and reported larger differences among less intelligible speakers. The authors inferred that these differences were due to the absence of semantic information in pseudowords, which led to increased sensitivity to acoustic-phonetic information, as reflected by the subword-level measure.

In addition, the AcP of the glottal consonant /ɦ/ in word lists surprisingly showed lower median values for healthy speakers and mild dysarthric speakers than those for moderate and severe dysarthric speakers. We applied a linear regression analysis to explore whether age and utterance were significant predictors to the AcP of glottal consonant. The analysis was conducted on AcP scores averaged per speaker per utterance (N=36). Note that having 36 samples instead of 54 samples was because we involved only two of the three word lists for the 18 speakers since the last one did not contain the glottal consonant. We observed that age and the interaction between age and utterance were significant ($p < .05$). We

further checked the word lists and found that one possible reason is the specific combinations of vowels and consonants that follow it. That is when /ɦ/, as an initial consonant in our word lists, was followed by /un/, /ɛin/, /ɪl/, /ɛn/, and /in/, which was the case mostly for healthy speakers and mild dysarthric speakers, it was often deleted or substituted. The vowels in these syllables were mainly front and closed vowels. However, when /ɦ/ was followed by /ɑn/, /an/, /yn/, /uf/, /oːn/, /oːs/, /ɔp/, /ʌut/, and /e/, which was the case mostly for moderate and severe dysarthric speakers, it was often perceived correctly. The vowels in these syllables were mainly back or front rounded vowels. We also checked the age of the speakers and found that some healthy speakers and speakers with mild dysarthria were older than 60, while speakers with moderate and severe dysarthria were always younger than 60 (see Table 3.1 in Chapter 3). Since aging can affect the functioning of the larynx (Kent et al., 1999; Linville, 1996; Weismer & Liss, 1991) and consequently the production of glottal consonants, this could partly explain the results we obtained for the glottal consonant.

Moreover, we observed very strong correlations (Mean >0.90) in the Sentence Experiment and relatively lower correlations (Mean >0.70) in the Word Experiment. These results seem to suggest that our phoneme-level measures showed similar distributions of scores when using meaningful sentences. This also indicates that these measures refer to a similar construct of intelligibility. In contrast, we observed slightly lower correlations when using word lists. This suggests that our measures refer to slightly different constructs of intelligibility. Further, the very strong correlations prevented us from observing significant differences in measures between speakers of different severity levels. Such limits may be because our datasets were a mixture of speakers with different types of dysarthria and were limited in size. Thus, further research should focus on larger datasets with sufficient numbers of speakers with different types of dysarthria.

Regarding the two classification tasks, word lists generated better results than meaningful sentences according to the higher accuracy values. One possible explanation for the different performance in classifying speakers in the two types of speech materials is the difference in some characteristics of the speech materials. As we explained earlier in Section 4.4, these two materials differ in their semantic cues, contextual

information, and word structures. These results also indicate that well-constructed word lists compared to meaningful sentences are more useful and successful in collecting phoneme-level measures for classifying speakers either to different speaker types or severity levels. Also, the finding that speakers were classified differently depending on the type of speech material used suggests that different speech materials are needed to obtain a comprehensive picture of speakers' intelligibility at the phoneme level.

The higher-level measures showed similar results in different speech materials regardless of the classification task. These results seem to suggest that the higher-level intelligibility measures, which are also considered as global intelligibility scores, refer to similar constructs of intelligibility regardless of the speech material (sentences or words). In contrast, the phoneme-level measures showed better results in the Word Experiment than in both sets of the Sentence Experiment. These results suggest that the phoneme-level intelligibility measures refer to different constructs of intelligibility in different speech materials. This might be due to, again, the different characteristics of the speech materials, such as different statistics of phonemes and amounts of contextual information contained.

The difference between AcP and PhonD was minor in terms of correlations and their distributions. This suggests that both types of measures can be used in clinical trials. However, since PhonD contains information on how different a transcribed phoneme is from a reference, more research is necessary to refine and further elaborate potential, detailed differences between PhonD and AcP.

Overall, the results seem to suggest that our two types of phoneme-level measures can be used for both classification tasks to a great extent when using word lists and for classifying speaker types to a certain extent when using meaningful sentences. Further, it may be too soon to conclude whether our phoneme-level measures can be used for classifying severity levels when using meaningful sentences. In future studies, it would be necessary to collect measures on larger datasets with more speakers at different severity levels.

4.4.4 Limitations and future work

The present study has some limitations that should be considered. First, the results obtained in the present study for expert listeners might not generalize

to naïve listeners who have no or limited experience in transcribing or are not trained for distinguishing sounds of minimal pairs of phonemes. In fact, differences between expert and naïve listeners have been reported not only on intelligibility measures (Carvalho et al., 2021; Monsen, 1983) but also on listening efforts (Maruthy & Raj, 2014). Thus, future work can investigate the same measures through transcriptions collected from naïve listeners and can thus compare the performance of measures between different types of listeners. Second, the two types of speech materials employed in the present study differ not only in the presence of contextual information but also with regard to fine-grained characteristics, such as the word structures and the number of syllables per word. Although the selection of speech materials was limited because we used the existing dataset COPAS, the lack of control over such characteristics may lead to results that differ from those obtained from well-controlled speech materials. For instance, Odell et al. (1991) found more errors in polysyllabic than monosyllabic words with both types of words being examined in isolation. However, it is difficult to draw the same conclusion from the current study. This is because words in the sentences consisted of polysyllabic and monosyllabic words, which is actually what happens in real life, whereas only monosyllabic words were contained in the word lists since these words were designed to assess target phonemes in CVC lexical structure. Therefore, more research is necessary to refine and further elaborate our novel findings.

4.4.5 Strengths

Our findings raise several important points for discussion. First, we contribute to the research on phoneme-level measures by examining the reliability and validity of two types of phoneme-level measures in two types of speech materials. The high reliability and validity of the two types of phoneme-level measures support their usage in clinical practice and research. Second, word lists perform better than meaningful sentences in classifying speakers in the two tasks. These results also seem to suggest that word lists are more sensitive to speakers' dysarthria type and age in detecting articulation impairment. These findings also indicate that speakers may articulate differently when receiving different levels of linguistic cues in speech materials. Therefore, as previous studies have suggested (Hustad,

2007; Yorkston & Beukelman, 1978), different speech materials should be considered when evaluating speech intelligibility, not only for word- and sentence-level measures but also for phoneme-level measures.

4.5 Conclusions

In conclusion, this study comprehensively analyzed two types of phoneme-level measures collected from experienced, well-trained expert listeners for two types of speech materials. Both types of phoneme-level measures are comparably reliable and valid. They show acceptable results in classifying speakers into two speaker types and four severity levels of dysarthria regardless of speech materials. Word lists show slightly higher levels of accuracy for both classification tasks compared to meaningful sentences. This is largely caused by different word structures, degrees of contextual information, and statistics of phonemes in the two speech materials, as well as individual differences. In turn, the different results in the two speech materials suggest that different speech materials are needed to evaluate speakers' intelligibility at the phoneme level. Our findings show an advantage for using word lists in clinical practice and research. On the other hand, meaningful sentences can be used as an alternative to word lists a) to distinguish speakers and b) to evaluate articulation problems, which can help diagnosis to a certain extent. To further refine our findings, future studies may focus on well-constructed sentences or sentences with phoneme and word structures comparable to those in word lists.


Meow!
Mow?

CHAPTER 5

SPEECH INTELLIGIBILITY OF DYSARTHRIC SPEECH: HUMAN SCORES AND ACOUSTIC-PHONETIC FEATURES

ABSTRACT

To gain a better understanding of how acoustic-phonetic correlates could be employed to obtain more objective measures of speech intelligibility and a better classification of dysarthric and non-dysarthric speakers, we studied the relation between different intelligibility measures and some important acoustic-phonetic correlates. We investigated speech intelligibility in dysarthric and non-dysarthric speakers as measured by two commonly used methods, ratings through the Visual Analogue Scale (VAS) and word accuracy (AcW) through orthographic transcriptions. We found that the two intelligibility measures are related, but distinct, and that they might refer to different components of the intelligibility construct. The acoustic-phonetic features showed no difference in the mean values between the two speaker types at the utterance level, but more than half of them played a role in classifying the two speaker types. We computed an acoustic-phonetic probability index (API) per speaker. API is moderately correlated to VAS ratings but not correlated to AcW. In addition, API and VAS complement each other in classifying dysarthric and non-dysarthric speakers. This suggests that the intelligibility measures assigned by human listeners and acoustic-phonetic features relate to different constructs of intelligibility.

This chapter is based on the following publication:

Xue, W., van Hout, R., Boogmans, F., Ganzeboom, M., Cucchiarini, C., & Strik, H. (2021). Speech intelligibility of dysarthric speech: human scores and acoustic-phonetic features. In *Proceedings of Interspeech 2021,* 2911-2915.

5.1 Introduction

Speech intelligibility is an important construct in speech pathology that is employed for diagnosis and for determining whether speech therapy has been effective. A common definition in the clinical practice of speech therapy is that proposed by Hustad (2008) "Intelligibility refers to how well a speaker's acoustic signal can be accurately recovered by a listener" (p. 562). In line with this definition, intelligibility has been measured by asking listeners to make orthographic transcriptions (OTs) of what they hear (Yorkston & Beukelman, 1981; Garcia & Cannito, 1996). Percentages of words correctly transcribed are then used as an intelligibility measure as in the Sentence Intelligibility Test (Yorkston et al., 1996a).

In the clinical field, intelligibility has also been measured by collecting scalar ratings from human listeners (Abur et al., 2019; Barreto & Ortiz, 2008; Ishikawa et al., 2020; Miller, 2013; Yorkton & Beukelman, 1978). For instance, by asking listeners to indicate the degree of intelligibility on an equal-appearing interval scale (or Likert scale; e.g., Yorkton & Beukelman, 1978), or a visual analogue scale (VAS; placing a point on a horizontal line; e.g., Finizia et al., 1998). It is common practice to check the reliability of these kinds of ratings before they can be used for research purposes (Beijer et al., 2014).

In a previous study (Ganzeboom et al., 2016), scalar ratings and OTs were used to obtain intelligibility scores of disordered speech at three different levels of granularity: utterance, word, and subword level. Utterance-level evaluations were obtained using subjective rating scales (VAS and Likert scale); word- and subword-level evaluations were obtained from orthographic transcriptions, using automatic alignment and conversion methods. The phoneme scoring thus obtained turned out to be feasible and reliable, and provided a more sensitive and informative measure of intelligibility. The results showed that the intelligibility measures at the different levels of granularity were fairly highly correlated, but performed differently. The orthography-based measures revealed higher intelligibility scores than the scalar rating measures. This appeared to be in line with previous research (Abur et al., 2019; Hustad, 2006), suggesting that in scalar ratings experts tend to underestimate the degree of intelligibility. Even in the case that listeners understand every word, they may still judge

intelligibility as less than perfect because of imprecisions and deviations that make the speech difficult to understand.

An important aspect of the study presented by Ganzeboom et al. (2016) was that all measurements were based on subjective scores provided by human listeners, which of course is in line with the definitions of the constructs themselves. However, it is known that these human-based procedures are subjective, error-prone, and extremely time-consuming, thus making intelligibility measurement extremely problematic in practice. For these reasons, researchers have been studying alternative ways of measuring intelligibility that do not rely on human judgments. Many have employed ASR algorithms to obtain automatic measures of pathological speech quality (Kim et al., 2015; Middag et al., 2009a; Schuster et al., 2006a). However, it is not clear how exactly these ASR-based measures are related to speech intelligibility and to properties of pathological speech that can be addressed in speech therapy.

Previous studies have shown that pitch (Feenaughty et al., 2014; Tjaden & Wilding, 2004, 2011), intensity (Bunton et al., 2000; Cannito et al., 2012; Holmes et al., 2000; Tjaden & Wilding, 2004, 2011), and formant frequencies (Feenaughty et al., 2014; Tjaden & Wilding, 2004; Weismer et al., 2001) are related to intelligibility and can contribute to distinguishing dysarthric speech from healthy speech. For example, Parkinson's patients have limited pitch and loudness variability (Holmes et al., 2000) in their voices. Their intelligibility can be improved by the Lee Silverman Voice Treatment (Cannito et al., 2012), in which speakers are instructed to speak louder. Bunton et al. (2000) found that a restricted intensity range tended to be associated with reduced speech intelligibility in amyotrophic lateral sclerosis speakers with moderate intelligibility.

Moreover, past research has shown that intelligibility measures at different levels of granularity turned out to perform differently (Ganzeboom et al., 2016; Xue et al., 2020, 2021b). Therefore, it is worth investigating the correlations of the features with intelligibility measures at different levels of granularity.

In the present research, we investigated speech intelligibility in dysarthric and non-dysarthric speakers as measured by VAS ratings and OTs and compared these measures to acoustic-phonetic correlates to determine

to what extent these a) can provide a basis for more objective evaluations of speech intelligibility and b) contribute to speaker classification.

5.2 Method

5.2.1 Datasets and speakers

We used two datasets collected within the CHASING project[7] (Ganzeboom et al., 2018) which was aimed at developing a serious game for conducting research on speech disorder treatment through ASR-based technology. The two datasets (Ganzeboom et al., 2018) contain speech of thirteen dysarthric speakers (10 male and 3 female) and of five non-dysarthric speakers (4 male and 1 female). The dysarthric speakers were aged between 53 and 75 (M = 64.2, SD = 6.4), ten of them had Parkinson's and three had had a Cerebral Vascular Accident (CVA). All non-dysarthric speakers were aged between 61 and 69 (M = 65.0, SD = 3.4).

5.2.2 Speech materials

From the datasets described in Section 5.2.1, we extracted, for each speaker, eight recordings covering three types of speech materials: four meaningful sentences, two semantically unpredictable sentences, and two three-word lists. The four meaningful sentences, were selected from the text "Papa en Marloes" ("Papa and Marloes" in English). We made sure that all the recordings were made before the game to avoid any impact of the treatment.

5.2.3 Listeners, intelligibility measures, and experimental setting

We recruited eleven speech therapists as listeners to participate in the listening experiment. All the selected listeners are native speakers of Dutch, have normal hearing and vision and do not have the Attention Deficit Disorder or problems with fine motor skills (typing). They were all graduated as speech therapists being either all-round speech therapists or specialized in neurorehabilitation. Seven of them had experience in dysarthria. In detail,

[7] http://hstrik.ruhosting.nl/chasing/

four of them followed the master speech-language pathology of Radboud University, two had training in Parkinson's and dysarthria, and one had followed a minor in neurorehabilitation. The number of years after they had graduated as speech therapists varied between 0.5 and 15 years (M = 3.6, SD = 5.0). They were all female and were aged between 23 and 38 years (M = 26.9, SD = 4.7).

In the listening experiment, each listener assigned a score through a Visual Analogue Scale (VAS) ranging from 0 (0% intelligible) to 100 (100% intelligible) and completed an Orthographic Transcription (OT) of the utterance allowing only existing words. Based on the transcriptions, we computed word accuracy (AcW), which is the percentage of correctly transcribed words[8]. Therefore, for each utterance, we obtained two intelligibility measures, a rating (VAS) and a word accuracy score (AcW).

During the experiments, the recordings to be assessed were presented with a loudness of 60 decibels through headphones. The researcher who collected the data of the intelligibility measures was always present to make sure that the recordings were presented with the same sound intensity and the experiments took place in a low-noise environment.

5.2.4 Acoustic-phonetic features

The acoustic-phonetic features were calculated using Praat (Boersma & Weenink, 2021). The features extracted are duration, minimal pitch (pitchMin), maximal pitch (pitchMax), mean value of pitch (pitchMean), standard deviation of pitch (pitchStd), the mean of the absolute values of the pitch slope (pitchSlopeMean), minimal intensity (intensityMin), maximal intensity (intensityMax), mean value of intensity (intensityMean), standard deviation of intensity (intensityStd), formants 1 to 4 (F1, F2, F3 and F4), and center of gravity (centerGravity).

5.2.5 Statistical analysis

We first explored the two intelligibility measures with respect to their distribution and interrater reliability at the utterance level. Interrater

[8] We used the asr-evaluation python module provided in https://github.com/belambert/asr-evaluation to calculate the word accuracy.

reliability was calculated by applying Generalizability Theory (Brennan, 2001) with a fully crossed model design, where all utterances from all speakers were assessed by all listeners. The Phi coefficient (D-coefficient) was then calculated as the reliability.

We calculated the means and standard deviations of the acoustic-phonetic features for the two types of speakers separately and conducted a *t*-test to establish whether the mean values of the acoustic-phonetic features were significantly different for the two types of speakers.

In the next step, we averaged our two intelligibility measures over utterances and listeners to obtain a score for each speaker. We then computed their reliabilities using Intraclass Correlation Coefficient (ICC) and their correlation. We made a scattergram distinguishing the dysarthric and non-dysarthric speakers.

The acoustic-phonetic features were submitted to a stepwise logistic regression[9] to predict speaker type (set '0' for 'non-dysarthric' and '1' for 'dysarthric') at the utterance level. The Akaike Information Criterion (AIC) was used in the stepwise procedure to decide which acoustic-phonetic features to include in the final regression model. The predicted speaker-type probabilities at the utterance level were averaged over the eight utterances per speaker as the overall acoustic-phonetic probability index (API). According to the above speaker type setting, high API scores of test items refer to 'dysarthric' and low scores refer to 'non-dysarthric'. The next step was to explore and interpret the correlations and scattergrams of the API and the two intelligibility measures. We used the packages *gtheory* (Moore, 2016), *stats* (R Core Team, 2020), *psych* (Revelle, 2019), *MASS* (Venables & Ripley, 2002), and the *ggplot2* (Wickham, 2016) for performing the analyses and making plots in RStudio (RStudio Team, 2020) with R version 4.0.1 (R Core Team, 2020).

5.3 Results

5.3.1 Intelligibility measures at the utterance level

Figure 5.1 shows the density plots for VAS and AcW, with different colors for the two speaker types. VAS scores have a broad distribution (M = 59.6, SD =

[9] We obtained similar models in forwards and backwards regression.

23.6), with the distribution of the non-dysarthric utterances having higher frequencies on the right part of the scale. AcW scores have an extremely skewed distribution (M = 95.7, SD = 11.3) with a very high concentration of maximum scores of 100. Both plots show an overlap between the two speaker types. The reliability (D-coefficient) values at the utterance level are 0.90 for VAS and 0.47 for AcW.

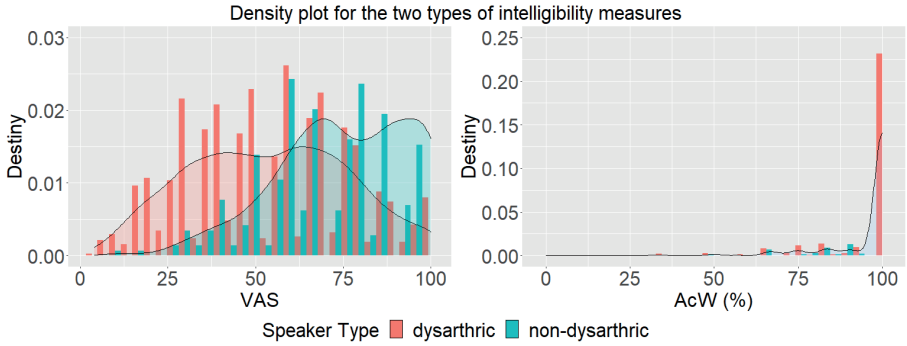


Figure 5.1. Density plots for the two intelligibility measures. **Best viewed in color.**

5.3.2 Acoustic-phonetic features at the utterance level

No significant difference in the mean values of the acoustic-phonetic features was found between the two types of speakers. The detailed results regarding the mean and standard deviation of the features for each speaker type can be found in Table 5.1.

5.3.3 Dysarthric and non-dysarthric speakers

5.3.3.1 Intelligibility measures

The reliability values using ICC (2, k) (items and listeners random with k = 11) are 0.93 for VAS and 0.85 for AcW. Figure 5.2 shows the scattergram between VAS and AcW, as well as the regression line. It can be seen that the points for non-dysarthric speakers are located close to the top-right corner, while those for dysarthric speakers are relatively scattered to the bottom-left. However, the ranges between the two measures differ considerably. The VAS scores range between 52 and 82, while the AcW scores range between 88 and 99. The VAS distinguishes dysarthric and non-dysarthric

speakers reasonably well, with an overlap in the higher regions of the VAS scores. AcW is not successful in distinguishing the two types of speakers. The correlation between the measures (.53) is significant and moderate.

Table 5.1. Mean (Standard Deviation) of the acoustic-phonetic features for the two types of speakers.

Feature	Non-dysarthric	Dysarthric
duration (ms)	3575.38 (2115.36)	3640.38 (2090.28)
pitchMin (Hz)	86.11 (15.57)	92.6 (17.63)
pitchMax (Hz)	289.98 (173.51)	279.5 (156.68)
pitchMean (Hz)	131.11 (27.45)	139.14 (22.82)
pitchStd (Hz)	40.58 (30.81)	34.41 (25.92)
pitchSlopeMean (Hz)	361.91 (353.02)	341.66 (236.26)
intensityMin (dB)	19.59 (10.41)	14.49 (36.15)
intensityMax (dB)	77.19 (5.9)	81.09 (4.3)
intensityMean (dB)	66.82 (6.83)	71.22 (4.52)
intensityStd (dB)	15.99 (4.97)	16.64 (5.99)
F1 (Hz)	756.92 (472.12)	805.55 (482.52)
F2 (Hz)	1793.3 (487.57)	1902.21 (466.82)
F3 (Hz)	2915.68 (402.85)	2934.33 (412.66)
F4 (Hz)	3913.72 (418.45)	3952.35 (431.59)
centerGravity (Hz)	685.12 (243.93)	533.5 (186.43)

5.3.3.2 Acoustic-phonetic features

Table 5.2 shows the summary of the final model of the stepwise logistic regression. All variables in Table 5.2 were relevant according to the AIC criterium. These results indicate that the chances of being dysarthric increase as pitchSlopeMean, intensityMax, F1, and F2 increases and decrease as pitchStd, intensityMin, intensityStd, F3, and centerGravity increase. Six of the variables are significant in the classification. The final model also reported a mean classification accuracy of 79% with 50% for non-dysarthric and 88% for dysarthric speech.

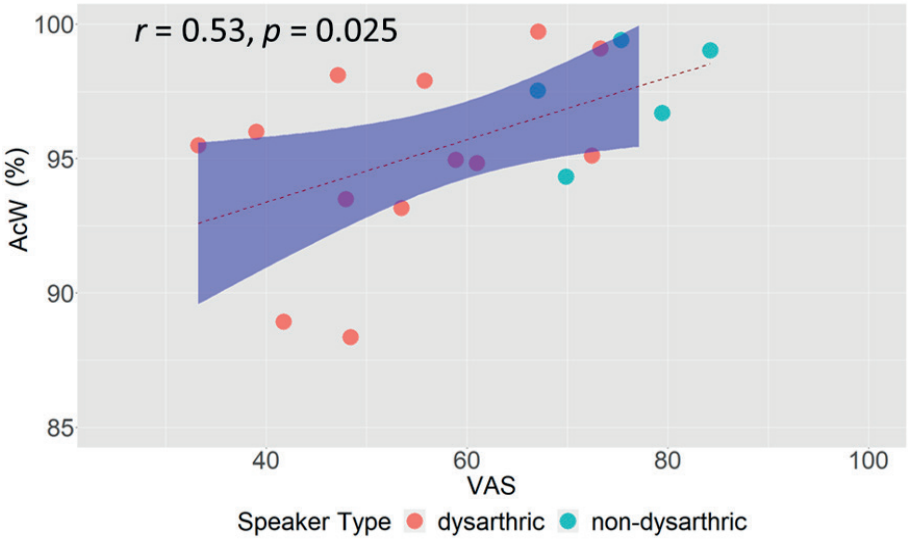


Figure 5.2. Scattergram of VAS and AcW (%), including the correlation *r*, *p* value, the regression line, and confidence intervals (95%). **Best viewed in color.**

Table 5.2. Summary of the final model in stepwise logistic regression with selected predictors' coefficients (Coef.), standard errors of the coefficients (Std. Err), the *z* value, the *p* value, and the significance levels (Sig. level; 0.001 – '***'; 0.01 – '**'; 0.05 – '*'; 0.1 – '.'; 1 – ' ').

Predictors	Coef.	Std. Err	*z* value	*p* value	Sig. Level
pitchStd	-0.039	0.0145	-2.074	0.012	**
pitchSlopeMean	0.0049	0.00157	3.135	0.00685	**
intensityMin	-0.011	0.0372	-2.943	0.00172	**
intensityMax	0.283	0.0648	4.375	0.00326	***
intensityStd	-0.2158	0.0768	-2.812	0.005	**
F1	0.00124	0.000765	1.624	0.104	
F2	0.00092	0.000631	1.458	0.145	
F3	-0.001627	0.000859	-1.893	0.058	.
centerGravity	-0.00363	0.0013	-2.784	0.00537	**

5.3.3.3 Correlating the intelligibility measures and the acoustic-phonetic probability index (API)

As explained in Section 5.2.5, the predicted speaker type probabilities at the utterance level were averaged over the eight utterances per speaker to compute the API. Figure 5.3 shows the scattergrams between the API, VAS, and AcW, with different colors for the two speaker types. It can be seen that the API makes a distinction between the two types of speakers, but that there is a clear overlap as well. The correlation of API with VAS is significant and moderate, while the correlation with AcW is non-significant.

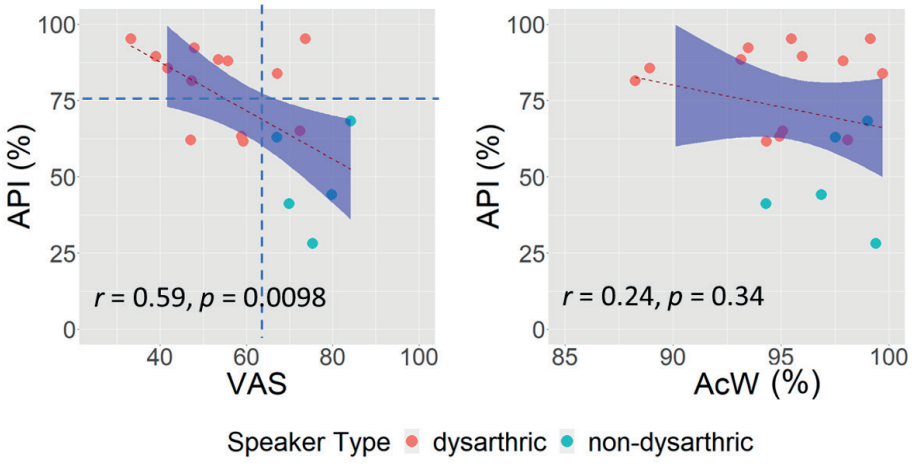


Figure 5.3. Scattergrams of VAS and AcW (%) with the API (%), including the correlation *r*, *p* value, the regression line, and confidence intervals (95%). VAS-API plot is split into four quadrants by dashed lines. **Best viewed in color.**

Interestingly, we can split the VAS-API plot into four quadrants along a vertical and a horizontal dashed line, as shown in the left panel of Figure 5.3. Most of the speakers are classified correctly as located in the top-left quadrant for 'dysarthric' (with high API scores and low VAS scores) and bottom-right quadrant for 'non-dysarthric' (with low API scores and high VAS scores) with only one red spot (as 'dysarthric') in the bottom-right quadrant. Divergent results were found for several dysarthric speakers. Specifically, we found three dysarthric speakers in the bottom-left quadrant. These speakers had relatively low VAS scores but were not classified as

'dysarthric' according to the low API scores. The opposite applies to two dysarthric speakers in the top-right quadrant, with high VAS scores and high API scores.

5.4 Discussion

In this study, speech of dysarthric and non-dysarthric speakers was evaluated on intelligibility by human listeners through VAS scalar ratings and word accuracy (AcW) based on Orthographic Transcriptions (OTs). Acoustic analyses were then conducted to study how acoustic-phonetic features are related to intelligibility measures and to what extent they contribute to classifying speakers as either dysarthric or non-dysarthric.

We observed high interrater reliability for the VAS ratings, which is in line with previous findings (Ganzeboom et al., 2016; Ishikawa et al., 2020; Miller, 2013; Xue et al., 2020). On the other hand, for AcW we found relatively low interrater reliability at the utterance level, which contrasts with previous findings (Miller, 2013; Xue et al., 2020; Tjaden, & Wilding, 2010). We further explored the data to gain insight into the possible causes of this low reliability index. We noted that most of the word accuracy scores indicate perfect intelligibility of 100, which implies that the variability in the scores was very limited. This seems to be plausible, as the speakers had mild dysarthria and thus displayed relatively little variability in their speech. In addition, only existing words were allowed in the transcriptions, which in turn further reduced the degree of variability. So, this would seem to suggest that the listeners did their job properly and transcribed the speech correctly. We consequently found that the correlation between the two intelligibility measures is significant, but not strong, partly due to this low variability.

The results presented above suggest that the two intelligibility measures investigated in this study are related, but have distinct qualities, and that they might refer to different constructs of the concept of intelligibility, even for the same speech material. Similar findings in the field of L2 pronunciation instruction, where speech intelligibility has also received considerable attention, led researchers (Munro & Derwing, 1995) to draw a distinction between speech intelligibility defined as "the extent

to which a speaker's message is actually understood by a listener" (p. 76) and comprehensibility, which stands for the degree of ease with which L2 speech can be understood. Accordingly, different operationalizations were adopted for the two constructs. To measure intelligibility listeners had to "write out carefully in standard orthography what they heard" (Munro & Derwing, 1995), while for comprehensibility listeners had to assign scale ratings.

In the clinical field, the term comprehensibility has also been used, but with different definitions like "contextual intelligibility" (Yorkston et al., 1996b) or a listeners' ability to recover the meaning of pathological speech utterances (Hustad & Beukelman, 2002). The results presented in this chapter might be seen as indications that also in clinical practice two constructs should be distinguished: intelligibility refers to the degree of actual understanding as measured by writing down what has been said, and comprehensibility refers to the ease of understanding as measured through rating scales.

The mean values of the acoustic-phonetic features did not differ significantly for the two types of speakers, suggesting that no single feature is highly related to speaker type. The logistic regression outcomes show that the most important and promising point seems to be that the three groups of features play a role in the final regression model selected in distinguishing dysarthric and non-dysarthric utterances: pitch, intensity, and formant frequencies. When we studied the correlations between the intelligibility measures on the one hand, and the overall acoustic-phonetic probability index (API) that a speaker is classified as either dysarthric or non-dysarthric, on the other, we found that the API showed moderate correlation with VAS and no correlation with AcW. With respect to the classification of speakers as either dysarthric or non-dysarthric, the results suggest that the intelligibility measures assigned by human listeners and the probabilities computed through the objective procedure based on acoustic-phonetic features are partly complementary to each other, as also found by Bunton et al. (2000). These results are also in line with previous findings that acoustic-phonetic features have correlations to speaker types or to speech intelligibility (Bunton et al., 2000; Tjaden & Wilding, 2004; Xue et al., 2019), to a certain extent.

Future research could investigate the precise impact of the individual acoustic-phonetic features in more detail and with more utterances from speakers with varied dysarthria severity. Additionally, other relevant acoustic-phonetic features such as F2 slope (Chiu et al., 2019; Tjaden & Wilding, 2004) and vowel space area (Weismer et al., 2001), and more robust metrics of features, such as percentiles could be considered as these might reveal different relations to speech intelligibility or may better contribute to classifying speaker types.


Meow!
Mow?

CHAPTER 6

ACOUSTIC CORRELATES OF INTELLIGIBILITY – THE USABILITY OF THE EGEMAPS FEATURE SET

ABSTRACT

Although speech intelligibility has been studied in different fields such as speech pathology, language learning, psycholinguistics, and speech synthesis, it is still unclear which concrete speech features most impact intelligibility. Commonly used subjective measures of speech intelligibility based on labor-intensive human ratings are time-consuming and expensive, so objective procedures based on automatically calculated features are needed. In this chapter, we investigate possible correlations between a set of acoustic features and speech intelligibility. Specifically, we study the usability of acoustic features in the eGeMAPS feature set for predicting Phoneme Intelligibility by using stepwise linear multiple regression analysis. The results showed that the acoustic features are potentially usable for predicting intelligibility. This finding may help to boost the development of automatic procedures to measure speech intelligibility with the underlying relevant acoustic-phonetic characteristics. Our analysis also covers the comparison between two speech types (dysarthric and normal), and between two different types of speech material (isolated words and running text). Finally, we discuss possible avenues for future research on speech intelligibility and implications for clinical practice.

This chapter is based on the following publication:

Xue, W., Cucchiarini, C., van Hout, R. & Strik, H. (2019). Acoustic correlates of speech intelligibility: the usability of the eGeMAPS feature set for atypical speech. In *Proceedings of SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education*, 48-52.

6.1 Introduction

Speech intelligibility is an important construct that has been studied in different fields like speech pathology (Hustad, 2008), second language (L2) pronunciation (Munro & Derwing, 1995), speech synthesis evaluation (Benoit et al., 1996; Gibbon et al., 1997), speech perception (Cutler et al., 2008; Cooke et al., 2013), and telecommunication, but from different perspectives. For instance, in telecommunication, speech intelligibility is defined in relation to the lossless of the transmission channel, while speech pathology mainly focuses on speaker-related properties such as the presence of a speech disorder. Hence, the approaches to measuring speech intelligibility are different as well. Telecommunication studies use measures such as the Speech Transmission Index (Steeneken & Houtgast, 1980), the diagnostic rhyme test (Voiers et al., 1973), and the modified rhyme test (House et al., 1965). Speech pathological research uses tests such as the Dutch Intelligibility Assessment (DIA; Middag et al., 2009b) and the Sentence Intelligibility Test (Yorkston et al., 1996a). In this chapter, we are interested in exploring the relation between acoustic features and the intelligibility of speech in which the intelligibility may be reduced by speaker-related properties, in particular by having dysarthria.

A common definition of speech intelligibility in the clinical practice of speech therapy was proposed by Hustad (2008) "Intelligibility refers to how well a speaker's acoustic signal can be accurately recovered by a listener" (p. 562). Munro and Derwing (2015) suggested a similar definition in L2 pronunciation research "the extent to which listeners' perceptions match speakers' intentions" (p. 14). According to these definitions, speech intelligibility cannot "be evaluated without some sort of reference to listener data" (Munro & Derwing, 2015, p. 13). As a result, measuring speech intelligibility has conventionally resorted to human listeners by asking them to provide ratings of intelligibility in different ways (Barreto & Ortiz, 2008; Miller 2013; Yorkston & Beukelman, 1978). Common procedures are to ask listeners to express scalar judgments on the degree of intelligibility of speech samples by using an equal-appearing interval scale like the Likert scale (Yorkston & Beukelman, 1978), or by using a horizontal line, on which a point is placed to indicate intelligibility, like a visual analogue scale (VAS; Finizia et al., 1998). Moreover, transcribing verbatim what listeners hear (Munro & Derwing, 1995; Hustad, 2006; Laures & Weismer, 1999) and

indicating how well individual phones in isolated words were realized (Van Nuffelen, 2009a; Middag et al., 2009a) were also explored.

In order to alleviate the effect of subjectivity in the above methods, ratings are usually collected from multiple listeners and then averaged for further analyses. Besides, reliability measures are also needed. Research has shown that these operations can help obtain reliable ratings (Ganzeboom et al., 2016) and, in fact, have been widely used in research and clinical practice. However, these operations are generally time-consuming and costly, and the need for multiple listeners makes this practice even more laborious and expensive. In addition, while it may be feasible to apply these rating procedures in a research context, they are still problematic in clinical practice, where easy-to-use tools are strongly preferred.

For these reasons, there is a need for valid procedures to obtain objective measures of intelligibility in an automated way that does not rely on intensive human efforts. Several researchers have employed ASR-based algorithms and ASR-free features to obtain automatic measures of pathological speech quality, which have been shown to be strongly correlated with human-based measures of speech intelligibility (Schuster et al., 2006a; Middag et al., 2009a; Berisha et al., 2013; Pellegrini et al., 2015; Kim et al., 2015). In the language learning domain, such objective and automated approaches to measuring speech intelligibility have not been undertaken, although the need is felt there as well (O'Brien et al., 2019). So far, it is not clear how exactly these ASR-based and ASR-free measures are related to properties of pathological speech that can be addressed in therapy, and whether these measures could be used to develop easy-to-use tools for clinical practice.

In this chapter, we investigate the usability of a new set of acoustic features called the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS; Eyben et al., 2015), which was recently developed for research in different areas of speech analysis. Specifically, we apply it to the prediction of speech intelligibility. The advantage of using this feature set is that it consists of a standardized, limited set of features, which were chosen based on their demonstrated theoretical relevance, their potential to distinguish important aspects of speech production, and their ease in the automatic calculation. As the authors suggest, the idea is "to provide a common baseline for evaluation of future research and eliminate differences caused by varying

parameter sets or even different implementations of the same parameters" (Eyben et al., 2015, p.190). The implementation is realized through the open-source openSMILE toolkit (Eyben et al., 2013), which guarantees standardized calculation of all parameters. In addition, we also included speech rate-related features to investigate if these features are complementary to the eGeMAPS feature set for predicting speech intelligibility.

The question we address in this chapter is to what extent these automatically calculated features are related to intelligibility scores of pathological and normal speech. If this relation is strong, then it might be also worthwhile to investigate whether these features can be employed to measure intelligibility in other types of atypical speech like L2 speech and to provide the possibility of developing an easy-to-use tool for clinical practice.

An important question in this connection is which intelligibility scores should be taken as the point of reference. As explained above, different types of human ratings have been employed in literature, and these often vary with respect to their degree of detail (Ganzeboom et al., 2016). In addition, not all sorts of human ratings are available for all speech databases. In other words, the choice of the speech database usually also determines which human ratings are to be used because they are the only ones available in that specific speech corpus. In this study, we chose to use the COPAS database (Middag, 2012) because it has many advantages, as will be explained in Section 6.2.1. In this database, the intelligibility scores of a speaker were calculated based on phoneme-level ratings and then averaged per speaker. These intelligibility scores will be taken as the point of reference to evaluate the automatically predicted scores based on eGeMAPS acoustic features.

The rest of the chapter is organized as follows. We first describe the database and our methods in Section 6.2. The experimental results are presented in Section 6.3, while their implications are discussed in Section 6.4, and the conclusions are provided in Section 6.5.

6.2 Method

6.2.1 Speakers and speech materials

The speech materials used in the study were selected from the COPAS database which was collected and used to develop a reliable ASR-based

speech assessment tool for pathological speech within the framework of the SPACE project. This speech database contains recordings, as shown in the COPAS document (Middag, 2012), collected from 197 pathological speakers with speech disorders such as dysarthria, cleft, articulation disorders, voice disorder, laryngectomy, and glossectomy and from a control group of 122 normal speakers whose speech is not disordered. The speech materials included in the COPAS dataset cover not only isolated words but also isolated sentences and short passages. In this chapter, we consider two of the materials, leading to two tasks: word reading and passage reading. The word reading task is the Dutch Intelligibility Assessment (DIA; De Bodt et al., 2006) with isolated-word material which contains 35 versions of 50 consonant-vowel-consonant (CVC) words and pseudowords organized in three subsets (i.e., subset A, B, and C). The three subsets were originally designed to evaluate intelligibility by assessing the initial, final, and central phonemes of isolated words as the target phonemes, respectively, which were also used in Chapters 2 through 4. The passage reading task is the phonetically balanced text known as Text Marloes (TM; Van de Weijer & Slis, 1991).

The subjects were selected on the basis of participation in both the DIA and TM tasks. There were 20 female and 29 male dysarthric speakers, and 48 female and 33 male normal speakers, giving a total of 49 dysarthric and 81 normal speakers.

6.2.2 Intelligibility ratings

The DIA task consists of three subsets containing different words designed for assessing specific phonemes (Middag, 2012). The Phoneme Intelligibility on each subset is achieved by asking the therapist whether the target phoneme was correct or not. Then the overall Phoneme Intelligibility of each speaker was calculated by averaging the percentages of correctly perceived target phonemes in these three subsets and used as the intelligibility score on both the DIA and TM tasks.

The frequency plot of the total Phoneme Intelligibility of the dysarthric and normal speakers is shown in Figure 6.1. As can be seen in Figure 6.1, the normal speakers and the dysarthric speakers showed an overlap in Phoneme Intelligibility from 82 to 100. The dysarthric speakers showed more variation in Phoneme Intelligibility than the normal speakers.

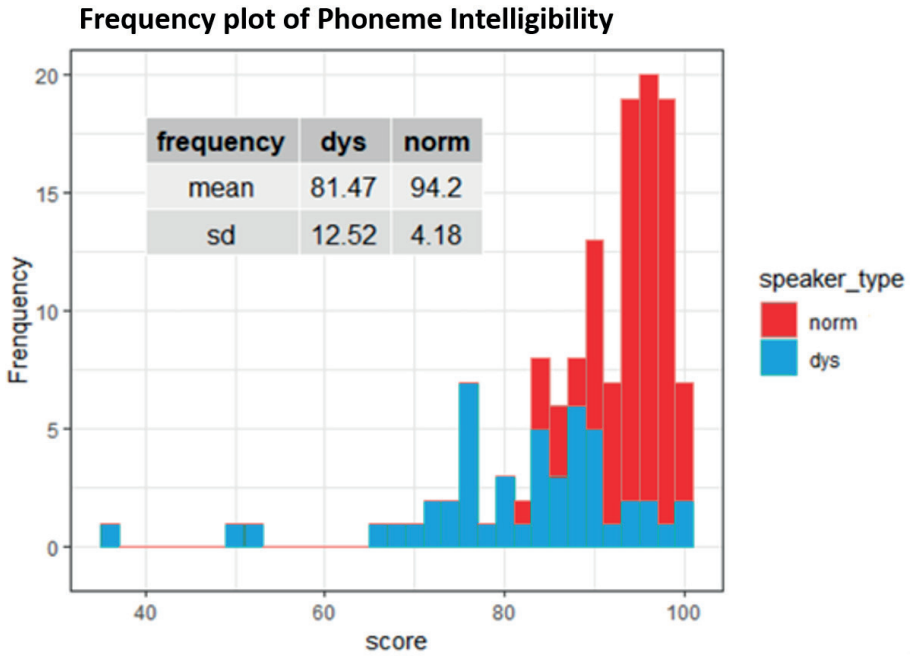


Figure 6.1. Frequency plot of Phoneme Intelligibility of dysarthric (dys) and normal speakers (norm) with a frequency table. **Best viewed in color.**

6.2.3 Automatic acoustic measures

In order to investigate possible acoustic correlates of speech intelligibility, we chose the eGeMAPS feature set because of its wide coverage of standardized relevant acoustic features that can be measured automatically. There are many acoustic features that might be correlated to speech intelligibility such as measures of voice (loudness, fundamental frequency F0, jitter) and articulation (Mel-frequency cepstral coefficient 1-4, vowel formant frequencies). More detailed descriptions can be found in Eyben et al. (2015). Specifically, the selected features were calculated for all the speech recordings using the default configuration file in openSMILE (Eyben et al., 2013). In addition, a Praat script (Boersma, 2001) was used to calculate a set of speech rate-related features such as the number of syllables, the number of pauses, phonation time, speech rate, articulation rate, and average syllable duration by automatically detecting syllable nuclei without using segmentation (De Jong & Wempe, 2009). This complementary feature

set may identify additional parameters related to intelligibility. All of the features in all subsets were averaged for each speaker in the regression analyses in predicting intelligibility.

6.2.4 Regression analysis

In order to investigate the correlation between acoustic features and speech intelligibility, a stepwise linear multiple regression (SLMR) algorithm (Efroymson, 1960; *p* in = 0.05, *p* out = .10) was applied to predicting Phoneme Intelligibility using the eGeMAPS features. This process was denoted as SLMR_1. Besides, we applied the second round of SLMR (SLMR_2) on the combination of eGeMAPS features and the speech rate-related features, to predict the residual values of SLMR_1. This was aimed to investigate whether speech rate-related features can complement the eGeMAPS features for better predicting the intelligibility scores. For each of the two speaker groups, this procedure was applied to the whole DIA task and TM separately. For DIA, the experiments were also conducted for three subsets separately. All of the features were normalized. The stepwise linear multiple regression was implemented by using the ols_step_both_p function in the R *olsrr* package (Hebbali, 2018).

6.3 Results

6.3.1 Correlation results

To explore the correlation between acoustic features and speech intelligibility, we first calculated the correlations between Phoneme Intelligibility and the eGeMAPS features of the DIA and TM tasks, respectively.

Table 6.1 shows the range of the ten highest correlations between Phoneme Intelligibility and eGeMAPS features and the overlaps of them between each pair of speaker groups. The highest correlation on the DIA task was found for dysarthric speech. We also found that only a small number of features in the ten highest correlated features were shared between each pair of speaker groups.

Table 6.1. Correlation ranges of the ten highest correlations between Phoneme Intelligibility and eGeMAPS features for DIA and TM tasks and the overlaps between each two speaker groups (dys.: dysarthric, norm.: normal, and comb.: combined).

Correlation	DIA	TM
Range dysarthric	0.294–0.492	0.266–0.395
Range normal	0.311–0.409	0.334–0.404
Range combined	0.210–0.423	0.313–0.528
Overlap dys. – norm.	0	2 features
Overlap comb. – dys.	3 features	3 features
Overlap comb. – norm.	1 feature	0

6.3.2 SLMR results on the DIA task and Text Marloes

The residual and fitted value plots of SLMR_1 for the dysarthric and normal speech on the DIA and TM tasks are shown in Figure 6.2. Normal speakers showed similar patterns on both tasks, while larger residual values were found for dysarthric speakers on TM.

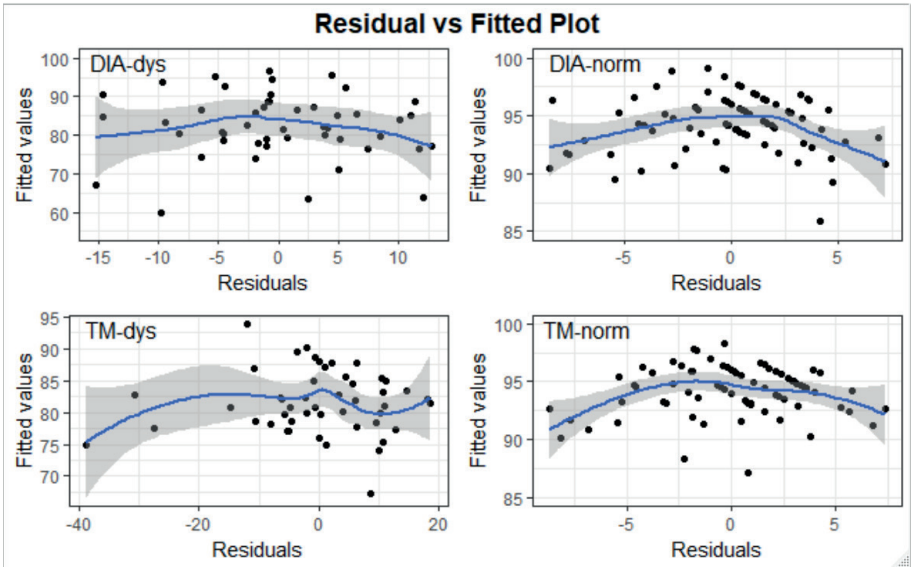


Figure 6.2. Residuals and fitted values plot of SLMR_1 for dysarthric (dys) and normal speech (norm) on the DIA and TM tasks. **Best viewed in color.**

Table 6.2 shows the multiple R-squared scores, which represent the proportion of the variance of the dependent variable that could be explained by the explanatory variables(s), of the final model in SLMR_1 on the whole DIA and TM tasks. The highest score was achieved for dysarthric speakers on DIA. The multiple R-squared score for the combined speakers on TM was larger than that of each separate speaker type.

Table 6.2. The multiple R-squared scores of the final model in SLMR_1 for three groups of speakers (dysarthric, normal, and combined speakers) on DIA (whole and three subsets A, B, and C) and TM (higher is better).

Multiple R-squared		**Dysarthric**	**Normal**	**Combined**
DIA	**Subset A**	0.6238	0.3712	0.2880
	Subset B	0.5011	0.2828	0.2669
	Subset C	0.7565	0.2821	0.6029
	Whole	0.6823	0.3297	0.3539
TM		0.1559	0.3318	0.4309

The numbers of selected features in the final models in SLMR_1 are shown in Table 6.3. The largest number of selected features was achieved for dysarthric speakers on DIA. The overlap between any two speaker groups was very small regardless of the type of speech material.

Table 6.3. The number of selected features of the final models in SLMR_1 for dysarthric, normal, and combined speakers on DIA (whole and three subsets A, B, and C) and TM tasks plus the overlaps between each two of three speaker groups (o_dys-norm, o_comb-dys, and o_comb-norm).

Selected features	**DIA**				**TM**
	A	**B**	**C**	**Whole**	
Dysarthric	4	5	9	7	1
Normal	4	5	3	4	3
Combine	5	4	14	4	5
o_dys-norm	0	0	0	1	0
o_comb_dys	2	1	3	1	0
o_comb_norm	0	1	1	0	0

Besides, we found that speech rate-related features were selected as predictors of the final model in SLMR_2 for both dysarthric and normal speakers on TM, but not on DIA. So, we applied SLMR with the combined feature set of eGeMAPS features and speech rate-related features for directly predicting the Phoneme Intelligibility on TM. The multiple R-squared scores of the resultant model were 0.3239, 0.4615, and 0.4714 for dysarthric, normal, and combined speakers, respectively. They were all higher than the scores in the models without speech rate-related features.

6.3.3 SLMR results on subsets of the DIA task

The results of SLMR_1 on each subset of the DIA task are shown in Table 6.2. The multiple R-squared scores of dysarthric speakers were all higher than those of normal speakers and showed a larger variation across three subsets. The scores for the combined speakers were smaller than those for each speaker type except in subset C.

6.4 Discussion

The present study was aimed at investigating the usability of the eGeMAPS feature set for predicting Phoneme Intelligibility. The medium-sized correlations between the features and Phoneme Intelligibility evidenced by the results in Table 6.1 but in particular the higher values of the multiple R-squared scores in Table 6.2 suggest that, in principle, these features could be employed to derive objective and automatically calculated measures of Phoneme Intelligibility for normal and dysarthric speech.

When comparing the results for the two types of speech, we see higher correlations for dysarthric speech than for normal speech in the case of isolated words. As shown in Table 6.2, the multiple R-squared scores for the dysarthric and normal speech on the DIA task are 0.6823 and 0.3297, respectively. In addition, we tried a fairer comparison by only considering the dysarthric speakers (a total of 28) whose Phoneme Intelligibility was between 82 to 100, which was the same range as that of the normal speakers. In this case, the multiple R-squared score was still higher than that of the normal speakers (0.5477 vs 0.3297). This suggests that on isolated words,

the eGeMAPS feature set is effective for predicting Phoneme Intelligibility in dysarthric speech, but probably not so effective in normal speech. It is conceivable that Phoneme Intelligibility in these two types of speech relies on different features and that further research is needed to gain more insights into these features.

As for the two different types of speech material, we can see a generally higher correlation for the isolated words than for the running text by a large margin (0.6823 vs 0.1559), as shown in Table 6.2. From Figure 6.2, we can also see that the model for dysarthric speech achieves the largest residual values on TM. These results are not surprising since the applied Phoneme Intelligibility scores, which were calculated based only on isolated words, do not generalize to running text. It could be the case that Phoneme Intelligibility only reveals one aspect of general speech intelligibility for dysarthric speech and that additional measures are required to get a more comprehensive picture.

For normal speech, the variation of the correlation results between these two materials as shown in Table 6.2 is small (0.3297 vs 0.3318). This indicates that the representation of speech intelligibility in normal speech does not vary much across different materials, as opposed to dysarthric speech.

The results of separate analyses for the three subsets of isolated words which addressed initial, central, and final phonemes provided further evidence of the substantial differences between normal speech and dysarthric speech. Specifically, we found that the central phonemes (vowels) play the most important role for dysarthric speech with a correlation score of 0.6238, while for normal speech, the initial consonants contribute the most with a score of 0.3712. This difference between the two speech types is also supported by the results in Table 6.1 and Table 6.3, where the overlap between the features appears to be very small.

After incorporating additional speech rate–related features, the multiple R-squared scores increased from 0.1559 to 0.3239 for dysarthric speech, and from 0.3318 to 0.4615 for normal speech, in the SLMR experiments on TM. This reveals that in the case of running text, speech rate is an important explanatory factor which could be employed to complement the eGeMAPS feature set, while this does not seem to be the case for isolated words.

On the one hand, these results might be seen as an indication that the eGeMAPS feature set does not contain features that are general predictors of Phoneme Intelligibility. On the other hand, it seems too early to draw such a conclusion. It is important to bear in mind that the measure of speech intelligibility adopted in the COPAS database is a very narrow measure of Phoneme Intelligibility. Before drawing conclusions about the usability of the eGeMAPS feature set, it is necessary to conduct further research with other measures of intelligibility at a higher level such as word and/or sentence level.

This study has shown, once again, that speech intelligibility is a complex construct and that further research is needed to get a better understanding of both the human-generated measures of intelligibility, as well as their more objective acoustic correlates. We have seen that it is important to include different types of speech materials, to explore the substantial differences between pathological and normal speech in more detail, and to employ various rating procedures (i.e., Likert, VAS, and orthographic transcriptions). The important insights that derive from this more comprehensive research will not be limited to the clinical domain, but may help us analyze speech intelligibility in the field of L2 pronunciation. The performance of the relations between acoustic features and speech intelligibility of L2 learners' speech may be different from the results in this chapter, but they may also reveal interesting overlaps and generalizations. Again, as we mentioned above, further exploring how the acoustic features are correlated with speech intelligibility is needed and may help researchers to understand the underlying processes and mechanisms that affect speech intelligibility.

6.5 Conclusions

Our experimental results showed moderate to medium correlations between Phoneme Intelligibility scores and the acoustic features in the eGeMAPS feature set, which seems promising for automatic speech intelligibility prediction. Our analyses also revealed important differences between dysarthric speech and normal speech, and between different types of speech material (isolated words and running text). These results indicate

new avenues for future research that are likely to benefit both the clinical and the language learning domains.


Meow!
Mow?

CHAPTER 7

MEASURING SPEECH INTELLIGIBILITY OF DYSARTHRIC SPEECH THROUGH AUTOMATIC SPEECH RECOGNITION IN A PLURICENTRIC LANGUAGE

ABSTRACT

Speech intelligibility is an essential though complex construct for evaluating dysarthric speech. Various procedures can be used to measure speech intelligibility, most of which are based on subjective ratings assigned by experts. Since these procedures are subjective and laborious, ASR has been proposed to obtain objective metrics of intelligibility. Although promising results have been reported, reliable ASR generally requires large amounts of data consisting of recorded and annotated speech. However, speech resources of dysarthric speech are inadequate, referring to the low-resource problem. In the present study, we explored the possibility of using dysarthric speech resources from the dominant language variety to improve the performance of ASR systems on the dysarthric speech of the non-dominant variety of the same pluricentric language. Dutch is used as an example of a pluricentric language, with Netherlandic Dutch considered the dominant and Flemish Dutch the non-dominant variety. The performance of ASR is evaluated by using two types of intelligibility metrics: orthographic transcriptions and global intelligibility assessments, both obtained from experts. Overall, the results show that dysarthric speech data from the dominant language variety can contribute to improving automatic human-comparable transcriptions and to developing objective, automatic global measures of speech intelligibility only when no data from the non-dominant variety are available for training ASR models.

This chapter is based on:

Xue, W., Cucchiarini, C., van Hout, R., & Strik, H. (submitted). Measuring Speech Intelligibility of Dysarthric Speech through Automatic Speech Recognition in a Pluricentric Language.

7.1 Introduction

Patients suffering from dysarthria, a motor speech disorder caused by neurological injury such as Parkinson's disease or stroke, experience difficulties in speech production that can lead to reduced speech intelligibility. Intensive speech therapy can be provided to limit the loss in speech intelligibility or even achieve some improvement. Important instruments in this process are reliable and valid measurements of speech intelligibility to establish a diagnosis and to evaluate therapy effectiveness. So far, the usual metrics for measuring speech intelligibility in research and clinical practice are based on subjective judgements obtained from human listeners. Because these metrics not only contain an element of subjectivity but are also rather time-consuming, researchers have been looking for more objective metrics based on acoustic measurements, possibly obtained semiautomatically or automatically. One of the technologies that have been used for this purpose is Automatic Speech Recognition (ASR; Rosen & Yampolsky, 2000). Although promising results have been obtained, it is clear that applying ASR for this specific purpose requires large amounts of data (Keshet, 2018) consisting of recordings of patients' speech with annotations such as orthographic transcriptions. However, obtaining such data is both ethically challenging and practically laborious. By ethically challenging, we refer to the difficulties in acquiring ethical approval and finding enough people willing to be recorded for academic research. These consequences are more severe in the case of languages that have relatively smaller resources like Dutch, which in general have fewer speech resources to train and test ASR algorithms than a large-size language like English.

In fact, both Dutch and English are pluricentric languages, which are defined as languages with distinct varieties belonging to different countries. These varieties can be expressed as dominant and non-dominant varieties according to their power asymmetries (Norrby et al., 2020). An interesting course of action that has been proposed for improving speech recognition on less-resourced languages that have multiple varieties, such as pluricentric languages, is to employ speech resources of different varieties. This course of action is even more relevant for "non-dominant" varieties. For instance, Dutch is a pluricentric language with two different standard language varieties, Netherlandic Dutch, spoken in the Netherlands,

and Flemish Dutch, spoken in Flanders, Belgium. These two varieties have lexical differences, but the most significant differences can be found in their phonetics (Verhoeven, 2005; Van de Velde et al., 2010), a distinction comparable to that in many other pluricentric languages. The pluricentric character of Dutch plays a major role in the language policy developed by the Netherlands and Flanders[10]. It has been also an important pillar in the national and bi-national initiatives aimed at strengthening the digital infrastructure for the Dutch language such as the Spoken Dutch Corpus (Oostdijk, 2000), the BLARK (Strik et al., 2002), and STEVIN (Spyns & Odijk, 2013). Thanks to these balanced programmes both varieties of Dutch have developed important language and speech technology resources that can be used for both research and development. However, for conducting ASR-based research in dysarthric speech, the amounts of speech data with corresponding annotations are still insufficient. A complicating factor is that ASR performance on dysarthric speech is notoriously lower than on non-dysarthric speech. To address this problem, Yilmaz et al. (2016b) studied whether combining resources of Netherlandic Dutch and Flemish non-dysarthric speech would help improve ASR performance on Flemish dysarthric speech. The results showed that combining resources led to improved ASR recognition on the Flemish dysarthric speech when comparing the ASR outputs to the prompts. These promising results open up opportunities for further research in developing dedicated ASR technology for the diagnosis and therapy of patients affected by speech disorders that reduce their speech intelligibility.

Currently, several languages can be defined as pluricentric, and resources for speech technology are available to a limited extent for these pluricentric languages. However, it seems that studies exploring whether existing resources from different varieties of a pluricentric language can be employed to the benefit of recognition of speech from other varieties are few and far between. Some studies did employ resources of different varieties of a pluricentric language, for instance, to study accented speech recognition. Winata et al. (2020) performed a cross-accented English speech recognition task as a benchmark for measuring the ability of the model to adapt to unseen accents using the CommonVoice corpus. This

[10] https://taalunie.org

corpus contains English read speech from 16 different areas such as Africa, Australia, Canada, England, Hong Kong, India, the United States, etc. In this study, the corpus was split into training and test sets for two settings: (1) mixed-region, and (2) cross-region. The focus was on evaluating a new approach, i.e., model-agnostic meta-learning (MAML) to develop a robust speech recognition system. Arsikere et al., (2019) studied a simple phone mapping approach to English multi-dialect acoustic model. In addition to the evaluation of their new approach, they discussed the gains in using resources from multiple varieties of English. The ASR model trained on four resources, i.e., American, Australian, Indian, and British English, performed better not only for accents with smaller amounts of training data but also for those which contribute more than half of the total training data.

Some other studies also used corpora from different varieties of a pluricentric language in training and test sets for dysarthric speech recognition. For instance, Librispeech, a corpus of American English non-dysarthric speech, was used as the training set, while TORGO, a corpus of Canadian English dysarthric speech, was used as the test set (Lin et al., 2020; Yue et al., 2020a). However, these studies focused more on employing a new approach, developing models with different architectures, and on the difference between non-dysarthric and dysarthric speech rather than on the added value of using speech resources from different varieties of the same pluricentric language.

It is important to note that the majority of studies on ASR for dysarthric speech recognition did not employ pluricentric language resources. Some of these studies are cross-lingual, in the sense that the ASR models are trained on dysarthric speech of one language and then tested on dysarthric speech of another language (Takashima et al., 2019a). This is different from studies employing pluricentric language resources, where speech data for training and testing ASR models are part of the same language although they originate from different varieties. Some studies are cross-speaker-type in which ASR models are trained on non-dysarthric speech and tested on dysarthric speech (Haderlein et al., 2011; Le et al., 2016; Middag et al., 2008; Middag et al., 2009c; Wang et al., 2021). Some others combine these two types of approaches (Bhat & Strik, 2020; Takashima et al., 2019b). Many of these studies showed moderate correlations between ASR-based feature

vectors and subjective intelligibility measures (Le et al., 2016; Middag et al., 2008; Middag et al., 2009c). For example, Middag et al. (2008) reported on training ASR systems to generate different features and then used different combinations of the features through an intelligibility prediction model to generate intelligibility scores automatically. High correlations were reported for combining different types of pathological speech and for a specific pathology type, 0.86 and 0.90, respectively. Haderlein et al. (2011) reported a human-machine correlation of r =0.85 for testing on German dysarthric speech with ASR trained on German non-dysarthric speech. However, it is difficult for speech-language pathologists in clinical practice to interpret such correlations, which are calculated between ASR-based feature vectors and subjective intelligibility measures. Speech-language pathologists lack the background knowledge to interpret these sets of features, such as those based on the Mel-Frequency Cepstral Coefficient (MFCC). In addition, different studies may use different sets of features, making it difficult for them to compare the results across studies. Moreover, these ASR models were trained and tested on the speech of only one language variety (Haderlein et al., 2011; Middag et al., 2008) or used specifically designed data, e.g., word lists used in the Dutch Intelligibility Assessment (DIA; Middag et al., 2009b). Such specifically designed data require a substantial amount of human effort in preparation, may not be available for other languages or varieties of a different pluricentric language, and are less ecologically valid. As described above, most of the studies focused on training specific acoustic models, adapting non-dysarthric speech models to dysarthric speech, and identifying alternative feature sets that are less dependent on mismatched acoustic models. In any case, they do not consider the resources from varieties of a pluricentric language, which might have potential benefit to the recognition of dysarthric speech or the relation with speech intelligibility of dysarthric speech.

In this chapter, we investigated whether speech corpora pertaining to the dominant variety of a pluricentric language can be used to develop objective, automatic measures of speech intelligibility of dysarthric speech in the non-dominant variety. We use dominant in the sense of having more speakers and larger resources. To broaden the scope of this chapter, we abstracted away from the specific scenario in which speech resources

are, to a certain extent, available for both varieties, the dominant and the non-dominant one. We hypothesized a scenario in which specific speech resources are not available for the non-dominant variety and use speech resources from the dominant variety instead. Subsequently, we compared the results obtained in the hypothetical scenario with those obtained in the realistic scenario in which specific speech resources from the non-dominant variety are actually available. As mentioned above, it is important to note that in the case of pathological speech in general and dysarthric speech in particular, we are always dealing with relatively small amounts of speech data.

In line with research on the intelligibility of dysarthric speech, different measures can be obtained at different levels of granularity through different measurement methods. An important distinction is that between rating-based measures, such as a global measure, and transcription-based measures through orthographic transcriptions, a sort of verbatim representation of the speech produced by a patient. These different measures have both advantages and disadvantages that we are not going to discuss here. Suffice it to say that both can be very useful for diagnosis and therapy. This means that for both types of measures, it is interesting to investigate to what extent speech intelligibility can be measured automatically by employing ASR technology. In turn, it is meaningful to investigate to what extent speech data of different varieties of the same pluricentric language can be usefully employed to develop such objective, automatic measures of speech intelligibility.

The Dutch language is used here as an example of a pluricentric language since several dysarthric speech resources are available for its dominant variety, Netherlandic, and the non-dominant variety, Flemish Dutch, albeit of limited size. Accordingly, we address the following two research questions:

- RQ1: To what extent can Netherlandic Dutch dysarthric speech data contribute to improving automatic transcriptions of Flemish dysarthric and non-dysarthric speech?
- RQ2: To what extent can Netherlandic Dutch dysarthric speech data contribute to developing objective, automatic global measures of speech intelligibility for Flemish dysarthric and non-dysarthric speech?

It is important to note that this chapter evaluates ASR models by using two types of intelligibility metrics: orthographic transcriptions and global intelligibility assessments, both obtained from experts. To address our first research question, the target references used in ASR systems were orthographic transcriptions obtained from experts including the pronunciation errors they transcribed. Differently, to address our second research question, canonical words were used as target references. The results were then compared with the global intelligibility assessment data, which were also collected from experts. Moreover, the ASR models were also compared between four severity levels of dysarthria to explore whether the additional Netherlandic Dutch dysarthric speech data play a role in the different severity levels.

7.2 Method

7.2.1 Experimental design

For each research question, we investigated two different scenarios: a) a hypothetical scenario in which Flemish dysarthric speech data are not available for training the ASR models so that one has to resort to dysarthric speech data of the Netherlandic Dutch; b) a realistic scenario in which some Flemish dysarthric speech data are available and can be used for training. In each scenario, we trained two ASR models on different data, with and without Netherlandic dysarthric speech, and compared their performance. For all four ASR models, we used the same language model that was trained on combined text materials of Flemish and Netherlandic Dutch in CGN corpus. Detailed information of the two scenarios about the corresponding training and test sets for each ASR model is presented in Table 7.1.

7.2.2 Speech data

We describe the speech corpora from which we selected the speech data for the training sets (FN, ND, and FD-train) and test set (FD-test). The speech segments with a non-speech sound produced by the speaker or with incomprehensible words were excluded from both the training and test sets.

Table 7.1. Detailed information of the two scenarios about the corresponding training and test sets for each ASR model. FN = Flemish non-dysarthric speech, ND = Netherlandic dysarthric speech, FD = Flemish dysarthric speech. The FD speech data were split into a training and a test part.

Scenario	Training set	Test set
hypothetical	FN	FD-test
	FN + ND	
realistic	FN + FD-train	
	FN + FD-train + ND	

For Flemish non-dysarthric speech (FN), we used the read speech components of Flemish speech in the CGN corpus (Oostdijk, 2000). It is a Netherlandic Dutch-Flemish speech corpus that contains representative collections of contemporary standard Dutch as spoken by non-dysarthric adults in the Netherlands and Flanders. This corpus contains 14 components, such as conversations (face-to-face), interviews, telephone conversations, lectures, and read speech. We only selected the read speech components in order to have speech material comparable to the dysarthric speech we have in the test set. The read speech is based on an open set of texts, resulting in recordings varying in length. Also, this read speech has less background noise compared to the other components and is thus rather clean for training ASR models. It also covers a large number of speakers. The duration of the selected training data from CGN is 6 hours and 42 minutes.

For Netherlandic Dutch dysarthric speech (ND), we combined three datasets, i.e., the EST dataset and two CHASING datasets. The EST dataset (Yılmaz et al., 2016a) contains Dutch dysarthric speech that was collected as a part of the e-health-based speech therapy (EST) research program (Beijer, 2012a). The collected data were annotated according to a common protocol to create a principled dysarthric speech corpus. This contains dysarthric speech from 16 patients, where ten of them had Parkinson's Disease (PD), four patients had had a Cerebral Vascular Accident (CVA), one patient had suffered Traumatic Brain Injury (TBI), and one patient was affected by dysarthria due to a birth defect. These patients were aged from 34 to 75 years with a median of 54.5 years. The level of dysarthria varied from mild to moderate, with seven at mild, eight at moderate, and

one at moderate-severe level. The speech tasks presented to the patients consisted of word and sentence lists with varying linguistic complexity. The database includes 12 Semantically Unpredictable Sentences (SUSs) with 6- and 13-word declarative sentences, 12 interrogative sentences with 6 words each, 13 Plomp and Mimpen sentences (Plomp and Mimpen, 1979), which are short, simple, and represent conversational speech, 5 short texts, 30 sentences with /t/, /p/, and /k/ in initial position and unstressed syllable, 15 sentences with /a/, /e/, and /o/ in unstressed syllables, production of 3 individual vowels /a/, /e/, and /o/, 15 bisyllabic words with /t/, /p/, and /k/ in initial position and unstressed syllable, and 25 words with alternating vowel-consonant combinations (CVC, CVCVCC, etc.). The duration of the selected training data from EST is 5 hours and 56 minutes.

The two CHASING datasets were collected in research aimed at developing a serious game for speech disorder treatment[11]. The first one, the CHASING01 dataset (Yılmaz et al., 2017), contains speech of five male patients who participated in speech training experiments and were tested at six different times during the treatment. The second one contains speech of five male and three female patients who were tested at three different times during the treatment. These patients were aged from 53 to 75 with a median of 63.5 years, ten of them having PD and three having had a CVA. For each test time, utterances of the following material were collected: 12 SUSs, 30 /p/, /t/, /k/ sentences[12] in which the first syllable of the last word is unstressed and starts with /p/, /t/ or /k/, 15 vowel sentences with the vowels /a/, /e/, and /o/ in stressed syllables, pronunciations of isolated vowels /a/, /o/, and /e/, 15 words with /p/, /t/, /k/ in word initial position and in unstressed syllable, and the "appeltaarttekst" ("apple cake recipe" in English), which consists of the recipe for making apple cake, in five parts, and spontaneous speech about specific topics (e.g., hobby and work). Since spontaneous speech was different from read speech, it was excluded from our training set, and the total duration of the selected training data from

[11] http://hstrik.ruhosting.nl/chasing/

[12] An example of /p/ sentence: 'de aard van deze man is optimistisch en positief' ('the nature of this man is optimistic and positive' in English). An example of /t/ sentence: 'op de salade ligt een vers ei en een knalrode tomaat' ('on the salad is a fresh egg and a bright red tomato' in English). An example of /k/ sentence: 'de jongen bekijkt de vissen en het schitterende koraal' ('the boy looks at the fish and the magnificent coral' in English).

these two datasets is 15 hours and 16 minutes. Note that the two CHASING datasets did not report severity level information. On the other hand, the speakers in the two CHASING datasets were of similar age to those in the EST dataset and had the same diseases as some of those in the EST dataset.

For the Flemish dysarthric speech (FD), we used speech from the COPAS corpus (Middag, 2012) of 49 speakers who read the text 'Papa en Marloes' (PM, 'Papa and Marloes' in English), which is a commonly used narrative for assessing the intelligibility of pathological speech. We selected four of nine sentences of the PM text as follows:

1. "Papa en Marloes staan op het station." (PM1, in English "Papa and Marloes are at the station."),
2. "Marloes kijkt naar links." (PM2, in English "Marloes looks to the left."),
3. "In de verte ziet ze de trein al aankomen." (PM3, in English "In the distance she can see the train coming."),
4. "Het is al vijf over drie dus het duurt nog vier minuten." (PM4, in English "It is already five past three so it will take another four minutes.").

These four sentences were selected according to our previous study in Chapter 3. The authors explained that these four sentences were chosen because they vary in length and contain the corner vowels, i.e., /a:/, /u/, and /i/. Further, for the training set (FD-train), we used speech of 23 speakers who were aged from 11 to 78 with a median of 44.5 years. These speakers varied in severity levels of dysarthria from mild to moderate, with 16 speakers at mild level, six at moderate level, and for one speaker the severity level was unknown. The total duration of the FD-train is 6.4 minutes of speech. For the test set (FD-test), we used speech of 36 speakers who were aged from 8 to 85 with a median of 46.5 years (see Table 3.1). FD-test includes 10 non-dysarthric speakers, 12 mild, 9 moderate, and 5 severe dysarthric speakers. The total duration of the FD-test is 9 minutes of speech. Note that the FD-test was built based on our previous study (Chapter 3). The study was conducted to collect various subjective intelligibility measures from experts, and these intelligibility measures were used to evaluate our ASR models, as explained below in Section 7.2.3. Moreover, to ensure that all four ASR models were speaker-independent, the dysarthric speakers in FD-train and FD-test were different. This means that excluding the 26

dysarthric speakers in the FD-test set based on Chapter 3, the other 23 of the total 49 dysarthric speakers were used in the FD-train set.

7.2.3 Evaluation

To evaluate the performance of ASR models, we calculated the word error rate (WER) between the ASR outputs and target references. Specifically, for the first research question, the human orthographic transcriptions at word level were used as the target references, and the distribution of the WERs over different utterances for different severity levels was analyzed. For the second research question, the prompts of the utterances were used as the target references, and the correlations between the WERs and the subjective global intelligibility measures were analyzed.

The human orthographic transcriptions and the subjective global intelligibility measures mentioned above for the FD-test were collected from five speech-language pathologists as expert listeners as explained in Chapter 3. The global intelligibility measures were collected through a Visual Analogue Scale (VAS) ranging from 0 (not intelligible) to 100 (intelligible) with tick marks for every ten scores (e.g., 10, 20, 30, etc.). The interrater reliabilities of the word accuracy (AcW) computed from the transcriptions and the global measure are 0.83 and 0.93, respectively. The construct validity was measured by the correlation between these two subjective intelligibility measures which is 0.75. The high interrater reliabilities and the validity show that it is legitimate to use these subjective intelligibility measures as the reference to evaluate the ASR models' performance and to address the research questions.

7.2.4 Implementation details of training

The ASR experiments were performed using the Kaldi ASR toolkit (Povey et al., 2011). A standard feature extraction scheme was used by applying Hamming windowing with a frame length of 25 ms and a frameshift of 10 ms. A conventional context-dependent GMM-HMM system with 20k Gaussians and 2500 triphone states was trained on the 39-dimensional MFCC features including the deltas and delta-deltas. We also trained a GMM-HMM system on the LDA-MLLT-SAT features, followed by training models with speaker adaptive training using FMLLR features. This system

was used to obtain the state alignments required for training the DNN as shown below.

The DNNs with 6 hidden layers and 1024 sigmoid hidden units at each hidden layer were trained on the 40-dimensional log-Mel filterbank features with the deltas and delta-deltas. The DNN training was done by mini-batch Stochastic Gradient Descent with an initial learning rate of 0.0015 and a minibatch size of 256. The default initial learning rate of 0.0015 was used in the first training stage. A trigram language model was trained on the text as we described in Section 7.2.1.

7.2.5 Statistical analyses

To compare the performance of different ASR models, we applied statistical analyses by using the *stats* (R Core Team, 2020), *base* (R Core Team, 2020), *car* (Fox & Weisberg, 2019), and *ggplot2* (Wickham, 2016) packages in RStudio (RStudio Team, 2020) with R version 4.0.2 (R Core Team, 2020). Specifically, to address the first research question, we calculated the mean and standard deviation of WER scores computed between ASR outputs and human orthographic transcriptions in the four ASR models per utterance and speakers' severity level. We applied *t* test to study the difference in WER scores between different models and their effect size based on Cohen's *d*. We also made a boxplot for WER by severity level. To address the second research question, we first calculated the correlation coefficients between ASR outputs (WER) and the subjective global intelligibility measures at utterance and speaker levels, as well as the significance of correlations. The ASR outputs (WER) were logit transformed as this is a proportional measure. The logit transform was calculated as the log of the proportion divided by one minus the proportion, where proportion refers to WER divided by 100. Note that in WER, 0 has been set as 1 and 100 as 99 before the transform as it was not possible to logit transform 0 and 100 values. To further study the relationship between perceptual and computed intelligibility scores, we applied regression analysis (Bocklet et al., 2012; Middag et al., 2008; Van Nuffelen et al., 2009a). We visualized the results at speaker level in scatterplots. We studied the residuals based on the linear regression for each severity level in the four ASR models and calculated Levene's test for the four models to test whether the variances of the residuals in the four severity groups of speakers were different.

7.3 Results

7.3.1 RQ1: To what extent can Netherlandic Dutch dysarthric speech data contribute to improving automatic transcriptions of Flemish dysarthric speech?

The mean and standard deviation (SD) of WER and of the numbers of errors, i.e., substitutions, deletions, and insertions, using the four ASR models are presented in Table 7.2. The values of the *t* tests (*df* = 719) between all model pairs were as follows: the WER of FN was greater than those of the other three models with *t* values of 9.14, 24.18, and 23.45, respectively; the WER of FN + ND was greater than those of the other two models with *t* values of 19.78 and 19.55, respectively; the WER of FN + FD-train was greater than that of FN + FD-train + ND with a *t* value of -2.71. All WER differences between all model pairs were significant (*t* test, $p < .05$).

Table 7.2. Mean (standard deviation) of word error rate (WER) and of the numbers of errors, i.e., substitutions, deletions, and insertions, computed between ASR outputs and human orthographic transcriptions (the lower the better). All WER differences between all model pairs were significant (*t* test, $p < .05$).

ASR model	WER (%)	Substitution	Deletion	Insertion
FN	54.80 (42.61)	3.27 (3.06)	0.75 (1.53)	0.64 (1.62)
FN + ND	46.12 (42.11)	2.83 (3.16)	0.64 (1.44)	0.69 (1.80)
FN + FD-train	22.71 (34.07)	0.96 (1.69)	0.46 (1.27)	0.29 (1.03)
FN + FD-train + ND	24.15 (35.41)	1.32 (2.16)	0.22 (0.71)	0.49 (1.46)

As can be seen in Table 7.2, the mean values of WER decrease substantially from FN to FN+ND and to FN+FD-train, especially largely from FN+ND to FN+FD-train. FN+FD-train+ND performs slightly worse than FN+FD-train. Although the difference was significant, it was small according to its effect size (Cohen's $d = 0.04$). The WERs for the different models seem to be reflected best by the substitutions. We need to ascertain that the WER reduction rates mirror a better performance of the ASR models in recognizing the words that speakers actually said.

Since the interrater reliability of the word accuracy using human orthographic transcriptions, as mentioned in Section 7.2.3, was very high, we did not plot the WERs across the listeners. Rather, we explored the

model differences in WERs over different severity levels for the four ASR models. The results are given in Figure 7.1.

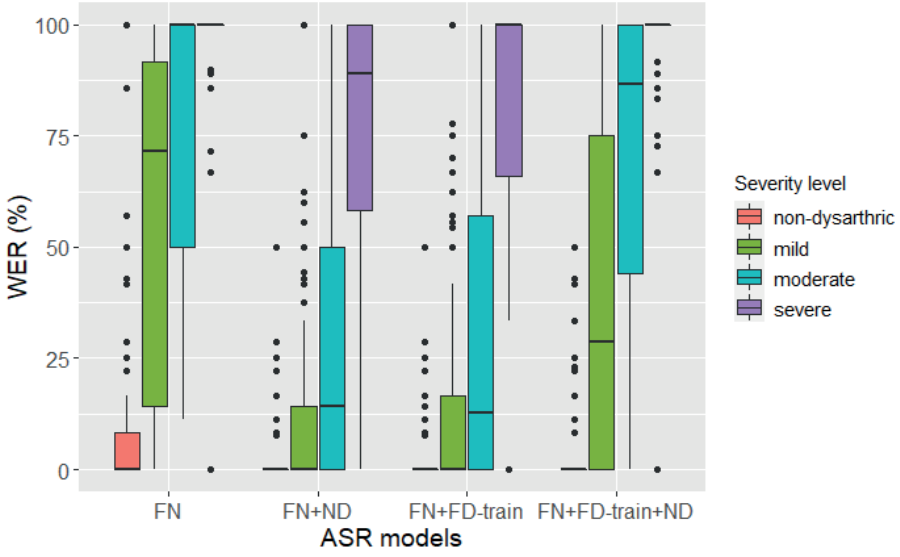


Figure 7.1. Boxplots for WER computed between ASR outputs and human orthographic transcriptions in the four ASR models per speakers' severity level. All WER differences between all model pairs were significant (*t* test, $p < .05$). **Best viewed in color.**

As can be seen in Figure 7.1, different models perform differently for severity levels. The recurrent pattern is that the medians in the boxplots rise along increasing severity levels. At the same time, we see smaller boxes for FN+FD-train and FN+FD-train+ND, especially for mild and moderate levels, and lower WER scores for all severity levels. These results showed that FN+FD-train and FN+FD-train+ND had smaller intersections between the severity levels than those for FN and FN+ND.

7.3.2 RQ2: To what extent can Netherlandic Dutch dysarthric speech data contribute to developing objective, automatic global measures of speech intelligibility for Flemish dysarthric speech?

As shown in Table 7.3, the magnitudes of the correlation coefficients gradually increase from FN to FN+FD-train+ND at both utterance and speaker levels although FN+FD-train+ND showed a slightly lower correlation coefficient

than FN+FD-train at speaker level. The correlations at speaker level for the last two models in Table 7.3 are very high. All correlation coefficients were significant ($p < .05$).

Table 7.3. Magnitudes of correlation coefficients between ASR outputs (with logit transform) and the subjective global intelligibility measures at utterance and speaker levels. All correlation coefficients were significant ($p < .05$).

ASR model	Utterance level	Speaker level
FN	0.50	0.65
FN + ND	0.56	0.73
FN + FD-train	0.61	0.87
FN + FD-train + ND	0.66	0.85

The scatterplots in Figure 7.2 show that the observed data points in FN+FD-train and FN+FD-train+ND are generally closer to the regression lines than those in FN and FN+ND, especially for non-dysarthric and mild dysarthric speakers. This reflects the higher correlations of FN+FD-train and FN+FD-train+ND. For moderate and severe dysarthric speakers, we observed slightly larger distances between the predicted and observed scores. Figure 7.2 also shows clearly that there is one larger overestimation and one larger underestimation of WERs on the basis of the VAS score in the moderate group.

We further investigated the residuals of the regression analyses for each severity level in the four ASR models. The results are shown in Table 7.4. Larger absolute values of the means mean that the speakers are further away from the regression lines in Figure 7.2. We also applied Levene's test on the four models to test whether the variances of the residuals in the four groups of speakers with varying severity levels were different. We observed no significant differences for model FN. Significant differences were found for the other three models ($p = .018$ for FN+ND; $p = .040$ for the last two models). These results indicate that the models are less successful in predicting WER by using VAS for dysarthric speakers, especially in the moderate and severe groups. This effect is also visible in Figure 7.2, as we mentioned above. Nevertheless, FN+ND better matches the two successfully trained models, i.e., FN+FD-train and FN+FD-train+ND, than FN.

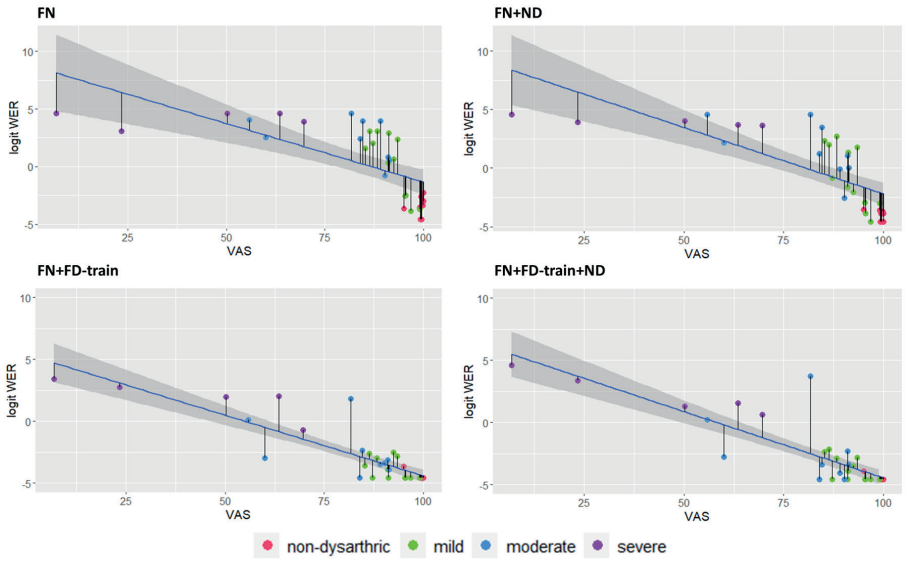


Figure 7.2. Scatterplots with regression lines for the subjective global intelligibility measures, through a Visual Analogue Scale (VAS), and WER (with logit transform) at speaker level for the four ASR models, with different colors for the severity levels. **Best viewed in color.**

Table 7.4. Mean and Standard Deviation (SD) of residuals of the regression analyses for each severity level in the four ASR models.

ASR model	Severity level			
	non-dysarthric	mild	moderate	severe
FN	-2.44 (0.91)	0.79 (2.34)	1.84 (1.79)	-0.33 (2.91)
FN+ND	-1.97 (0.41)	0.55 (2.26)	1.64 (1.90)	-0.31 (2.73)
FN+FD-train	-0.17 (0.15)	-0.22 (0.79)	0.10 (1.92)	0.69 (1.62)
FN+FD-train+ND	-0.16 (0.08)	-0.18 (0.75)	0.07 (2.56)	0.63 (1.31)

7.4 Discussion

In this chapter, we have investigated whether speech resources pertaining to the dominant variety of a pluricentric language are useful to obtain objective, automatic, speech technology-based intelligibility measures for dysarthric speech of the non-dominant variety. Our study focused on two types of intelligibility metrics that are often employed in dysarthric speech

research, orthographic transcription and a global measure, as addressed by our two research questions. For each research question, a hypothetical and a realistic scenario are discussed separately.

7.4.1 RQ1: To what extent can Netherlandic Dutch dysarthric speech data contribute to improving automatic transcriptions of Flemish dysarthric speech?

The results presented in Section 7.3.1 indicate that in the hypothetical scenario, adding the ND speech data leads to a lower word error rate. This means that the automatic transcriptions are more in line with those made by expert transcribers. Improvements are observed for all utterances and all severity levels as well as for non-dysarthric speech. These results are plausible. Dysarthric speech deviates considerably from non-dysarthric speech. When dysarthric speech data from the non-dominant variety are not available for training, adding dysarthric speech of the dominant variety helps because this added speech is similar to the test speech pertaining to the non-dominant variety. Similar results can also be found for accented speech in Winata et al. (2020) where the performance of ASR models for unseen accented speech, e.g., English speech in Wales, was better when other accented speech was added for pre-training.

In the realistic scenario, we see that the FN+FD-train and FN+FD-train+ND models perform comparably in all cases. This seems to indicate that once you have some dysarthric speech of the non-dominant variety for training, adding dysarthric speech of the dominant variety does not make a significant contribution. However, the marginal differences in results between these two models might have to do with a possible mismatch between the two types of dysarthric speech over and above the fact that they are from different varieties. In this specific case, there is a speaker mismatch between ND and FD-test. The ND data were recorded from speakers aged between 34 to 75, while the FD-test set also contains young children. Besides, the ND data contains speakers who are mild to moderate dysarthric, while the FD-test set also contains non-dysarthric and severe dysarthric speakers. Another explanation could be that solely adding FD-train, which is similar to FD-test in terms of language variety and text of the recordings, already led to a large boost. Therefore, further adding

ND, which is very different from FD-test, will not contribute much useful information. This is confirmed by the fact that the FN+FD-train model performs better than the FN+ND model.

Overall, the results show that adding Netherlandic Dutch dysarthric speech can help obtain better automatic transcriptions of Flemish dysarthric speech if no Flemish dysarthric speech data are available for training. Note that Yilmaz et al. (2016b) also studied the ASR models for Flemish dysarthric speech but added the Netherlandic Dutch non-dysarthric speech rather than dysarthric speech for training. Also, they evaluated the ASR models by comparing the automatic transcriptions with prompts, whereas we evaluated the ASR models by comparing the automatic transcriptions with orthographic transcriptions obtained from human experts.

7.4.2 RQ2: To what extent can Netherlandic Dutch dysarthric speech data contribute to developing objective, automatic global measures of speech intelligibility for Flemish dysarthric speech?

In the hypothetical scenario, as described in Section 7.3.2, the correlations between WER and the subjective global intelligibility measure (VAS) for the FN+ND model are clearly higher than those for the FN model at both utterance and speaker levels. Similarly, the scatterplots, as well as the mean and standard deviations of residuals of regression analyses, show that the FN+ND model performs better at speaker level. This seems to be in line with the finding for RQ1 in the same scenario. The FN+ND model yields outputs that are more similar to the transcriptions by expert transcribers for all severity levels as compared to the FN model.

In the realistic scenario, the correlations for the FN+FD-train+ND model are higher than those for the FN+FD-train model at utterance level. The correlations at speaker level are comparable in the two models when combining all speakers. Similarly, the scatterplots, as well as the mean and standard deviations of residuals of regression analyses, show that the two models perform similarly at speaker level for speaker at all severity levels and for non-dysarthric speakers. The marginal difference between the two models might be due to the mismatch between the ND and the FD-test sets, as described above, concerning speaker, text of recordings, and

language variety. In addition, these two models both outperform the FN and the FN+ND models, especially for non-dysarthric and mild dysarthric speakers, since the added FD-train set is very similar to the FD-test set.

Many studies reported very high correlations between ASR outputs and subjective measures of intelligibility of dysarthric speech, but only when the models were specifically designed or the training and test sets have similar distributions (Kim et al., 2015; Middag et al., 2008; Middag et al., 2009c). For example, Middag et al. (2008) studied ASR systems to align the speech to generate phonemic and/or phonological features and used the features to predict the Phoneme Intelligibility for speakers through a prediction model. Both the ASR systems and the prediction model were trained and tested on data from COPAS. This study targeted only the word lists that were specifically designed for assessing intelligibility of phonemes, i.e., DIA task. Their other study (Middag et al., 2009c) used a similar system as described above with ASR systems trained on CGN read speech and CoCGN corpus (Demuynck et al. 1997) together with the prompts. Although their training set also included speech data from different varieties of Dutch, the training data were very similar to those in the test. The individual contribution of each of the two varieties was not studied. Our study did not aim to explore a new feature or a model structure. Rather, we chose a very common design and implementation of the ASR models and explored the influence of using speech resources in the training set that are of a different variety from those in the test set.

In general, adding Netherlandic Dutch dysarthric speech can contribute to developing objective, automatic global measures of speech intelligibility for non-dysarthric and dysarthric Flemish speech when no dysarthric speech Flemish data are available for training. Moreover, we studied ASR models by using orthographic transcriptions including errors as target references. This is different from typical robust ASR systems where canonical words are used as target references. By applying such analyses, we were able to explore the possibility of replacing human transcribers with ASR models, resulting into a fully automatic assessment of intelligibility for dysarthric speech. Again, as we discussed above, further exploration on the condition that the Netherlandic Dutch dysarthric speech in the training set and the Flemish dysarthric speech in the test set are matched in terms of

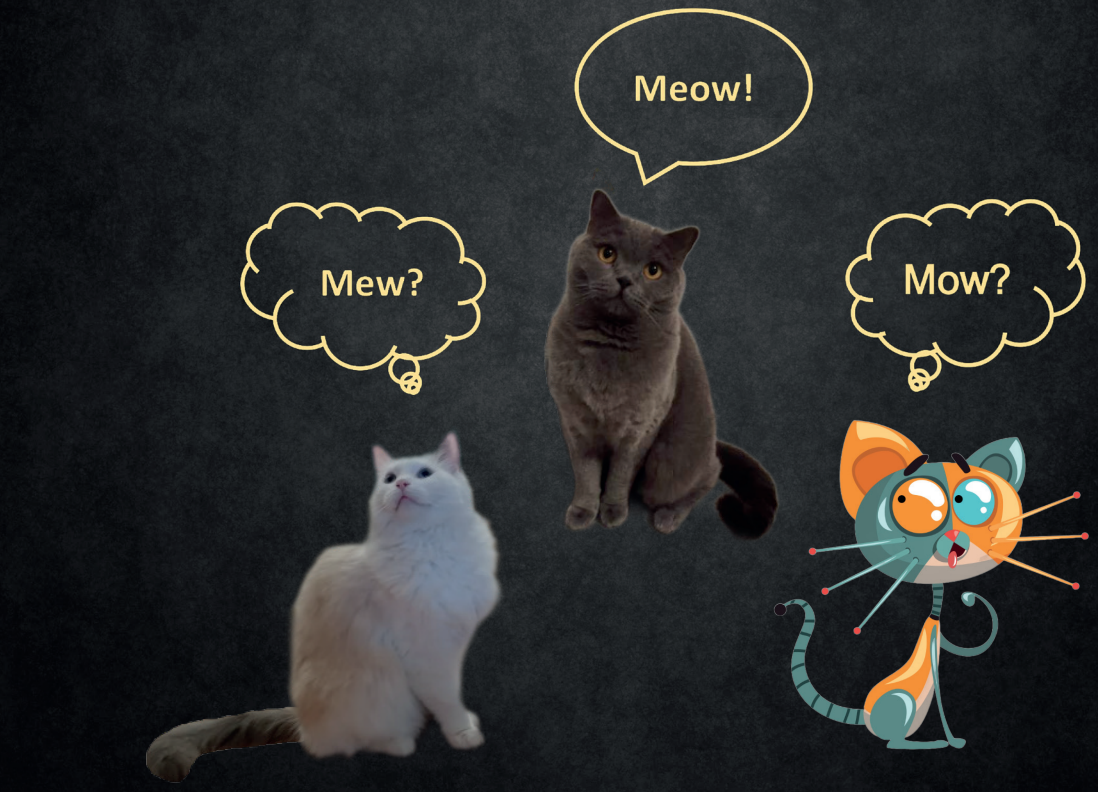
speaker and text is needed. Note that Dutch is in a very specific situation where speech resources are available for both varieties, the dominant and the non-dominant one. In reality, this specific situation may not hold true for other pluricentric languages, which is exactly consistent with the hypothetical scenario we discussed. Therefore, our findings in this study provide important insight for these pluricentric languages. It is feasible to use the speech resources from the dominant variety to improve dysarthric speech recognition for the non-dominant variety, especially to improve the correlation with global intelligibility measures of dysarthric speech.

7.5 Conclusions

In this chapter, we investigated the potential contribution of dysarthric speech resources from the dominant variety of a pluricentric language as additional training data for improving automatic speech recognition of dysarthric speech of the non-dominant variety. The performance of automatic speech recognition was evaluated by two types of intelligibility metrics: orthographic transcriptions and global intelligibility assessments, both obtained from experts. Overall, the results show that adding dysarthric speech of the dominant variety, i.e., Netherlandic Dutch dysarthric speech, can help obtain better automatic transcriptions and global intelligibility measures for dysarthric speech of the non-dominant variety, i.e., Flemish Dutch dysarthric speech. However, this positive effect is overpowered due to other factors when dysarthric speech resources from the non-dominant variety are available. In the future, it is worth exploring the condition in which the Netherlandic Dutch dysarthric speech in the training set and the Flemish dysarthric speech in the test set are matched in terms of speaker and text. Given the high correlation between the ASR outcomes and the subjective intelligibility measures, further research is also needed to develop ASR models as a tool to estimate and diagnose the intelligibility of dysarthric speech.

Interestingly, exploiting pluricentric language resources is not only applicable to assessing the intelligibility of dysarthric speech but also to other domains of speech research, particularly when the resources are scarce or asymmetric with much larger speech resources available in the dominant

variety. For example, it may help improve the robustness of ASR for non-native speech recognition in the non-dominant variety. Another domain is aphasic speech. However, the success of a pluricentric approach depends on the linguistic differences between the varieties of the pluricentric languages involved. As we earlier pointed out, phonetic differences are a main component of distinguishing Netherlandic and Flemish Dutch.


Meow!
Mew?
Mow?

DISCUSSION AND CONCLUSIONS

The aim of the research reported in this dissertation was to advance our understanding of the intelligibility assessment of pathological speech by investigating both subjective and objective procedures and to eventually provide guidelines for developing valid measurement procedures. Specifically, for the investigation of subjective procedures, we aimed to comprehensively study the effects of different factors (i.e., speech materials, measurement methods, granularity levels, and listener characteristics) on intelligibility measures. For the investigation of objective procedures, we focused on acoustic correlates of intelligibility and on addressing the low-resource problem in a pluricentric language when ASR models are used. This chapter discusses the findings related to the three research questions:

1. For subjective procedures, how do the factors in subjective procedures influence speech intelligibility? (**Chapters 2, 3, and 4**)
2. For objective procedures based on acoustic features, how do different features correlate to speech intelligibility? (**Chapters 5 and 6**)
3. For objective procedures based on ASR models, how can the low-resource problem in assessing speech intelligibility of a pluricentric language be addressed? (**Chapter 7**)

Section 8.1 discusses the effects of the factors in subjective procedures, as well as the reliability and validity of intelligibility measures. Section 8.2 discusses the findings related to objective procedures. In particular, Section 8.2.1 discusses the practical implications of using acoustic features to assess speech intelligibility and to classify different types of speech, and Section 8.2.2 discusses the possibility of incorporating speech resources from different varieties of the pluricentric language, Dutch in this case, into ASR systems to assess speech intelligibility and to generate human-comparable transcriptions. Section 8.3 provides guidelines for measuring intelligibility, and Section 8.4 provides recommendations for future work. Lastly, Section 8.5 summarizes the main conclusions.

8.1 Subjective procedures

In this subsection, we discuss the effects of different factors in subjective procedures as well as the reliability and validity of intelligibility measures

based on our investigations in the three listening experiments. These three listening experiments covered different speech materials, measurement methods, and granularity levels of intelligibility measures. Specifically, three types of speech materials varying in length, morphosyntactic complexity, and semantic predictability were employed. The intelligibility measures were collected through two categories of measurement methods, specifically through VAS and orthographic transcriptions, respectively. For orthographic transcriptions, our novel form of transcription that allows pseudowords was compared with the typical form of transcription. Various intelligibility measures were extracted at different granularity levels, i.e., utterance, word, and subword (grapheme and phoneme). Five expert listeners were recruited to give assessments of speech intelligibility for speakers with varying severity levels of dysarthria, dysarthria type, gender, age, etc. The assessments by the five expert listeners were compared with those in **Chapter 5**, which had been assigned by naïve listeners, to study the effects of listener experience. **Chapter 2** presented a comprehensive analysis of eight measures in the three listening experiments. **Chapter 3** further studied two measures at the utterance and word levels, and focused on addressing the reliability issues. Moreover, the usability of our novel pseudoword-allowing form of transcription was examined in depth. **Chapter 4** expanded the study of two types of phoneme-level measures and explored the possibility of using them to classify speakers.

8.1.1 The effects of speech materials

By comparing the results of the three experiments in **Chapters 2 through 4**, we investigated the effects of *speech materials* on intelligibility measures in general and, more specifically, with respect to the degree of semantic predictability and the sample length. First, speech materials with a higher degree of semantic predictability led to higher intelligibility regardless of speakers' severity levels of dysarthria. Past research has shown that intelligibility scores generally increase when the degree of semantic predictability in speech materials increases (e.g., Hustad, 2007; Miller 2013; Yorkston & Beukelman, 1978). In line with this finding, **Chapter 2** clearly showed that the means of scalar judgments and transcription-based accuracy scores increased from word lists, to semantically unpredictable sentences,

and to meaningful sentences, with declines in measures of distance and change, which examined intelligibility from the error perspective. **Chapter 4** also reported significant differences in the means of the two types of phoneme-level measures between meaningful sentences and word lists. Though the differences in the standard deviations of the measures between the three materials were marginal in **Chapter 2**, we observed significantly different variances between meaningful sentences and word lists in **Chapter 4**. Further, such an increase in intelligibility scores was observed for all severity levels based on the analysis of concurrent validity in **Chapter 3**, where two intelligibility measures were studied in relation to severity levels of dysarthria, with the intelligibility scores being higher in meaningful sentences compared to word lists. Also, such higher scores in meaningful sentences were particularly evident for the transcription-based, word-level measure when using our novel pseudoword-allowing form of transcription. These findings indicate that the listeners can profit from the contextual cues in the speech material irrespective of the speakers' severity level, and that they tend to overestimate the intelligibility in meaningful sentences. It is noteworthy that speakers with more severe dysarthria also benefited from the contextual cues in the sentences, which differs from those of previous studies (Miller, 1951; Hustad, 2007). This may be due to the form of transcription: we used our novel pseudoword-allowing form of transcription, whereas previous studies used the typical form of transcription. This finding seems to suggest that our novel form of transcription is better capable of capturing the subtle differences caused by different speech materials.

Second, different speech materials were found to lead to intelligibility measures that refer to different constructs of intelligibility. This was demonstrated in the analysis of construct validity in **Chapter 3,** where we studied the correlations of measures between meaningful sentences and word lists. This may be because the level of communicating information in a list of words is different from that in a meaningful sentence, and thus, listeners may rely on different perceptual cues in assessments of intelligibility.

Third, the presence of the context for the target unit to be examined, which can be seen as a change in the degree of semantic predictability, was also found to lead to intelligibility measures that refer to different

constructs of intelligibility. In **Chapter 4**, we examined the relation between our phoneme-level measures and an external variable, i.e., Phoneme Intelligibility, that contained different contexts for the target unit to be examined. Our phoneme-level measures were obtained through full transcriptions of word lists, which required writing down all phonemes composing a word, while the Phoneme Intelligibility scores were obtained through partial transcriptions, which required writing down only the target phoneme in a word with the remaining phonemes (the context) being presented to listeners. The correlations between these two types of measures were not very strong (< .75). This indicates that presenting phonemes other than the target phoneme in a word does affect the intelligibility measures. It may be because procedures examining only the target phonemes draw listeners' attention to specific, context-rich phonemes, whereas in procedures where no phoneme is presented to listeners, co-articulation or other aspects may affect the listeners' perception of individual phonemes. In other words, Phoneme Intelligibility appears to focus on the articulation of phonemes, whereas our phoneme-level measures appear to additionally take into account the influence of other speech properties, such as co-articulation and suprasegmental cues (Miller, 2013). A similar tendency was observed in a study by Yorkston and Beukelman (1978) although the authors examined the effect for a measure at the word level, i.e., the number of correctly identified words, in sentences.

Fourth, the sample length of speech materials, i.e., the number of words or syllables of individual speech samples, did not show an impact on speech intelligibility. We studied the effect of the sample length, i.e., the number of words or syllables of individual utterances, by applying Generalizability Theory and, in particular, by conducting a G study to analyze variance. The results of the G study for the three speech materials, as exemplified by the VAS, are presented in **Appendix A**. We observed relatively low percentages of the total variance related to *Utterance*, indicating that *Utterance*, which was categorized by the number of words or syllables, does not play an important role in explaining the variance in the intelligibility scores in each type of speech material.

Overall, the examination of reliability in **Chapters 2 through 4** showed that all three speech materials can provide reliable intelligibility measures

regardless of granularity levels and measurement methods. The type of speech material used to measure speech intelligibility in clinical practice and research should be determined based on both practical and theoretical considerations.

8.1.2 The effects of measurement methods and granularity levels

The effects of *measurement methods* and *granularity levels* were studied across three different speech materials in **Chapters 2 through 4**. The results suggest that the measures studied in this dissertation, varied in granularity levels and measurement methods, refer to different constructs of intelligibility. More specifically, we have the following observations.

First, different transcription-based, subword-level (grapheme- and phoneme-level) measures referred to slightly different constructs of intelligibility. We observed in **Chapter 2** that the correlations between the measures at the grapheme and phoneme levels were very strong regardless of speech materials. These results align with those of Ganzeboom et al. (2016). The authors also reported a very strong correlation between grapheme- and phoneme-level measures derived from orthographic transcriptions for various speech materials. These results suggest that these subword-level measures can be interchangeably used. This phenomenon was anticipated since these measures examined transcriptions in a similar fine-grained manner. However, the individual subword-level measures performed slightly differently when they were studied in relation to external variables.

Second, the subword-level measures showed different relations to different higher-level measures. We observed the correlations of subword-level measures were higher with the word-level measures than with the scalar judgments in **Chapters 2 and 4**. Similar results were also reported in Ganzeboom et al. (2016), where two subword-level scorings of the orthographic transcriptions showed stronger correlations with a word-level scoring of the orthographic transcriptions than with two scale ratings. This is understandable as the measures at the subword and word levels were extracted from the same transcriptions and were different from the scalar judgments.

Third, the abovementioned relations between various intelligibility measures were found to be associated with the degree of semantic predictability in speech materials. **Chapter 2** showed that the correlations between the phoneme-level measures and the scalar judgments were found to increase from word lists to meaningful sentences, i.e., as the degree of semantic predictability increased. A similar tendency was also observed between the scalar judgments and the transcription-based, word-level measures, yet their correlations were consistently below 0.85. These results suggest that when speech materials become more similar to spontaneous speech in daily speech communication, the constructs of intelligibility measured by scalar judgments and transcription-based measures appear to be more similar. One particular observation is that semantically unpredictable sentences led to lower correlations than word lists and meaningful sentences. This might be because the presence of suprasegmental cues in semantically unpredictable sentences compared to word lists makes the scalar judgments more different from transcription-based measures in terms of the constructs of intelligibility they refer to. In addition, although suprasegmental cues were also contained in meaningful sentences, their influence tends to be reduced by the richer contextual information. However, these findings of the speech materials' effects may be somewhat limited by the different distributions of speakers across the three speech materials, such as the different distributions of speakers' severity level of dysarthria, age, and gender. Therefore, future research should consider using the same distributions of speakers to validate whether the results are similar.

Fourth, the intelligibility measures collected through our specific experimental settings can be used interchangeably as speaker-level indicators. As shown in **Chapters 3 and 4**, the correlations between each pair of the measures at the utterance, word, and phoneme levels were very strong within the same speech material when the scores were averaged per speaker. The results of the comparison between the scalar judgments and the transcription-based, word-level measures across speech materials in **Chapter 3** showed that the scalar judgments seem to be a more stable and robust indicator of intelligibility, as the scalar judgments showed higher reliability values regardless of speech materials and higher correlations with

the external variables. These results are consistent with those of previous studies (Abur et al., 2019; Stipancic et al., 2016), which demonstrated that VAS is a suitable alternative to transcription-based intelligibility measures for sentences.

Fifth, scalar judgments seem to be a better proxy for dysarthria than transcription-based intelligibility measures. As intelligibility measures have frequently been used as a proxy for dysarthria (e.g., for a brief overview see Stipancic et al., 2021; Barreto & Ortiz, 2016; Kent, 1989; Hustad, 2007, 2008), we also examined the relation between our various measures and an external variable, i.e., severity levels of dysarthria, in **Chapters 2 through 4**. Results showed that the scalar judgments were more strongly correlated to the severity levels of dysarthria than the transcription-based measures, either at the word or subword level, and showed better discriminations of the severity levels of dysarthria. We also observed in **Chapter 3** that the correlations between the severity levels of dysarthria and Phoneme Intelligibility, which examines intelligibility at the phoneme level, were very low, suggesting that these two measures refer to different constructs of intelligibility. These findings support the view of Sussman and Tjaden (2012) that transcription-based intelligibility measures are not a perfect proxy for overall severity levels of speech although they are related. The authors reasoned that transcription-based intelligibility measures do not represent suprasegmental parameters, such as speaking rate, prosody, and naturalness, which can have a great impact on perceived severity levels of speech (Tjaden et al., 2014a). These findings indicate that the severity level of speech is a global perception of dysarthria and is similar to scalar judgments, whereas transcription-based measures focus more on articulation precision, which reflects one aspect of dysarthria.

Sixth, our novel pseudoword-allowing form of transcription was shown to be a valuable tool for obtaining reliable measures and for reducing the impact of contextual cues. We compared our novel form of transcription with the typical form of transcription in the two types of speech materials of sentences in **Chapters 2 and 3**. The results of **Chapter 2** showed that our novel form of transcription led to lower intelligibility, i.e., lower accuracy scores and higher error scores, indicating that listeners in the typical form of transcription tend to overestimate intelligibility, and that our novel

form of transcription seems to help reduce the impact of contextual cues in the meaningful sentences. The results of the reliability examination in **Chapters 2 and 3** indicate that our novel form of transcription can provide more reliable measures. This was more evident in the case where the listeners differed significantly in their transcriptions due to the design of the material. An example is one of the four meaningful sentences discussed in detail in Section 3.3.3 of **Chapter 3**. It is worth noting that when using our novel form of transcription, the instructions should be carefully presented, which is crucial to obtain reliable measures.

The above observations can be used to choose a measure of intelligibility in research or in clinical practice. **Chapter 3** demonstrated that scalar judgments are more reliable and stable in different speech materials compared to transcription-based, word-level measures. On the other hand, transcription-based measures, both at the word and subword levels, better represent the accuracy of listeners' perception of phonemes than scalar judgments. **Chapters 2 and 4** demonstrated that reliable and valid subword-level measures can be programmatically computed, with the advantage of greatly reducing human efforts and labor. Among the intelligibility measures that vary in granularity levels, the phonetic distance is the most informative, but also the most complex one to calculate, as it requires a grapheme-to-phoneme convertor and phonetic distance matrices. Although such resources are readily available for Dutch, they may be more difficult to obtain for other languages. In contrast, transcription-based, word-level measures are simple to calculate and show comparable performance to phoneme-level measures, especially when our novel pseudoword-allowing form of transcription is used. The comparable performance for the measures at the word and phoneme levels in our novel pseudoword-allowing form of transcription may be because this form of transcription can restrict the listeners' focus on segmental-level spelling. Further, applying our novel form of transcription could enrich these types of comparisons by providing additional information about perceptual accuracy at the segmental level, and expand the comparison to materials containing pseudowords compared to the typical form of transcription that allows only existing words. It is worth mentioning that our novel form of transcription may require additional manual corrections of the transformed phonemes.

8.1.3 The effects of listener characteristics

The listener characteristics in terms of *listener familiarity* with speakers and *listener experience* with dysarthric speech were considered when designing the experiments. The *listener familiarity* with speakers was controlled in our experiments to prevent its impact on our measures. Specifically, we did not conduct any speaker familiarization interventions to the listeners and allowed listeners to listen to each recording only for a limited number of times. Furthermore, although the speakers involved in the later experiment on word lists overlapped with those in the former experiment on meaningful sentences, no systematic increase in the intelligibility scores was observed. This may be because these two speech materials were too different in terms of semantic predictability. Another possible explanation is that the monosyllabic words and pseudowords in the word lists helped mask the identity of the speakers to a great extent.

The *listener experience* with dysarthric speech showed an impact on the reliability of intelligibility measures, especially for the transcription-based, word-level measure. In this dissertation, we indirectly compared the results between the expert listeners (**Chapters 2 through 4**) and naïve listeners (**Chapter 5**). Results convincingly showed that although only five expert listeners were recruited for our study in **Chapters 2 through 4**, they were sufficient to provide reliable measures of speech intelligibility regardless of granularity levels, speech materials, and measurement methods, as well as forms of transcriptions for the method of orthographic transcription. In contrast, the study in **Chapter 5**, where a larger number (eleven) of naïve listeners were recruited compared to that of expert listeners, showed acceptable reliability for the scalar judgments, but very low reliability for the transcription-based, word-level measure, i.e., word accuracy. This implies that future studies planning to extend our work to naïve listeners would need to recruit a large number of listeners to obtain reliable and valid results. However, it should be noted that having more listeners may inevitably increase human efforts in analyzing transcriptions, especially when multiple listening experiments are conducted to compare different procedures.

The differences in the reliability of intelligibility measures between the two types of listeners regarding their experience partly support the

idea that *listener experience* can affect intelligibility measures as suggested by Carvalho et al. (2021). The authors observed significant differences in intelligibility scores between healthcare professionals and naïve listeners in both word lists and sentences.

It is noteworthy that although the listeners in **Chapter 5** were also speech-language pathologists, we considered them to be naïve because they did not have extensive experience with disordered speech compared to expert listeners. This consideration is also supported by a previous study of Smith et al. (2019), which did not observe significant differences between 'trained' and 'untrained' listeners. The 'trained' listeners were similar to the listeners in **Chapter 5** since they were students majoring in speech and language therapy and received training in relevant aspects (e.g., perception, transcription, and phonological disorders).

8.1.4 Reliability and validity examinations

Generalizability Theory was found to be effective in systematically dealing with reliability issues in our experimental designs. The results in **Chapters 2 and 3** provided suggestions for clinical practice that five expert listeners were sufficient to provide reliable measures of intelligibility. Moreover, scalar judgments, in particular the VAS, were more reliable irrespective of speech materials when compared to the word-level measures obtained through either form of transcription. **Chapters 2 and 4** showed that phoneme-level measures were more reliable than the word-level measures, and that their reliability was comparable to scalar judgements collected through VAS irrespective of speech materials.

Generalizability Theory allows to study the optimal numbers of utterance samples and listeners to obtain reliable results, i.e., passing the criterion of 0.90 for professionally developed high-stake tests. **Chapter 3** showed that to obtain reliable results of scalar judgments, three samples per speaker in combination with four listeners were sufficient regardless of speech materials, while the results for the transcription-based, word-level measures were found to be different for different speech materials. Specifically, for word lists, having two expert listeners in combination with two samples was sufficient to obtain reliable measures, whereas for meaningful sentences, more listeners and much more samples were required. The analyses of one of

the four meaningful sentences in Section 3.3.3 of **Chapter 3** provided possible reasons for this difference. The cumulation of four pronunciation variants in this particular sentence made its transcriptions less reliable than the other three. This implies that these pronunciation variants should be avoided when selecting sentences for listening experiments, especially when the accuracy of transcribed words is used as a measure of speech intelligibility. Moreover, future research could extend our exploration of Generalizability Theory to semantically unpredictable sentences. Since naïve listeners may exhibit low reliable transcription-based measures on meaningful sentences, using semantically unpredictable sentences may help improve the reliability in that population of listeners.

Along with reliability examination, we also examined the validity of the intelligibility measures either by studying the correlations between the measures (construct validity) or by studying their correlations with external variables (concurrent validity) in **Chapters 2 through 5**. The validity examination helped us investigate the relation between the intelligibility measures that were derived through different subjective procedures, as well as the effects of the factors, as shown above in Sections 8.1.1 and 8.1.2.

Overall, the findings in **Chapters 2 through 5** aligned with the consensus that speech intelligibility is a complex construct. Therefore, it should be evaluated through different procedures, with different measurement methods (e.g., VAS and orthographic transcriptions), at different granularity levels (utterance, word, and phoneme), and in different speech materials. One would expect increased intelligibility measures when the degree of semantic predictability in speech materials increases. Global measures of intelligibility obtained through scalar judgments are more reliable and seem to capture the suprasegmental-level characteristics in speech, while transcription-based measures are more informative about the listeners' perception of phonemes.

8.2 Objective procedures

In this subsection, we discuss the results of our investigations of objective procedures. Specifically, we extended previous research on acoustic correlates of speech intelligibility to different intelligibility measures by

applying stepwise regression models in **Chapters 5 and 6**. We explored the possibility of addressing the low-resource problem of the non-dominant variety of the pluricentric language, Dutch, by using resources from the dominant variety in **Chapter 7**.

8.2.1 Acoustic correlates of intelligibility

Though many acoustic features have been studied and used to predict speech intelligibility or to classify speakers, we explored the relation between different acoustic features and different intelligibility measures through stepwise regression models to select the most relevant features. We also considered different speech materials. Specifically, **Chapter 5** studied a fairly small set of acoustic-phonetic features that contained three groups of features related to pitch, intensity, and formant frequencies. We investigated them in relation to speaker types (dysarthric vs healthy) by applying a stepwise logistic regression model. Based on the outcomes of the regression model, we calculated an acoustic-phonetic probability index (API) and studied its relation with two subjective intelligibility measures, i.e., scalar judgements at the utterance level and transcription-based measure at the word level. In addition to the global acoustic-phonetic features studied in **Chapter 5,** we studied other features related to Vowel Space Area (VSA) on three datasets that differ in speech material, intelligibility measure, and language, in **Appendix E**. **Chapter 6** extended the exploration to a larger feature set, eGeMAPS (Eyben et al., 2015), consisting of eighty-eight features that can be easily computed by using the opensource software openSMILE (Eyben et al., 2013). The measure of intelligibility used was a valid, phoneme-level measure, i.e., Phoneme Intelligibility. The relations between the acoustic features and Phoneme Intelligibility were investigated on two types of speech materials with varying degrees of semantic predictability to study whether the relations are material-dependent. Also, the number of speakers involved in **Chapter 6** was much larger than that in **Chapter 5**.

For the investigations in **Chapter 5**, we had the following findings. First, we found no significant difference between the mean values of acoustic-phonetic features for healthy and dysarthric speakers. This finding is in agreement with those of Feenaughty et al. (2014), who reported no significant differences for four relevant acoustic features between

healthy speakers and speakers with Parkinson's disease. In line with the interpretation in Feenaughty et al. (2014), this result may be ascribed to the mild level of dysarthria in most of the dysarthric speakers involved. On the other hand, this indicated that no single feature studied was highly related to the type of speaker, and thus, that the deviations in speech sound may occur in different aspects of speech production. Second, when applying a logistic regression model to dysarthric and healthy speakers, more than half of the acoustic-phonetic features functioned in the classification of speakers, and these functional features covered all three groups of features. Third, the API had a moderate correlation with the scalar judgments from human listeners, suggesting that these measures complement each other in classifying dysarthric and healthy speakers, while for the transcription-based intelligibility measure, no correlation was found, which may be due to the relatively low reliability of this measure. Future work could incorporate deep learning algorithms to further explore relevant acoustic features and to provide insights for diagnosis and speech therapy.

For the investigations in **Appendix E,** we found that intelligibility measures at different granularity levels were associated with different acoustic features. Weaker correlations were found between Phoneme Intelligibility and acoustic features compared to those for intelligibility measures at higher levels. Specifically, pitch variability was shown to be less related to Phoneme Intelligibility (at the phoneme level) than to the intelligibility measures at higher levels. In comparison, VSA-related features appeared to be related to intelligibility measures irrespective of granularity levels. In particular, the distance between the corner vowels and the centroid of VSA was found to be more related to the intelligibility measures than the size of VSA since the distance features showed higher correlations with the intelligibility measures and were more frequently selected by the regression models to predict the intelligibility measures. These findings for objective procedures further support the results obtained for subjective procedures that granularity levels can have an impact on the measure of intelligibility. Note that the language under study may affect the direction and strength of the relation. For example, FCR showed a stronger correlation with intelligibility in English, while VSA had stronger correlations in Dutch. As different intelligibility measures were used for

Dutch and English, future research could consider refining our findings by using the same intelligibility measures to make a fairer comparison of these acoustic features between the two languages.

For the investigations in **Chapter 6**, we had the following observations. First, acoustic features selected for dysarthric speech were different from those for healthy speech when predicting Phoneme Intelligibility. Evidence for this was that among the ten highest correlations between acoustic features and Phoneme Intelligibility, there was little overlap between the acoustic features of dysarthric and healthy speech. Another piece of evidence was that the results of the stepwise regression models were found to be different for dysarthric and healthy speech. Specifically, for dysarthric speech, the eGeMAPS feature set was effective for predicting Phoneme Intelligibility, given that more than half of the variation in scores could be explained by the features selected by the regression model. In contrast, the eGeMAPS feature set may not be so effective for healthy speech, given that only a low proportion of variance in scores could be explained by the model-selected features. Future work should explore what other feature sets are effective for healthy speech.

Second, the relation between acoustic features and Phoneme Intelligibility was shown to be material-dependent. When predicting Phoneme Intelligibility, features related to speech rate were selected by the regression models for running text, but not for isolated words. Moreover, for running text, after incorporating these additional speech rate-related features, the regression models can better explain the variation in scores for both types of speakers. These results indicate that speech rate is an important explanatory factor in connected speech and, thus, could be employed to complement the eGeMAPS feature set, while this does not seem to be the case for word lists. We also found that Phoneme Intelligibility that was extracted from the word lists does not seem to generalize to other materials. For example, the results for the two different types of speech materials showed that the correlations between Phoneme Intelligibility and acoustic features were generally higher for isolated words than for running text. In addition, the regression model for dysarthric speech achieved the largest residual values on running text. These findings observed in objective procedures further support the results for subjective procedures that speech

materials can have an impact on the measure of intelligibility. The finding that the relation was material-dependent seemed to be different from the results of Liu et al. (2022), who reported consistent, significant correlations between eGeMAPS features and speech intelligibility for three types of materials (i.e., 0.59 for five vowels, 0.66 for read speech of sentences, and 0.54 for spontaneous speech). However, this may be because in Liu et al. (2022), speech intelligibility and eGeMAPS features were derived for each type of material, whereas in our study, Phoneme Intelligibility used for running text was derived from a different type of speech material (i.e., word lists).

8.2.2 Machine learning-based models

Chapter 7 investigated the possibility to address the low-resource problem in a pluricentric context. Specifically, dysarthric speech resources from the dominant variety of the pluricentric language, Dutch, were used as additional training data. The influence on the non-dominant variety of Dutch, Flemish Dutch, was evaluated with respect to generating human-comparable transcriptions and predicting global measures of intelligibility. In general, the results suggest that the dysarthric speech resources from the dominant variety of Dutch, Netherlandic Dutch, can contribute to reducing differences with human orthographic transcriptions and increasing relevance to global comprehensibility measures. These findings were particularly evident when there were no dysarthric speech resources from the non-dominant variety, Flemish Dutch. When dysarthric speech resources from the non-dominant variety were available, the word error rate with human orthographic transcriptions as the target increased slightly due to a slight increase in substitutions and insertions. This may be due to more variations in acoustic characteristics introduced by the additional Netherlandic Dutch dysarthric speech resources. Another reason is that the speech materials of the additional Netherlandic Dutch dysarthric speech resources did not match the material in the test set. Future work can be conducted to validate these results when speech materials between the Netherlandic Dutch dysarthric speech in the training set and the Flemish dysarthric speech in the test set are matched. In general, our findings may also help future research that involves pluricentric languages beyond Dutch

and possibly other types of atypical speech, such as children speech or non-native speech.

8.3 Guidelines for measuring the intelligibility of pathological speech

Based on the research reported in this dissertation about assessing the intelligibility of pathological speech, we propose several guidelines for both *subjective* and *objective procedures*. With respect to *subjective procedures*, first, it is necessary to collect intelligibility measures using different speech materials varying in the degree of semantic predictability. This is because different speech materials can result in intelligibility measures that refer to different constructs of intelligibility. The presence or absence of the context of the target unit to be examined can also be seen as a variation in the degree of semantic predictability.

Second, measurement methods and granularity levels of intelligibility should be determined based on specific purposes in clinical practice and in research. For rapidly obtaining valid measures of speech intelligibility, scalar judgments based on VAS seem to be a promising approach. For assessments focusing on articulation, it is better to collect transcription-based intelligibility measures. These measures are also largely reliable and valid regardless of granularity levels. They can be programmatically derived with great reductions in human efforts involved in the derivation process. When comparing between speakers, such as to rank speakers or to assess their severity levels of dysarthria, intelligibility measures varied in granularity levels can be used interchangeably.

Third, it is recommended to use our novel pseudoword-allowing form of transcription to obtain transcription-based intelligibility measures from expert listeners. This form of transcription allows pseudowords, and therefore, can help to reduce the effect of contextual cues to some extent, for example, in meaningful sentences. Also, this novel form of transcription seems to provide more reliable intelligibility measures compared to the typical form of transcription, which allows only existing words. However, when recruiting naïve listeners, one should be careful in deciding whether to use this novel form of transcription. This is because naïve listeners

generally lack the ability to discern subtle pronunciation differences and, thus, may provide less reliable transcriptions.

Fourth, it is necessary to collect multiple speech samples per speaker in combination with several listeners to obtain reliable and valid intelligibility measures. For scalar judgments, having at least three samples in combination with four expert listeners is needed for reliable results regardless of speech materials. For transcription-based, word-level measures, two word lists containing around fifteen words each assessed by two expert listeners would be sufficient, whereas more samples and more listeners are required when using meaningful sentences.

Furthermore, the application of Generalizability Theory is recommended because it can provide a systematic analysis of the reliability of intelligibility measures and has the advantage of being able to handle more complex experimental designs compared to previously used methods such as ICC.

With respect to *objective procedures*, first, it is feasible to use stepwise regression models to study the relation between acoustic features and intelligibility measures. Second, acoustic features should be selected depending on the target intelligibility measure and speech material. For example, temporal features are more related to connected speech than to words uttered individually. Third, in the context of a pluricentric language, resources from dominant varieties can be used to benefit the assessment of speech intelligibility in non-dominant varieties.

8.4 Recommendations for future work

The research reported on in this dissertation has several limitations that could be addressed in future research. First, future studies could further investigate the suitability of our novel pseudoword-allowing form of transcription compared to the typical form using only existing words, e.g., across speech materials, by using the same groups of speakers. In this dissertation, we were only able to compare the two forms of transcriptions in meaningful and semantically unpredictable sentences (**Chapter 2**) and to compare the performance of our novel form of transcription between word lists and sentences (**Chapters 3 and 4**). Whether our novel form of transcription consistently outperforms across all speech materials

is unclear. Second, future research could refine the investigation of the listener experience's impact on intelligibility measures and their reliability by directly comparing different types of listeners. Third, more insights could be obtained through a more comprehensive analysis of various acoustic features either across languages or across granularity levels of intelligibility.

An important direction of future work focusing on subjective procedures would be to expand on the studies in **Chapters 2 through 4** by considering specific pathologies or diseases. This dissertation considered speakers with different pathologies as a whole due to the limitation of using an existing dataset. This limitation also resulted in the fact that we could only compare speakers at different severity levels of dysarthria. However, different dysarthria types or diseases can exhibit different speech disorders. For example, flaccid dysarthria can lead to breathy and nasal sounds as well as to difficulties in differentiating pronounced consonants, whereas hypokinetic dysarthria is notable for decreased loudness and tonal monotony (Kirshner, 2021, pp. 149–150). Such differences between dysarthria types or diseases may lead to changes in intelligibility measures that are more relevant to particular speech disorders (Tjaden et al., 2014b; Kim et al., 2011b, p. 426). Thus, studying specific pathologies or diseases and comparing them may provide more concrete insights relevant to speech therapy. In connection with this, acoustic features used in objective procedures could be tailored in terms of the specific pathologies, thus helping to improve performance in using stepwise regression models.

One direction of future work focusing on objective procedures would be to expand on using machine learning techniques in **Chapter 7** by involving pseudowords. Specifically, the ASR systems investigated in **Chapter 7** focused on generating transcriptions with only existing words. However, the results of **Chapters 2 through 4** showed that our novel form of transcription allowing also pseudowords (i.e., AWTrans) seems to be more reliable and valid than the typical form of transcription (i.e., EWTrans), which allows only existing words. Therefore, future work may consider involving AWTrans in training sets of ASR systems to generate transcriptions with pseudowords. In this way, ASR systems could be seen as a tool that can completely replace the role of human listeners in assessing intelligibility.

Related to allowing pseudowords in transcriptions, ASR outcomes could also be used to derive intelligibility measures at detailed subword levels, which concentrate more on articulations than the higher-level measures.

In addition, the experimental designs, statistical methods, various intelligibility measures, and ASR systems investigated in this dissertation can be applied for research in other disciplines or about other perceptual assessments. For instance, *intelligibility* is also a core concept in second language research (L2; Wheeler & Saito, 2022). In terms of subjective procedures, the four factors (i.e., speech materials, measurement methods, granularity levels, and listener characteristics) studied in this dissertation that showed an impact on the intelligibility of pathological speech may also affect the intelligibility of L2 speech (Kang et al., 2018). Other factors that were not manipulated in the present experiments may also be involved. Aside from the concept of *intelligibility*, other concepts in speech assessment, such as *comprehensibility* and *accentedness*, can also be investigated in research on atypical speech including pathological speech (e.g., Pommée et al., 2022) and L2 speech (e.g., Kennedy & Trofimovich, 2008; Munro & Derwing, 2001).

8.5 Conclusions

In the dissertation, we investigated various subjective and objective procedures for measuring the intelligibility of pathological speech. The effects of different factors were investigated in a comprehensive way for subjective procedures, while the acoustic correlates of intelligibility and Automatic Speech Recognition (ASR) systems were investigated for objective procedures.

For subjective procedures, we demonstrated that all four examined factors (speech materials, granularity levels, measurement methods, and listener characteristics) had an impact on the measure of intelligibility. Specifically, for speech materials, the intelligibility measures generally increased when the degrees of semantic predictability increased. For granularity levels, different intelligibility measures can be used interchangeably when averaged per speaker, but not when averaged per utterance. In particular, the scalar judgments through Visual Analogue Scales were more reliable

and robust in different speech materials compared to transcription-based, word-level measures. Phoneme-level measures were generally reliable and valid, indicating a successful reduction in human effort in deriving these measures in a programmatic manner. For measurement methods, our novel pseudoword-allowing form of transcription was shown to be a valuable tool for obtaining reliable measures and for reducing the impact of contextual cues. For listener characteristics, expert listeners seemed to provide more reliable intelligibility measures than naïve listeners. In addition, the newly applied Generalizability Theory was presented as a valuable method for studying the reliability of intelligibility measures since it can accommodate all relevant factors in experiment designs. In order to obtain reliable measures, scalar judgments require three samples per speaker in combination with four listeners irrespective of speech materials, but transcription-based, word-level measures require only two samples and two listeners in word lists.

For objective procedures, we investigated the relation between different sets of acoustic features and different intelligibility measures. We demonstrated that, first, the scalar judgments from human listeners and the acoustic-phonetic probability index seemed to complement each other in classifying dysarthric and healthy speakers. Second, the eGeMAPS feature set was shown to be effective for predicting Phoneme Intelligibility in dysarthric speech but was not in healthy speech. Third, the relation between acoustic features and intelligibility measures seemed to be material-dependent. Last, intelligibility measures at different granularity levels were associated with different acoustic features. We also investigated how to address the low-resource problem of ASR models in the pluricentric context of Dutch. We demonstrated that using dysarthric speech resources from the dominant variety of Dutch can benefit the analysis of dysarthric speech from the non-dominant variety in terms of assessing intelligibility and generating human-comparable transcriptions.

Taken together, the research in this dissertation provides valuable scientific insights and useful guidelines in developing valid procedures for measuring the intelligibility of pathological speech, which could be helpful for clinical practice and research.

REFERENCES

Abur, D., Enos, N. M., & Stepp, C. E. (2019). Visual analog scale ratings and orthographic transcription measures of sentence intelligibility in Parkinson's disease with variable listener exposure. *American Journal of Speech-Language Pathology, 28*(3), 1222–1232. https://doi.org/10.1044/2019_AJSLP-18-0275

Adank, P., Van Hout, R., & Van de Velde, H. (2007). An acoustic description of the vowels of Northern and Southern Standard Dutch II: Regional varieties. *Journal of the Acoustical Society of America, 121*(2), 1130–1141. https://doi.org/10.1121/1.1779271

Allison, K. M., Annear, L., Policicchio, M., & Hustad, K. C. (2017). Range and precision of formant movement in pediatric dysarthria. *Journal of Speech, Language, and Hearing Research, 60*(7), 1864–1876. https://doi.org/10.1044/2017_JSLHR-S-15-0438

Ansel, B. M., & Kent, R. D. (1992). Acoustic-phonetic contrasts and intelligibility in the dysarthria associated with mixed cerebral palsy. *Journal of Speech, Language, and Hearing Research, 35*(2), 296-308. https://doi.org/10.1044/jshr.3502.296

Arias-Vergara, T., Vásquez-Correa, J. C., & Orozco-Arroyave, J. R. (2017). Parkinson's disease and aging: analysis of their effect in phonation and articulation of speech. *Cognitive Computation*, *9*(6), 731-748. https://doi.org/10.1007/s12559-017-9497-x

Arias-Vergara, T., Vasquez-Correa, J. C., Orozco-Arroyave, J. R., Klumpp, P., & Nöth, E. (2018a). Unobtrusive monitoring of speech impairments of Parkinson's disease patients through mobile devices. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018*, 6004-6008. https://doi.org/10.1109/ICASSP.2018.8462332

Arias-Vergara, T., Vásquez-Correa, J. C., Orozco-Arroyave, J. R., & Nöth, E. (2018b). Speaker models for monitoring Parkinson's disease progression considering different communication channels and acoustic conditions. *Speech Communication, 101*, 11–25. https://doi.org/10.1016/j.specom.2018.05.007

Arsikere, H., Sapru, A., & Garimella, S. (2019). Multi-Dialect Acoustic Modeling Using Phone Mapping and Online i-Vectors. In *Proceedings of Interspeech 2019*, 2125–2129. https://doi.org/10.21437/Interspeech.2019-2881

Balzan, P., Vella, A., & Tattersall, C. (2019). Assessment of intelligibility in dysarthria: development of a Maltese word and phrase list. *Clinical Linguistics & Phonetics, 33*(10–11), 965-977. https://doi.org/10.1080/02699206.2019.1594383

Barefoot, S. M., Bochner, J. H., Johnson, B. A. and Eigen, B. A. (1993). Rating deaf speakers' comprehensibility: an exploratory investigation. *American Journal of Speech–Language Pathology*, *2*(3), 31–35. https://doi.org/10.1044/1058-0360.0203.31

Barreto, S. dos S., & Ortiz, K. (2008). Intelligibility measurements in speech disorders: A critical review of the literature. *Pró-Fono Revista de Atualização Científica, 20*(3), 201–206. https://doi.org/10.1590/s0104-56872008000300011

Barreto, S. D. S., & Ortiz, K. Z. (2010). Intelligibility: effects of transcription analysis and speech stimulus. *Pró-Fono Revista de Atualização Científica, 22*(2), 125-130. https://doi.org/10.1590/s0104-56872010000200010

Barreto, S. dos S., & Ortiz, K. (2016). Protocol for the evaluation of speech intelligibility in dysarthrias: evidence of reliability and validity. *Folia Phoniatrica et Logopaedica, 67*(4), 212–218. https://doi.org/10.1159/000441929

Beijer, L. (2012a). *E-learning based Speech Therapy (EST): Exploring the potentials of e-health in dysarthric speakers* [Doctoral dissertation, Radboud University]. Radboud Repository. https://hdl.handle.net/2066/101662

Beijer, L. J., Clapham, R. P., & Rietveld, A. C. M. (2012b). Evaluating the suitability of orthographic transcription and intelligibility scale rating of semantically unpredictable sentences (SUS) for speech training efficacy research in dysarthric speakers with Parkinson's disease. *Journal of Medical Speech-Language Pathology, 20*(2), 17–35.

Beijer, L. J., Rietveld, A. C. M., Ruiter M. B., & Geurts, A. C. H. (2014). Preparing an E-learning-based Speech Therapy (EST) efficacy study: Identifying suitable outcome measures to detect within-subject changes of speech intelligibility in dysarthric speakers. *Clinical Linguistics & Phonetics, 28*(12), 927–950. https://doi.org/10.3109/02699206.2014.936627

Benoît, C., Grice, M., & Hazan, V. (1996). The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. *Speech Communication, 18*(4), 381–392. https://doi.org/10.1016/0167-6393(96)00026-X

Berisha, V., Utianski, R., & Liss, J. (2013). Towards a clinical tool for automatic intelligibility assessment. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2013,* 2825–2828. https://doi.org/10.1109/ICASSP.2013.6638172

Beukelman, D. R., & Yorkston, K. M. (1979). The relationship between information transfer and speech intelligibility of dysarthric speakers. *Journal of Communication Disorders, 12*(3), 189–196. https://doi.org/10.1016/0021-9924(79)90040-6

Beukelman, D. R., & Yorkston, K. M. (1980). Influence of passage familiarity on intelligibility estimates of dysarthric speech. *Journal of Communication Disorders, 13*(1), 33–41. https://doi.org/10.1016/0021-9924(80)90019-2

Bhat, C., & Strik, H. (2020). Automatic Assessment of Sentence-Level Dysarthria Intelligibility Using BLSTM. *IEEE Journal of Selected Topics in Signal Processing, 14*(2), 322–330. https://doi.org/10.1109/JSTSP.2020.2967652

Bhat, C., Vachhani, B., & Kopparapu, S. K. (2017). Automatic assessment of dysarthria severity level using audio descriptors. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2017,* 5070–5074. https://doi.org/10.1109/ICASSP.2017.7953122

Bocklet, T., Haderlein, T., Hönig, F., Rosanowski, F., & Nöth, E. (2009). Evaluation and assessment of speech intelligibility on pathologic voices based upon acoustic speaker models. In *Proceedings of the 3rd advanced voice function assessment international workshop*, 89–92. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.639.2892&rep=rep1&type=pdf

Bocklet, T., Riedhammer, K., Nöth, E., Eysholdt, U., & Haderlein, T. (2012). Automatic intelligibility assessment of speakers after laryngeal cancer by means of acoustic modeling. *Journal of Voice*, *26*(3), 390–397. https://doi.org/10.1016/j.jvoice.2011.04.010

Boersma, P., & Weenink, D. (2021). *Praat: doing phonetics by computer* [Computer program]. Version 6.1.41, 2021, http://www.praat.org/.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

Borrie, S. A., McAuliffe, M. J., Liss, J. M., Kirk, C., O'Beirne, G. A., & Anderson, T. (2012). Familiarisation conditions and the mechanisms that underlie improved recognition of dysarthric speech. *Language and Cognitive Processes*, *27*(7–8), 1039–1055. https://doi.org/10.1080/01690965.2011.610596

Botelho, C., Teixeira, F., Rolland, T., Abad, A., & Trancoso, I. (2020). Pathological speech detection using x-vector embeddings. *arXiv preprint arXiv:2003.00864*. https://www.researchgate.net/profile/Francisco-Teixeira/publication/339641784_Pathological_speech_detection_using_x-vector_embeddings/links/5e7f57d592851caef4a7994c/Pathological-speech-detection-using-x-vector-embeddings.pdf

Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, *20*(3–4), 255–272. https://doi.org/10.1016/S0167-6393(96)00063-5

Brennan, R. L. (2001). *Generalizability theory*. Springer. https://doi.org/10.1007/978-1-4757-3456-0

Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of School Psychology*, *52*(1), 13–35. https://doi.org/10.1016/j.jsp.2013.11.008

Bunton, K., Kent, R. D., Kent, J. F., & Rosenbek, J. C. (2000). Perceptuo-acoustic assessment of prosodic impairment in dysarthria. *Clinical Linguistics & Phonetics*, *14*(1), 13–24. https://doi.org/10.1080/026992000298922

Bunton, K., Kent, R. D., Kent, J. F., & Duffy, J. R. (2001). The effects of flattening fundamental frequency contours on sentence intelligibility in speakers with dysarthria. *Clinical Linguistics & Phonetics*, *15*(3), 181–193. https://doi.org/10.1080/02699200010003378

Cannito, M. P., Suiter, D. M., Beverly, D., Chorna, L., Wolf, T., & Pfeiffer, R. M. (2012). Sentence intelligibility before and after voice treatment in speakers with idiopathic Parkinson's disease. *Journal of Voice*, *26*(2), 214–219. https://doi.org/10.1016/j.jvoice.2011.08.014

Carvalho, J., Cardoso, R., Guimarães, I., & Ferreira, J. J. (2021). Speech intelligibility of Parkinson's disease patients evaluated by different groups of healthcare professionals and naïve listeners. *Logopedics Phoniatrics Vocology*, *46*(3), 141–147. https://doi.org/10.1080/14015439.2020.1785546

Chandrashekar, H. M., Karjigi, V., & Sreedevi, N. (2019). Spectro-temporal representation of speech for intelligibility assessment of dysarthria. *IEEE Journal of Selected Topics in Signal Processing*, *14*(2), 390–399. https://doi.org/10.1109/JSTSP.2019.2949912

Chiu, Y. F., Forrest, K., & Loux, T. (2019). Relationship between F2 slope and intelligibility in Parkinson's disease: Lexical effects and listening environment. *American Journal of Speech-Language Pathology, 28*(2S), 887-894. https://doi.org/10.1044/2018_AJSLP-MSC18-18-0098

Cooke, M., Mayo, C., Valentini-Botinhao, C., Stylianou, Y., Sauert, B., & Tang, Y. (2013). Evaluating the intelligibility benefit of speech modifications in known noise conditions. *Speech Communication, 55*(4), 572-585. https://doi.org/10.1016/j.specom.2013.01.001

Cucchiarini, C. (1993). *Phonetic transcription: a methodological and empirical study* [Doctoral dissertation, Radboud University]. Radboud Repository. https://hdl.handle.net/2066/157784

Cucchiarini, C. (1996). Assessing transcription agreement: methodological aspects. *Clinical Linguistics & Phonetics, 10*(2), 131-155. https://doi.org/10.3109/02699209608985167

Cummins, N., Pan, Y., Ren, Z., Fritsch, J., Nallanthighal, V. S., Christensen, H., ... & Härmä, A. (2020). A comparison of acoustic and linguistics methodologies for Alzheimer's dementia recognition. In *Proceedings of Interspeech 2020*, 2182-2186. https://doi.org/10.21437/Interspeech.2020-2635

Cutler, A., Garcia Lecumberri, M. L., & Cooke, M. (2008). Consonant identification in noise by native and non-native listeners: Effects of local context. *The Journal of the Acoustical Society of America, 124*(2), 1264-1268. https://doi.org/10.1121/1.2946707

Dagenais, P. A., Brown, G. R., & Moore, R. E. (2006). Speech rate effects upon intelligibility and acceptability of dysarthric speech. *Clinical Linguistics & Phonetics,* 20(2-3), 141-148. https://doi.org/10.1080/02699200400026843

De Bodt, M. S., Huici, M. E. H. D., & Van De Heyning, P. H. (2002). Intelligibility as a linear combination of dimensions in dysarthric speech. *Journal of Communication Disorders, 35*(3), 283-292. https://doi.org/10.1016/S0021-9924(02)00065-5

De Bodt, M., Guns, C., & Van Nuffelen, G. (2006). *NSVO: Nederlandstalig Spraakverstaanbaarheidsonderzoek: handleiding*. Vlaamse Vereniging voor Logopedie: Herentals, Tech. Rep.

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods, 41*(2), 385-390. https://doi.org/10.3758/BRM.41.2.385

Demuynck, K., Van Compernolle, D., Van Hove, C., & Martens, J. P. (1997). *CoGeN een Corpus gesproken Nederlands voor spraaktechnologisch Onderzoek*. Final Report of CoGeN Project, Katholieke Universiteit Leuven and Universiteit Gent.

DePaul, R., & Kent, R. D. (2000). A longitudinal case study of ALS: Effects of listener familiarity and proficiency on intelligibility judgments. *American Journal of Speech-Language Pathology*, 9(3), 230-240. https://doi.org/10.1044/1058-0360.0903.230

D'Innocenzo, J., Tjaden, K., & Greenman, G. (2006). Intelligibility in dysarthria: Effects of listener familiarity and speaking condition. *Clinical Linguistics & Phonetics, 20*(9), 659-675. https://doi.org/10.1080/02699200500224272

Dongilli, P. (1994). Semantic context and speech intelligibility. In J. Till, K. Yorkston, & D. Beukelman (Eds.), *Motor speech disorders: Advances in assessment and treatment* (pp. 175–191). Baltimore, MD: Paul H. Brookes.

Efroymson, M. A. (1960). Multiple regression analysis. In A. Ralston & H. S. Wilf (Eds.), *Mathematical methods for digital computers* (pp. 191–203). Wiley: New York.

Ellis, L. W., & Fucci, D. J. (1991). Magnitude-estimation scaling of speech intelligibility: effects of listeners' experience and semantic-syntactic context. *Perceptual and Motor Skills, 73*(1), 295–305. https://doi.org/10.2466/pms.1991.73.1.295

Ellis, L., Fucci, D., Reynolds, L., & Benjamin, B. (1996). Effects of gender on listeners' judgments of speech intelligibility. *Perceptual and Motor Skills, 83*(3), 771–775. https://doi.org/10.2466/pms.1996.83.3.771

Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences*. Pacific Grove, CA: Brooks/Cole Publishing Company.

Elffers, B., Van Bael, C., & Strik, H. (2005). *ADAPT: Algorithm for Dynamic Alignment of Phonetic Transcriptions.* Technical Report at the Department of Language and Speech, Radboud University, The Netherlands.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., … & Truong, K. P. (2015). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing, 7*(2), 190–202. https://doi.org/10.1109/TAFFC.2015.2457417

Eyben, F., Weninger, F., Gross, F., & Schuller, B. (2013). Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, 835–838. https://doi.org/10.1145/2502081.2502224

Feenaughty, L., Tjaden, K., & Sussman, J. (2014). Relationship between acoustic measures and judgments of intelligibility in Parkinson's disease: A within-speaker approach. *Clinical Linguistics & Phonetics, 28*(11), 857–878. https://doi.org/10.3109/02699206.2014.921839

Feng, S., Kudina, O., Halpern, B. M., & Scharenborg, O. (2021). Quantifying bias in automatic speech recognition. *arXiv preprint arXiv:2103.15122*. https://doi.org/10.48550/arXiv.2103.15122

Ferrier, L., Shane, H., Ballard, H., Carpenter, T., & Benoit, A. (1995). Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition. *Augmentative and Alternative Communication, 11*(3), 165–175. https://doi.org/10.1080/07434619512331277289

Finizia, C., Lindström, J., & Dotevall, H. (1998). Intelligibility and perceptual ratings after treatment for laryngeal cancer: laryngectomy versus radiotherapy. *The Laryngoscope, 108*(1), 138–143. https://doi.org/10.1097/00005537-199801000-00027

Fisher, R. A. (1992). Statistical methods for research workers. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics* (pp. 66–70). Springer Series in Statistics. Springer, New York, NY. https://doi.org/10.1007/978-1-4612-4380-9_6

Fritsch, J., & Magimai-Doss, M. (2021). Utterance verification-based dysarthric speech intelligibility assessment using phonetic posterior features. *IEEE Signal Processing Letters, 28*, 224-228. https://doi.org/10.1109/LSP.2021.3050362

Ford, A. L., & Johnson, L. D. (2021). The use of Generalizability Theory to inform sampling of educator language used with preschoolers with Autism Spectrum Disorder. *Journal of Speech, Language, and Hearing Research, 64*(5), 1748-1757. https://doi.org/10.1044/2021_JSLHR-20-00586

Fox, J. & Weisberg, S. (2019). *An R Companion to Applied Regression* (3rd ed.). Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Furia, C. L., Kowalski, L. P., Latorre, M. R., Angelis, E. C., Martins, N. M., Barros, A. P., & Ribeiro, K. C. (2001). Speech intelligibility after glossectomy and speech rehabilitation. *Archives of Otolaryngology–Head & Neck Surgery, 127*(7), 877-883.

Ganzeboom, M., Bakker, M., Cucchiarini, C., & Strik, H. (2016). Intelligibility of disordered speech: global and detailed scores. In *Proceedings of Interspeech 2016*, 2503–2507. https://doi.org/10.21437/Interspeech.2016-1448

Ganzeboom, M., Bakker, M., Beijer, L., Rietveld, T., & Strik, H. (2018). Speech training for neurological patients using a serious game. *British Journal of Educational Technology, 49*(4), 761-774. https://doi.org/10.1111/bjet.12640

Garcia, J. M., & Cannito, M. P. (1996). Influence of verbal and nonverbal contexts on the sentence intelligibility of a speaker with dysarthria. *Journal of Speech, Language, and Hearing Research, 39*(4), 750-760. https://doi.org/10.1044/jshr.3904.750

Gibbon, D., Moore, R., & Winski, R. (Eds.). (1997). *Handbook of standards and resources for spoken language systems*. Walter de Gruyter.

Gu, L., Harris, J. G., Shrivastav, R., & Sapienza, C. (2005). Disordered speech assessment using automatic methods based on quantitative measures. *EURASIP Journal on Advances in Signal Processing, 2005*(9), 1400-1409. https://doi.org/10.1155/ASP.2005.1400

Gurugubelli, K., & Vuppala, A. K. (2020). Analytic phase features for dysarthric speech detection and intelligibility assessment. *Speech Communication, 121*, 1-15. https://doi.org/10.1016/j.specom.2020.04.006

Haderlein, T., Moers, C., Möbius, B., Rosanowski, F., & Nöth, E. (2011). Intelligibility Rating with Automatic Speech Recognition, Prosodic, and Cepstral Evaluation. In I. Habernal & V. Matoušek (Eds.), *Text, Speech and Dialogue* (vol. 6836, pp. 195–202). Springer Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-23538-2_25

Halpern, B. M., Feng, S., van Son, R., van den Brekel, M., & Scharenborg, O. (2022). Low-resource automatic speech recognition and error analyses of oral cancer speech. *Speech Communication, 141*, 14-27. https://doi.org/10.1016/j.specom.2022.04.006

Hammen, V. L., Yorkston, K. M., & Minifie, F. D. (1994). Effects of temporal alterations on speech intelligibility in Parkinsonian dysarthria. *Journal of Speech, Language, and Hearing Research, 37*(2), 244-253. https://doi.org/10.1044/jshr.3702.244

Harrell, F. E., (2021). *Hmisc: Harrell Miscellaneous* [Computer software]. R package version 4.6-0. https://CRAN.R-project.org/package=Hmisc

Hashemi Hosseinabad, H., Ishikawa, K., & Washington, K. (2021). Agreements between speech language pathologists and naïve listeners' judgements of Intelligibility in children with cleft palate. *Clinical Linguistics & Phonetics*, 1-19. https://doi.org/10.1080/02699206.2021.1983021

Hebbali, A. (2018). *Package olsrr* [Computer software]. https://cran.r-project.org/web/packages/olsrr/index.html

Hermann, E., & Doss, M. M. (2020). Dysarthric speech recognition with lattice-free MMI. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2020,* 6109-6113. https://doi.org/10.1109/ICASSP40776.2020.9053549

Hirsch, M. E., Thompson, A., Kim, Y., & Lansford, K. L. (2022). The Reliability and Validity of Speech-Language Pathologists' Estimations of Intelligibility in *Dysarthria. Brain Sciences, 12*(8), 1011. https://doi.org/10.3390/brainsci12081011

Hodge, M. M., & Gotzke, C. L. (2014). Construct-related validity of the TOCS measures: Comparison of intelligibility and speaking rate scores in children with and without speech disorders. *Journal of Communication Disorders, 51*, 51–63. https://doi.org/10.1016/j.jcomdis.2014.06.007

Holmes, J. R., Oates, J. M., Phyland, D. J., & Hughes, A. J. (2000). Voice characteristics in the progression of Parkinson's disease. *International Journal of Language & Communication Disorders, 35*(3), 407-418. https://doi.org/10.1080/136828200410654

Hollo, A., Staubitz, J. L., & Chow, J. C. (2020). Applying Generalizability Theory to optimize analysis of spontaneous teacher talk in elementary classrooms. *Journal of Speech, Language, and Hearing Research, 63*(6), 1947-1957. https://doi.org/10.1044/2020_JSLHR-19-00118

Hosseini-Kivanani, N., Vásquez-Correa, J. C., Stede, M., & Nöth, E. (2019). Automated cross-language intelligibility analysis of Parkinson's disease patients using speech recognition technologies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop,* 74-80. https://doi.org/10.18653/v1/P19-2010

House, A. S., Williams, C. E., Hecker, M. H., & Kryter, K. D. (1965). Articulation-testing methods: Consonantal differentiation with a closed-response set. *The Journal of the Acoustical Society of America, 37*(1), 158-166. https://doi.org/10.1121/1.1909295

Hubers, F., Cucchiarini, C., Strik, H., & Dijkstra, T. (2019). Normative data of Dutch idiomatic expressions: Subjective judgments you can bank on. *Frontiers in Psychology, 10*, Article 1075. https://doi.org/10.3389/fpsyg.2019.01075

Hustad, K. C. (2006). Estimating the intelligibility of speakers with dysarthria. *Folia Phoniatrica et Logopaedica, 58*(3), 217–228. https://doi.org/10.1159/000091735

Hustad, K. C. (2007). Effects of speech stimuli and dysarthria severity on intelligibility scores and listener confidence ratings for speakers with Cerebral Palsy. *Folia Phoniatrica et Logopaedica, 59*(6), 306–317. https://doi.org/10.1159/000108337

Hustad, K. C. (2008). The relationship between listener comprehension and intelligibility scores for speakers with dysarthria. *Journal of Speech, Language, and Hearing Research, 51*(3), 562–573. https://doi.org/10.1044/1092-4388(2008/040

Hustad, K. C., & Beukelman, D. R. (2001). Effects of linguistic cues and stimulus cohesion on intelligibility of severely dysarthric speech. *Journal of Speech, Language, and Hearing Research, 44*(3), 497–510. https://doi.org/10.1044/1092-4388(2001/039)

Hustad, K. C., & Beukelman, D. R. (2002). Listener Comprehension of Severely Dysarthric Speech. *Journal of Speech, Language, and Hearing Research, 45*(3), 545–558. https://doi.org/10.1044/1092-4388(2002/043)

Hustad, K. C., & Cahill, M. A. (2003). Effects of presentation mode and repeated familiarization on intelligibility of dysarthric speech. *American Journal of Speech-Language Pathology, 12*(2), 198–208. https://doi.org/10.1044/1058-0360(2003/066)

Ishikawa, K., de Alarcon, A., Khosla, S., Kelchner, L., Silbert, N., & Boyce, S. (2018). Predicting intelligibility deficit in dysphonic speech with cepstral peak prominence. *Annals of Otology, Rhinology & Laryngology, 127*(2), 69–78. https://doi.org/10.1177/0003489417743518

Ishikawa, K., Webster, J., & Ketring, C. (2020). Agreement between transcription- and rating-based intelligibility measurements for evaluation of dysphonic speech in noise. *Clinical Linguistics & Phonetics, 35*(10), 1–13. https://doi.org/10.1080/02699206.2020.1852602

Katz, L., & Frost, R. (1992). Reading in different orthographies: The orthographic depth hypothesis. In R. Frost & L. Katz (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 67–84). Amsterdam: New Holland.

Kang, O., Thomson, R. I., & Moran, M. (2018). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension. *Language Learning, 68*, 115–146. https://doi.org/10.1111/lang.12270

Kempler, D., & Van Lancker, D. (2002). Effect of speech task on intelligibility in dysarthria: A case study of Parkinson's disease. *Brain and Language, 80*(3), 449–464. https://doi.org/10.1006/brln.2001.2602

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review, 64*(3), 459–489.

Kent, R. D., & Kim, Y. (2011). The assessment of intelligibility in motor speech disorders. In A. Lowit & R. D. Kent (Eds.), *Assessment of motor speech disorders* (pp. 21–37). San Diego, CA: Plural.

Kent, R. D., Weismer, G., Kent, J. F., & Rosenbek, J. C. (1989). Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders, 54*(4), 482–499. https://doi.org/10.1044/jshd.5404.482

Kent, R. D., Weismer, G., Kent, J. F., Vorperian, H. K., & Duffy, J. R. (1999). Acoustic studies of dysarthric speech: Methods, progress, and potential. *Journal of communication disorders, 32*(3), 141–186.

Keshet, J. (2018). Automatic speech recognition: A primer for speech-language pathology researchers. *International Journal of Speech-Language Pathology, 20*(6), 599–609. https://doi.org/10.1080/17549507.2018.1510033

Kim, H., Hasegawa-Johnson, M., & Perlman, A. (2011a). Vowel contrast and speech intelligibility in dysarthria. *Folia Phoniatrica et Logopaedica, 63*(4), 187–194. https://doi.org/10.1159/000318881

Kim, Y., Kent, R. D., & Weismer, G. (2011b). An acoustic study of the relationships among neurologic disease, dysarthria type, and severity of dysarthria. *Journal of Speech, Language, and Hearing Research, 54, 417–429*. https://doi.org/10.1044/1092-4388(2010/ 10-0020)

Kim, M. J., & Kim, H. (2012). Combination of multiple speech dimensions for automatic assessment of dysarthric speech intelligibility. In *Proceedings of Interspeech 2012*, 1323-1326. https://doi.org/10.21437/Interspeech.2012-317

Kim, M. J., Kim, Y., & Kim, H. (2015). Automatic Intelligibility Assessment of Dysarthric Speech Using Phonologically-Structured Sparse Linear Model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23*(4), 694–704. https://doi.org/10.1109/TASLP.2015.2403619

Kim, H., & Nanney, S. (2014). Familiarization effects on word intelligibility in dysarthric speech. *Folia Phoniatrica et Logopaedica, 66*(6), 258-264. https://doi.org/10.1159/000369799

Kim, Y., Weismer, G., Kent, R. D., & Duffy, J. R. (2009). Statistical models of F2 slope in relation to severity of dysarthria. *Folia Phoniatrica et Logopaedica, 61*(6), 329-335. https://doi.org/10.1159/000252849

Kirshner, H. S. (2021). Dysarthria and apraxia of speech. *Bradley's Neurology in Clinical Practice E-Book*.

Laaridh, I., Kheder, W., Fredouille, C., & Meunier, C. (2017). Automatic Prediction of Speech Evaluation Metrics for Dysarthric Speech. In *Proceedings of Interspeech 2017*, 1834-1838. https://doi.org/10.21437/Interspeech.2017-1363

Landerl, K., & Reitsma, P. (2005). Phonological and morphological consistency in the acquisition of vowel duration spelling in Dutch and German. *Journal of Experimental Child Psychology, 92*(4), 322-344. https://doi.org/10.1016/j.jecp.2005.04.005

Lansford, K. L., & Liss, J. M. (2014). Vowel acoustics in dysarthria: Speech disorder diagnosis and classification. *Journal of Speech, Language, and Hearing Research, 57*(1), 57-67. https://doi.org/10.1044/1092-4388(2013/12-0262)

Laures, J. S., & Weismer, G. (1999). The effects of a flattened fundamental frequency on intelligibility at the sentence level. *Journal of Speech, Language, and Hearing Research, 42*(5), 1148-1156. https://doi.org/10.1044/jslhr.4205.1148

Laures-Gore, J., Russell, S., Patel, R., & Frankel, M. (2016). The Atlanta motor speech disorders corpus: motivation, development, and utility. *Folia Phoniatrica et Logopaedica, 68*(2), 99-105. https://doi.org/10.1159/000448891

Le, D., Licata, K., Persad, C., & Provost, E. M. (2016). Automatic Assessment of Speech Intelligibility for Individuals with Aphasia. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24*(11), 2187–2199. https://doi.org/10.1109/TASLP.2016.2598428

Levy, E. S., Leone, D., Moya-Gale, G., Hsu, S. C., Chen, W., & Ramig, L. O. (2016). Vowel intelligibility in children with and without dysarthria: An exploratory study. *Communication Disorders Quarterly, 37*(3), 171-179. https://doi.org/10.1177/1525740115618917

Levy, E. S., Moya-Galé, G., Chang, Y. H. M., Freeman, K., Forrest, K., Brin, M. F., & Ramig, L. A. (2020). The effects of intensive speech treatment on intelligibility in Parkinson's disease: a randomised controlled trial. *EClinicalMedicine, 24*, Article 100429. https://doi.org/10.1016/j.eclinm.2020.100429

Li, M., Shavelson, R. J., Yin, Y., Wiley, E. (2015). Generalizability Theory. In R. L. Cautin & S. O. Lilienfeld (Eds.), *The Encyclopedia of Clinical Psychology* (pp. 1-19). Hoboken, NJ: John Wiley & Sons. Inc. https://doi.org/10.1002/9781118625392.wbecp352

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News, 2*(3), 18–22.

Lin, Y., Wang, L., Dang, J., Li, S., & Ding, C. (2020). End-to-End Articulatory Modeling for Dysarthric Articulatory Attribute Detection. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2020*, 7349–7353. https://doi.org/10.1109/ICASSP40776.2020.9054233

Linville, S.E. (1996). The sound of senescence. *Journal of Voice, 10*, 190–200.

Liss, J. M., Spitzer, S. M., Caviness, J. N., & Adler, C. (2002). The effects of familiarization on intelligibility and lexical segmentation in hypokinetic and ataxic dysarthria. *The Journal of Acoustical Society of America, 112*(6), 3022–3030. https://doi.org/10.1121/1.1515793

Liss, J. M., Spitzer, S., Caviness, J. N., Adler, C., & Edwards, B. (1998). Syllabic strength and lexical boundary decisions in the perception of hypokinetic dysarthric speech. *The Journal of the Acoustical Society of America, 104*(4), 2457-2466. https://doi.org/10.1121/1.423753

Liss, J. M., Spitzer, S. M., Caviness, J. N., Adler, C., & Edwards, B. W. (2000). Lexical boundary error analysis in hypokinetic and ataxic dysarthria. *The Journal of the Acoustical Society of America, 107*(6), 3415-3424. https://doi.org/10.1121/1.429412

Liu, H. M., Tsao, F. M., & Kuhl, P. K. (2005). The effect of reduced vowel working space on speech intelligibility in Mandarin-speaking young adults with cerebral palsy. *The Journal of the Acoustical Society of America, 117*(6), 3879-3889. https://doi.org/10.1121/1.1898623

Liu, Y., Reddy, M. K., Penttilä, N., Ihalainen, T., Alku, P., & Räsänen, O. (2022). Automatic Assessment of Parkinson's Disease Using Speech Representations of Phonation and Articulation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31*, 242–255. https://doi.org/10.1109/TASLP.2022.3212829

López-Pabón, F. O., Arias-Vergara, T., & Orozco-Arroyave, J. R. (2020). Cepstral analysis and Hilbert-Huang transform for automatic detection of Parkinson's disease. *TecnoLógicas*, *23*(47), 91-106. https://doi.org/10.22430/22565337.1401

Maier, A., Schuster, M., Batliner, A., Nöth, E., & Nkenke, E. (2007). Automatic scoring of the intelligibility in patients with cancer of the oral cavity. In *Proceedings of Interspeech 2007*, 1206-1209. https://doi.org/10.21437/Interspeech.2007-388

Martínez, H., Van Nuffelen, G., De Bodt, M. (2011). NSVO-Z: Nederlandstalig Spraakverstaanbaarheidsonderzoek - Zinsniveau. *Vlaamse Vereniging voor Logopedisten.* Belsele.

Martínez, D., Green, P., & Christensen, H. (2013). Dysarthria intelligibility assessment in a factor analysis total variability space. In *Proceedings of Interspeech 2013*, 2133-2137. https://doi.org/10.21437/Interspeech.2013-505

Martínez, D., Lleida, E., Green, P., Christensen, H., Ortega, A., & Miguel, A. (2015). Intelligibility assessment and speech recognizer word accuracy rate prediction for dysarthric speakers in a factor analysis subspace. *ACM Transactions on Accessible Computing, 6*(3), 1–21. https://doi.org/10.1145/2746405

Maruthy, S., & Raj, N. (2014). Relationship between speech intelligibility and listener effort in Malayalam-speaking individuals with hypokinetic dysarthria. *Speech, Language and Hearing, 17*(4), 237-245. https://doi.org/10.1179/2050571X14Z.00000000057

Max, M. (2022). *caret: Classification and Regression Training* [Computer software]. R package version 6.0-91 . https://CRAN.R-project.org/package=caret

McRae, P. A., Tjaden, K., & Schoonings, B. (2002). Acoustic and perceptual consequences of articulatory rate change in Parkinson disease. *Journal of Speech, Language, and Hearing Research, 45*(1), 35-50. https://doi.org/10.1044/1092-4388(2002/003)

Mencke, E. O., Ochsner, G. J., & Testut, E. W. (1983). Listener judges and the speech intelligibility of deaf children. *Journal of Communication Disorders, 16*(3), 175-180. https://doi.org/10.1016/0021-9924(83)90031-X

Mendoza Ramos, V., Paulyn, C., Van den Steen, L., Hernandez-Diaz Huici, M. E., De Bodt, M., & Van Nuffelen, G. (2021a). Effect of boost articulation therapy (BArT) on intelligibility in adults with dysarthria. *International Journal of Language & Communication Disorders, 56*(2), 271-282. https://doi.org/10.1111/1460-6984.12595

Mendoza Ramos, V., Vasquez-Correa, J. C., Cremers, R., Van Den Steen, L., Nöth, E., De Bodt, M., & Van Nuffelen, G. (2021b) Automatic boost articulation therapy in adults with dysarthria: Acceptability, usability and user interaction. *International Journal of Language & Communication Disorders, 56*(5), 892-906. https://doi.org/10.1111/1460-6984.12647

Mendoza Ramos, V., Hernandez-Diaz, H. A. K., Huici, M. E. H. D., Martens, H., Van Nuffelen, G., & De Bodt, M. (2020). Acoustic features to characterize sentence accent production in dysarthric speech. *Biomedical Signal Processing and Control, 57*, Article 101750. https://doi.org/10.1016/j.bspc.2019.101750

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. and Leisch F. (2021). *e1071: Misc Functions of the Department of Statistics* [Computer software]. R package version 1.7-6. https://CRAN.R-project.org/package=e1071

Middag, C. (2012). *Automatic analysis of pathological speech* [Doctoral dissertation, Ghent University]. Ghent University. http://hdl.handle.net/1854/LU-3007443

Middag, C., Martens, J., Van Nuffelen, G., & De Bodt, M. (2009a). Automated intelligibility assessment of pathological speech using phonological features. *EURASIP Journal on Advances in Signal Processing, 2009*, Article 629030. https://doi.org/10.1155/2009/629030

Middag, C., Martens, J. P., Van Nuffelen, G., & De Bodt, M. (2009b). DIA: a tool for objective intelligibility assessment of pathological speech. In *Proceedings of 6th International workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 165-167. https://biblio.ugent.be/publication/828696/file/828771

Middag, C, Saeys, Y., & Martens, J. P. (2009c). Speech technology-based assessment of phoneme intelligibility in dysarthria. *International Journal of Language & Communication Disorders, 44*(5),716-730. https://doi.org/10.1080/13682820802342062

Middag, C., Van Nuffelen, G., Martens, J.-P., & De Bodt, M. (2008). Objective intelligibility assessment of pathological speakers. In *Proceedings of Interspeech 2008*, 1745-1748. https://doi.org/10.21437/Interspeech.2008-481

Middag, C., Bocklet, T., Martens, J. P., & Nöth, E. (2011). Combining phonological and acoustic ASR-free features for pathological speech intelligibility assessment. In *Proceedings of Interspeech 2011*, 3005-3008. https://doi.org/10.21437/Interspeech.2011-752

Middag, C., Saeys, Y., & Martens, J. P. (2010). Towards an ASR-free objective analysis of pathological speech. In *Proceedings of Interspeech 2010*, 294-297. https://doi.org/10.21437/Interspeech.2010-114

Miller G. A., Heise G. A., & Lichten W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology, 41*, 329–335. https://doi.org/10.1037/h0062491

Miller, N. (2013). Measuring up to speech intelligibility. *International Journal of Language & Communication Disorders, 48*(6), 601–612. https://doi.org/10.1111/1460-6984.12061

Miller, N., & Bloch, S. (2017). A survey of speech–language therapy provision for people with post-stroke dysarthria in the UK. *International Journal of Language & Communication Disorders, 52*(6), 800-815. https://doi.org/10.1111/1460-6984.12316

Miyakoda, H. (2003). Speech errors in normal and pathological speech: evidence from Japanese. *Journal of Multilingual Communication Disorders, 1*(3), 210-221. https://doi.org/10.1080/1476967031000091006

Moore, C. T. (2016). *gtheory: Apply generalizability theory with R*. https://CRAN.R-project.org/package=gtheory

Monsen, R. (1983). The oral speech intelligibility of hearing-impaired talkers. *Journal of Speech and Hearing Disorders, 48*(3),286-296. https://doi.org/10.1044/jshd.4803.286

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 45*(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech the role of speaking rate. *Studies in second language acquisition, 23*(4), 451–468.

Munro, M. J., & Derwing, T. M. (2015). A prospectus for pronunciation research in the 21st century: A point of view. *Journal of Second Language Pronunciation, 1*(1), 11–42. https://doi.org/10.1075/jslp.1.1.01mun

Muñoz-Vigueras, N., Prados-Román, E., Valenza, M. C., Granados-Santiago, M., Cabrera-Martos, I., Rodríguez-Torres, J., & Torres-Sánchez, I. (2021). Speech and language therapy treatment on hypokinetic dysarthria in Parkinson disease: Systematic review and meta-analysis. *Clinical Rehabilitation, 35*(5), 639–655. https://doi.org/10.1177/0269215520976267

Nakayama, K., Yamamoto, T., Oda, C., Sato, M., Murakami, T., & Horiguchi, S. (2020). Effectiveness of Lee Silverman Voice Treatment® LOUD on Japanese-speaking patients with Parkinson's disease. *Rehabilitation Research and Practice, 2020*, Article 6585264. https://doi.org/10.1155/2020/6585264

Narendra, N. P., & Alku, P. (2018). Dysarthric Speech Classification Using Glottal Features Computed from Non-words, Words and Sentences. In *Proceedings of Interspeech 2018*, 3403–3407. https://doi.org/10.21437/Interspeech.2018-1059

Narendra, N. P., & Alku, P. (2020). Glottal source information for pathological voice detection. *IEEE Access, 8*, 67745–67755. https://doi.org/10.1109/ACCESS.2020.2986171

Neel, A. T. (2009). Effects of loud and amplified speech on sentence and word intelligibility in Parkinson disease. *Journal of Speech, Language, and Hearing Research, 52*(4), 1021–1033. https://doi.org/10.1044/1092-4388(2008/08-0119)

Nicoras, R., Gotowiec, S., Hadley, L. V., Smeds, K., & Naylor, G. (2022). Conversation success in one-to-one and group conversation: a group concept mapping study of adults with normal and impaired hearing. *International Journal of Audiology*, 1–9. https://doi.org/10.1080/14992027.2022.2095538

Nishio, M., & Niimi, S. (2001). Speaking rate and its components in dysarthric speakers. *Clinical Linguistics & Phonetics, 15*(4), 309–317. https://doi.org/10.1080/02699200010024456

Norrby, C., Lindström, J., Nilsson, J. & Wide, C. (2020). Pluricentric Languages. In J. Östman & J. Verschueren (Eds.), *Handbook of Pragmatics* (vol. 23, pp. 201–220). Amsterdam: Benjamins. https://doi.org/10.1075/hop.23.plu1

O'Brian, S., Packman, A., Onslow, M., & O'Brian, N. (2003). Generalizability Theory II: Application to perceptual scaling of speech naturalness in adults who stutter. *Journal of Speech, Language, and Hearing Research, 46*(3), 718–723. https://doi.org/10.1044/1092-4388(2003/057)

O'Brien, M. G., Derwing, T. M., Cucchiarini, C., Hardison, D. M., Mixdorff, H., Thomson, R. I., ... & Levis, G. M. (2018). Directions for the future of technology in pronunciation research and teaching. *Journal of Second Language Pronunciation, 4*(2), 182–207. https://doi.org/10.1075/jslp.17001.obr

Odell, K., McNeil, M. R., Rosenbek, J. C., & Hunter, L. (1991). Perceptual characteristics of vowel and prosody production in apraxic, aphasic, and dysarthric speakers. *Journal of Speech, Language, and Hearing Research, 34*(1), 67–80. https://doi.org/10.1044/jshr.3401.67

Oostdijk, N. (2000). The Spoken Dutch Corpus. Overview and first Evaluation. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. http://www.lrec-conf.org/proceedings/lrec2000/pdf/110.pdf

Pan, Y., Mirheidari, B., Tu, Z., O'Malley, R., Walker, T., Venneri, A., … & Christensen, H. (2020). Acoustic feature extraction with interpretable deep neural network for neurodegenerative related disorder classification. In *Proceedings of Interspeech 2020*, 4806–4810. https://doi.org/10.21437/Interspeech.2020-2684

Parra-Gallego, L. F., Arias-Vergara, T., Vásquez-Correa, J. C., Garcia-Ospina, N., Orozco-Arroyave, J. R., & Nöth, E. (2018). Automatic intelligibility assessment of Parkinson's disease with diadochokinetic exercises. In J. C. Figueroa-García, J. G. Villegas, J. R. Orozco-Arroyave, & P. A. Maya Duque, P. (Eds.), *Applied Computer Sciences in Engineering: 5th Workshop on Engineering Applications, WEA 2018*. Communications in Computer and Information Science, 916, 223–230. Springer, Cham. https://doi.org/10.1007/978-3-030-00353-1_20

Pellegrini, T., Fontan, L., Mauclair, J., Farinas, J., Alazard-Guiu, C., Robert, M., & Gatignol, P. (2015). Automatic assessment of speech capability loss in disordered speech. *ACM Transactions on Accessible Computing, 6*(3), 1–14. https://doi.org/10.1145/2739051

Platt, L. J., Andrews, G., & Howie, P. M. (1980). Dysarthria of adult cerebral palsy: II. Phonemic analysis of articulation errors. *Journal of Speech, Language, and Hearing Research, 23*(1), 41–55. https://doi.org/10.1044/jshr.2301.41

Plomp, R. & Mimpen, A. M. (1979). Improving the reliability of testing the speech reception threshold for sentences. *Audiology, 18*, 43–52. https://doi.org/10.3109/00206097909072618

Pommée, T., Balaguer, M., Mauclair, J., Pinquier, J., & Woisard, V. (2022). Intelligibility and comprehensibility: A Delphi consensus study. *International Journal of Language & Communication Disorders, 57*(1), 21–41. https://doi.org/10.1111/1460-6984.12672

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlıcˇek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. In *IEEE 2011 workshop on Automatic Speech Recognition & Understanding (ASRU)* (pp. 1–4). IEEE Signal Processing Society.

Quintas, S., Mauclair, J., Woisard, V., & Pinquier, J. (2020). Automatic Prediction of Speech Intelligibility Based on X-Vectors in the Context of Head and Neck Cancer. In *Proceedings of Interspeech 2020*, 4976–4980. https://doi.org/10.21437/Interspeech.2020-1431

Ramig, L. O., Bonitati, C., Lemke, J., & Horii, Y. (1994). Voice treatment for patients with Parkinson disease: Development of an approach and preliminary efficacy data. *Journal of Medical Speech-Language Pathology*, 2, 191–209.

Ramig, L. O., Countryman, S., Thompson, L. L., & Horii, Y. (1995). Comparison of two forms of intensive speech treatment for Parkinson disease. *Journal of Speech, Language, and Hearing Research, 38*(6), 1232–1251. https://doi.org/10.1044/jshr.3806.1232

Riedhammer, K., Stemmer, G., Haderlein, T., Schuster, M., Rosanowski, F., Noth, E., & Maier, A. (2007). Towards robust automatic evaluation of pathologic telephone speech. In *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)* (pp. 717–722). IEEE. https://doi.org/10.1109/ASRU.2007.4430200

Rietveld, T. (2020) *Human Measurement Techniques in Speech and Language Pathology: Methods for Research and Clinical Practice*. Routledge.

R Core Team. (2020). *R: A language and environment for statistical computing* [Computer software]. http://www.R-project.org/

Revelle, W. (2019). *psych: Procedures for Psychological, Psychometric, and Personality Research* [Computer software]. https://CRAN.R-project.org/package=psych

Rosen, K., & Yampolsky, S. (2000). Automatic speech recognition and a review of its functioning with dysarthric speech. *Augmentative and Alternative Communication, 16*(1), 48–60. https://doi.org/10.1080/07434610012331278904

RStudio Team. (2020). *RStudio: Integrated Development for R* [Computer software]. http://www.rstudio.com/

Rudzicz, F., Namasivayam, A.K., & Wolff, T. (2012). The TORGO database of acoustic and articulatory speech from speakers with dysarthria. *Language Resources and Evaluation, 46*(4), 523––541. https://doi.org/10.1007/s10579-011-9145-0

Sapir, S., Ramig, L. O., & Fox, C. M. (2011). Intensive voice treatment in Parkinson's disease: Lee Silverman voice treatment. *Expert Review of Neurotherapeutics, 11*(6), 815–830. https://doi.org/10.1586/ern.11.43

Schiavetti, N. (1992). Scaling procedures for the measurement of speech intelligibility. In R. D. Kent (Ed.), *Intelligibility in speech disorders: Theory, measurement and management* (pp. 11–34). Amsterdam and Philadelphia: John Benjamins. https://doi.org/10.1075/sspcl.1.02sch

Schuster, M., Maier, A., Haderlein, T., Nkenke, E., Wohlleben, U., Rosanowski, F., ... & Nöth, E. (2006a). Evaluation of speech intelligibility for children with cleft lip and palate by means of automatic speech recognition. *International Journal of Pediatric Otorhinolaryngology, 70*(10), 1741–1747. https://doi.org/10.1016/j.ijporl.2006.05.016

Schuster, M., Haderlein, T., Nöth, E., Lohscheller, J., Eysholdt, U., & Rosanowski, F. (2006b). Intelligibility of laryngectomees' substitute speech: automatic speech recognition and subjective rating. *European Archives of Oto-Rhino-Laryngology and Head & Neck, 263*(2), 188–193. https://doi.org/10.1007/s00405-005-0974-6

Shahamiri, S. R., & Salim, S. S. B. (2014). Artificial neural networks as speech recognisers for dysarthric speech: Identifying the best-performing set of MFCC parameters and studying a speaker-independent approach. *Advanced Engineering Informatics, 28*(1), 102–110. https://doi.org/10.1016/j.aei.2014.01.001

Shavelson, R. J., & Webb, N. M. (2006). Generalizability Theory. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of Complementary Methods in Education Research* (pp. 309–322). Mahwah, NJ: Lawrence Erlbaum Associates.

Signorell, A., et al. (2021). *DescTools: Tools for Descriptive Statistics* [Computer software]. https://cran.r-project.org/package=DescTools

Smith, C. H., Patel, S., Woolley, R. L., Brady, M. C., Rick, C. E., Halfpenny, R., … & Sackley, C. M. (2019). Rating the intelligibility of dysarthric speech amongst people with Parkinson's Disease: a comparison of trained and untrained listeners. *Clinical Linguistics & Phonetics, 33*(10–11), 1063–1070. https://doi.org/10.1080/02699206.2019.1604806

Steeneken, H. J., & Houtgast, T. (1980). A physical method for measuring speech transmission quality. *The Journal of the Acoustical Society of America, 67*(1), 318–326. https://doi.org/10.1121/1.384464

Stipancic, K. L., Tjaden, K., & Wilding, G. (2016). Comparison of intelligibility measures for adults with Parkinson's disease, adults with multiple sclerosis, and healthy controls. *Journal of Speech, Language, and Hearing Research, 59*(2), 230–238. https://doi.org/10.1044/2015_JSLHR-S-15-0271

Stipancic, K. L., Palmer, K. M., Rowe, H. P., Yunusova, Y., Berry, J. D., & Green, J. R. (2021). "You say severe, I say mild": Toward an empirical classification of dysarthria severity. *Journal of Speech, Language, and Hearing Research*, *64*(12), 4718–4735. https://doi.org/10.1044/2021_JSLHR-21-00197

Strik, H., Daelemans, W., Binnenpoorte, D., Sturm, J., de Vriend, F., & Cucchiarini, C. (2002). Dutch HLT resources: From BLARK to priority lists, In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 1549–1552. Denver, USA. http://lands.let.kun.nl/TSpublic/strik/publications/

Spitzer, S. M., Liss, J. M., Caviness, J. N., & Adler, C. (2000). An exploration of familiarization effects in the perception of hypokinetic and ataxic dysarthric speech. *Journal of Medical Speech-Language Pathology, 8*(4), 285–293.

Spyns, P., & Odijk, J. (2013). *Essential speech and language technology for Dutch: Results by the STEVIN programme*. Springer Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-30910-6

Sussman, J. E., & Tjaden, K. (2012). Perceptual measures of speech from individuals with Parkinson's disease and multiple sclerosis: Intelligibility and beyond. *Journal of Speech, Language, and Hearing Research, 55*(4), 1208–1219. https://doi.org/10.1044/1092-4388(2011/11-0048)

Takashima, Y., Takiguchi, T., & Ariki, Y. (2019a). End-to-end Dysarthric Speech Recognition Using Multiple Databases. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2019*, 6395–6399. https://doi.org/10.1109/ICASSP.2019.8683803

Takashima, Y., Takashima, R., Takiguchi, T., & Ariki, Y. (2019b). Knowledge Transferability Between the Speech Data of Persons with Dysarthria Speaking Different Languages for Dysarthric Speech Recognition. *IEEE Access*, 7, 164320–164326. https://doi.org/10.1109/ACCESS.2019.2951856

Thomas-Stonell, N., Kotler, A. L., Leeper, H., & Doyle, P. (1998). Computerized speech recognition: Influence of intelligibility and perceptual consistency on recognition accuracy. *Augmentative and Alternative Communication, 14*(1), 51–56. https://doi.org/10.1080/07434619812331278196

Tjaden, K. (2007). Segmental articulation in motor speech disorders. In Weismer, G. (Ed.), *Motor speech disorders: Essays for Ray Kent*, 151–186. San Diego: Plural Pub.

Tjaden, K., Lam, J., & Wilding, G. (2013). Vowel acoustics in Parkinson's disease and multiple sclerosis: Comparison of clear, loud, and slow speaking conditions. *Journal of Speech, Language, and Hearing Research, 56*(5), 1485-1502. https://doi.org/10.1044/1092-4388(2013/12-0259)

Tjaden, K. K., & Liss, J. M. (1995a). The role of listener familiarity in the perception of dysarthric speech. *Clinical Linguistics & Phonetics, 9*(2), 139–154. https://doi.org/10.3109/02699209508985329

Tjaden, K., & Liss, J. M. (1995b). The influence of familiarity on judgments of treated speech. *American Journal of Speech-Language Pathology, 4*(1), 39–48. https://doi.org/10.1044/1058-0360.0401.39

Tjaden, K., Kain, A., & Lam, J. (2014a). Hybridizing conversational and clear Speech to investigate the source of increased intelligibility in speakers with Parkinson's disease. *Journal of Speech, Language, and Hearing Research, 57*(4), 1191–1205. https://doi.org/10.1044/2014_JSLHR-S-13-0086

Tjaden, K., Sussman, J. E., & Wilding, G. E. (2014b). Impact of clear, loud, and slow speech on scaled intelligibility and speech severity in Parkinson's disease and multiple sclerosis. *Journal of Speech, language, and hearing research, 57*(3), 779–792. https://doi.org/10.1044/2014_JSLHR-S-12-0372

Tjaden, K., Richards, E., Kuo, C., Wilding, G., & Sussman, J. (2013). Acoustic and perceptual consequences of clear and loud speech. *Folia Phoniatrica et Logopaedica, 65*(4), 214-220. https://doi.org/10.1159/000355867

Tjaden, K., & Wilding, G. E. (2004). Rate and loudness manipulations in dysarthria. *Journal of Speech, Language, and Hearing Research, 47*(4), 766-783. https://doi.org/10.1044/1092-4388(2004/058)

Tjaden, K., & Wilding, G. (2010). Effects of speaking task on intelligibility in Parkinson's disease. *Clinical Linguistics & Phonetics, 25*(2), 155–168. https://doi.org/10.3109/02699206.2010.520185

Tjaden, K., & Wilding, G. (2011). The impact of rate reduction and increased loudness on fundamental frequency characteristics in dysarthria. *Folia Phoniatrica et Logopaedica, 63*(4), 178-186. https://doi.org/10.1159/000316315

Tripathi, A., Bhosale, S., & Kopparapu, S. K. (2021). Automatic speaker independent dysarthric speech intelligibility assessment system. *Computer Speech & Language, 69*, Article 101213. https://doi.org/10.1016/j.csl.2021.101213

Turner, G. S., Tjaden, K., & Weismer, G. (1995). The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research, 38*(5), 1001-1013. https://doi.org/10.1044/jshr.3805.1001

Van de Weijer, J., & Slis, I. (1991). Nasaliteitsmeting met de Nasometer. *Logopedie en Foniatrie, 63*, 97–101.

Van de Velde, H., Kissine, M., Tops, E., van der Harst., S., & van Hout, R. (2010). Will Dutch become Flemish? Antonomous developments in Belgian Dutch. *Multilingua, 29*(3-4), 385-416. https://doi.org/10.1515/mult.2010.019

Van Lierde, K. M., De Bodt, M., Van Borsel, J., Wuyts, F. L., & Van Cauwenberge, P. van de Weijer, J., & Slis, I. (1991). Nasaliteitsmeting met de Nasometer. *Logopedie en Foniatrie, 63*, 97-101.

Van Haaften, L., Diepeveen, S., van den Engel-Hoek, L., Jonker, M., de Swart, B., & Maassen, B. (2019). The psychometric evaluation of a speech production test battery for children: the reliability and validity of the computer articulation instrument. *Journal of Speech, Language, and Hearing Research, 62*(7), 2141-2170. https://doi.org/10.1044/2018_JSLHR-S-18-0274

Van Nuffelen, G., De Bodt, M., Guns, C., Wuyts, F., & Van de Heyning, P. (2008). Reliability and clinical relevance of segmental analysis based on intelligibility assessment. *Folia Phoniatrica et Logopaedica, 60*(5), 264-268. https://doi.org/10.1159/000153433

Van Nuffelen, G., De Bodt, M., Vanderwegen, J., Van de Heyning, P., & Wuyts, F. (2010). Effect of rate control on speech production and intelligibility in dysarthria. *Folia Phoniatrica et Logopaedica, 62*(3), 110–119. https://doi.org/10.1159/000287209

Van Nuffelen, G., Middag, C., De Bodt, M., & Martens, J. P. (2009a). Speech technology-based assessment of phoneme intelligibility in dysarthria. *International Journal of Language & Communication Disorders, 44*(5), 716-730. https://doi.org/10.1080/13682820802342062

Van Nuffelen, G., De Bodt, M., Wuyts, F., & Van de Heyning, P. (2009b). The effect of rate control on speech rate and intelligibility of dysarthric speech. *Folia Phoniatrica et Logopaedica, 61*(2), 69-75. https://doi.org/10.1159/000208805

Van Riper, C., & Emerick, L. L. (1984). *Speech correction: An introduction to speech pathology and audiology*. Prentice Hall. Englewood Cliffs, NJ: Prentice-Hall.

Van Son, R., Middag, C., & Demuynck, K. (2018). Vowel Space as a Tool to Evaluate Articulation Problems. In *Proceedings of Interspeech 2018*, 357-361. https://doi.org/10.21437/Interspeech.2018-68

Vásquez-Correa, J. C., Arias-Vergara, T., Orozco-Arroyave, J. R., & Nöth, E. (2018a). A Multitask Learning Approach to Assess the Dysarthria Severity in Patients with Parkinson's Disease. In *Proceedings of Interspeech 2018*, 456-460. https://doi.org/10.21437/Interspeech.2018-1988

Vásquez-Correa, J. C., Arias-Vergara, T., Orozco-Arroyave, J. R., Eskofier, B., Klucken, J., & Nöth, E. (2019a). Multimodal assessment of Parkinson's disease: a deep learning approach. *IEEE journal of biomedical and health informatics, 23*(4), 1618-1630. https://doi.org/10.1109/JBHI.2018.2866873

Vásquez-Correa, J. C., Arias-Vergara, T., Rios-Urrego, C. D., Schuster, M., Rusz, J., Orozco-Arroyave, J. R., & Nöth, E. (2019b). Convolutional neural networks and a transfer learning strategy to classify Parkinson's disease from speech in three different languages. In I. Nyström, Y. Hernández Heredia, V. Milián Núñez (Eds.), *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: Iberoamerican Congress on Pattern Recognition, CIARP 2019*. Lecture Notes in Computer Science, 11896, 697-706. Springer, Cham. https://doi.org/10.1007/978-3-030-33904-3_66

Vásquez-Correa, J. C., Arias-Vergara, T., Schuster, M., Orozco-Arroyave, J. R., & Nöth, E. (2020). Parallel representation learning for the classification of pathological speech: studies on Parkinson's disease and cleft lip and palate. *Speech Communication, 122*, 56-67. https://doi.org/10.1016/j.specom.2020.07.005

Vásquez-Correa, J. C., Orozco-Arroyave, J. R., Bocklet, T., & Nöth, E. (2018b). Towards an automatic evaluation of the dysarthria level of patients with Parkinson's disease. *Journal of Communication Disorders*, *76*, 21-36. https://doi.org/10.1016/j.jcomdis.2018.08.002

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* [Computer software]. https://www.stats.ox.ac.uk/pub/MASS4/

Verhoeven, J. (2005). Belgian standard dutch. *Journal of the International Phonetic Association, 35*(2), 243-247. https://doi.org/10.1017/S0025100305002173

Voiers, W. D., Sharpley, A. D., & Hehmsoth, C. J. (1973). *Research on diagnostic evaluation of speech intelligibility*. Research Report AFCRL-72-0694, Air Force Cambridge Research Laboratories. Bedford, Massachusetts.

Wang, D., Yu, J., Wu, X., Sun, L., Liu, X., & Meng, H. (2021). Improved End-to-End Dysarthric Speech Recognition via Meta-learning Based Model Re-initialization. In *Proceedings of 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. https://doi.org/10.1109/ISCSLP49672.2021.9362068

Watson, P. J., & Schlauch, R. S. (2008). The effect of fundamental frequency on the intelligibility of speech with flattened intonation contours. *American Journal of Speech-Language Pathology, 17*(4), 348-355. https://doi.org/10.1044/1058-0360(2008/07-0048)

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. *Handbook of Statistics, 26*, 81-124. https://doi.org/10.1016/S0169-7161(06)26004-8

Weismer, G. (2009). Speech intelligibility. In M. J. Ball, M. R. Perkins, N. Muller, & S. Howard (Eds.), *The Handbook of Clinical Linguistics* (pp. 568–582). Oxford, UK: Blackwell. https://doi.org/10.1002/9781444301007.ch35

Weismer, G., & Laures, J. S. (2002) Direct magnitude estimates of speech intelligibility in dysarthria: effects of a chosen standard. *Journal of Speech, Language, and Hearing Research, 45*(3), 421-433. https://doi.org/10.1044/1092-4388(2002/033)

Weismer, G., Laures, J. S., Jeng, J. Y., Kent, R. D., & Kent, J. F. (2000). Effect of speaking rate manipulations on acoustic and perceptual aspects of the dysarthria in amyotrophic lateral sclerosis. *Folia Phoniatrica et Logopaedica, 52*(5), 201–219. https://doi.org/10.1159/000021536

Weismer, G., & Liss J.M. (1991). Age and speech motor control. In D. Ripich (Ed.), *Handbook of Aging and Communication* (pp. 205–226). Austin, TX: PRO-ED.

Weismer, G., Jeng, J. Y., Laures, J. S., Kent, R. D., & Kent, J. F. (2001). Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders. *Folia Phoniatrica et Logopaedica, 53*(1), 1–18. https://doi.org/10.1159/000052649

Wells, C. S., & Wollack, J. A. (2003). *An instructor's guide to understanding test reliability*. Wisconsin: Testing & Evaluation Services, University of Wisconsin.

Wheeler, P., & Saito, K. (2022). Second Language Speech Intelligibility Revisited: Differential Roles of Phonological Accuracy, Visual Speech, and Iconic Gesture. *The Modern Language Journal, 106*(2), 429–448. https://doi.org/10.1111/modl.12779

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* [Computer software]. https://ggplot2.tidyverse.org

Winata, G. I., Cahyawijaya, S., Liu, Z., Lin, Z., Madotto, A., Xu, P., & Fung, P. (2020). Learning Fast Adaptation on Cross-Accented Speech Recognition. In *Proceedings of Interspeech 2020*, 1276–1280. https://doi.org/10.21437/Interspeech.2020-45

Xue, W., Cucchiarini, C., van Hout, R. & Strik, H. (2019). Acoustic correlates of speech intelligibility: the usability of the eGeMAPS feature set for atypical speech. In *Proceedings of SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education*, 48–52.

Xue, W., Ramos, V. M., Harmsen, W., Cucchiarini, C., van Hout, R. W. N. M., & Strik, H. (2020). Towards a comprehensive assessment of speech intelligibility for pathological speech. In *Proceedings of Interspeech*, 3146–3150. https://doi.org/10.21437/Interspeech.2020-2693

Xue, W., van Hout, R., Boogmans, F., Ganzeboom, M., Cucchiarini, C., & Strik, H. (2021a). Speech Intelligibility of Dysarthric Speech: Human Scores and Acoustic-Phonetic Features. In *Proceedings of Interspeech 2021*, 2911–2915. https://doi.org/10.21437/Interspeech.2021-1189

Xue, W., van Hout, R., Cucchiarini, C., & Strik, H. (2021b). Assessing speech intelligibility of pathological speech: test types, ratings and transcription measures. *Clinical Linguistics & Phonetics*. Advance online publication. https://doi.org/10.1080/02699206.2021.2009918

Yılmaz, E., Ganzeboom, M., Beijer, L., Cucchiarini, C., & Strik, H. (2016a). A Dutch Dysarthric Speech Database for Individualized Speech Therapy Research. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, 792–795. https://aclanthology.org/L16-1127

Yılmaz, E., Ganzeboom, M., Cucchiarini, C., & Strik, H. (2016b). Combining Non-Pathological Data of Different Language Varieties to Improve DNN-HMM Performance on Pathological Speech. In *Proceedings of Interspeech 2016*, 218–222. https://doi.org/10.21437/Interspeech.2016-109

Yılmaz, E., Ganzeboom, M., Cucchiarini, C., & Strik, H. (2017). Multi-Stage DNN Training for Automatic Recognition of Dysarthric Speech. In *Proceedings of Interspeech 2017*, 2685–2689. https://doi.org/10.21437/Interspeech.2017-303

Yoho, S. E., Borrie, S. A., Barrett, T. S., & Whittaker, D. B. (2019). Are there sex effects for speech intelligibility in American English? Examining the influence of talker, listener, and methodology. *Attention, Perception, & Psychophysics, 81*, 558-570. https://doi.org/10.3758/s13414-018-1635-3

Yorkston, K. M., & Beukelman, D. R. (1978). A comparison of techniques for measuring intelligibility of dysarthric speech. *Journal of Communication Disorders, 11*(6), 499-512. https://doi.org/10.1016/0021-9924(78)90024-2

Yorkston, K. M., & Beukelman, D. R. (1980). A clinician-judged technique for quantifying dysarthric speech based on single-word intelligibility. *Journal of Communication Disorders, 13*(1), 15-31. https://doi.org/10.1016/0021-9924(80)90018-0

Yorkston, K. M., & Beukelman, D. R. (1981). Communication efficiency of dysarthric speakers as measured by sentence intelligibility and speaking rate. *Journal of Speech and Hearing Disorders, 46*(3), 296-301. https://doi.org/10.1044/jshd.4603.296

Yorkston, K., Beukelman, D., & Tice, R. (1996a). *Sentence intelligibility test* [Measurement instrument]. Lincoln, NE: Tice Technologies.

Yorkston, K. M., Hammen, V. L., Beukelman, D. R., & Traynor, C. D. (1990). The effect of rate control on the intelligibility and naturalness of dysarthric speech. *Journal of Speech and Hearing Disorders, 55*(3), 550–560. https://doi.org/10.1044/jshd.5503.550

Yorkston, K. M., Strand, E. A., & Kennedy, M. R. (1996b). Comprehensibility of dysarthric speech: Implications for assessment and treatment planning. *American Journal of Speech-Language Pathology, 5*(1), 55–66. https://doi.org/10.1044/1058-0360.0501.55

Yuan, F., Guo, X., Wei, X., Xie, F., Zheng, J., Huang, Y., … & Wang, Q. (2020). Lee Silverman Voice Treatment for dysarthria in patients with Parkinson's disease: a systematic review and meta-analysis. *European Journal of Neurology, 27*(10), 1957-1970. https://doi.org/10.1111/ene.14399

Yue, Z., Christensen, H., & Barker, J. (2020a). Autoencoder Bottleneck Features with Multi-Task Optimisation for Improved Continuous Dysarthric Speech Recognition. In *Proceedings of Interspeech 2020*, 4581–4585. https://doi.org/10.21437/Interspeech.2020-2746

Yue, Z., Xiong, F., Christensen, H., & Barker, J. (2020b). Exploring appropriate acoustic and language modelling choices for continuous dysarthric speech recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2020*, 6094–6098. IEEE. https://doi.org/10.1109/ICASSP40776.2020.9054343

Yunusova, Y., Green, J. R., Greenwood, L., Wang, J., Pattee, G. L., & Zinman, L. (2012). Tongue movements and their acoustic consequences in amyotrophic lateral sclerosis. *Folia Phoniatrica et Logopaedica, 64*(2), 94-102. https://doi.org/10.1159/000336890

Appendix A (Chapter 2)

Table A.1. G study results for VAS in Experiment 1 with a fully crossed design *Utterance × Speaker × Listener* by using G Theory.

Effect	DF	SS	MS	var	Percent
Speaker	35	323070.5	9230.586	437.1895	65.2
Utterance	3	4281.079	1427.026	4.999707	0.7
Listener	4	18164.36	4541.091	30.17296	4.5
Speaker:Utterance	105	49394.18	470.4208	69.45239	10.4
Speaker:Listener	140	19540.87	139.5777	4.106433	0.6
Listener:Utterance	12	2155.963	179.6636	1.569845	0.2
Residual	420	51723.91	123.1522	123.1524	18.4

Note. DF – the degrees of freedom in the source; SS – the sum of squares due to the source; MS – the mean sum of squares due to the source; var – the variance components estimate; Percent – percentage of the total variance.

Table A.2. G study results for VAS in Experiment 2 with a nested design *(Utterance : Speaker) × Listener* by using G Theory.

Effect	DF	SS	MS	var	Percent
Speaker	17	112301	6605.939	414.2327	61.3
Listener	4	15941.1	3985.274	68.74486	10.2
Speaker:Utterance	36	8505.2	236.2556	23.92584	3.5
Speaker:Listener	68	18541.57	272.6702	52.01473	7.7
Residual	144	16794.13	116.6259	116.6264	17.3

Note. DF – the degrees of freedom in the source; SS – the sum of squares due to the source; MS – the mean sum of squares due to the source; var – the variance components estimate; Percent – percentage of the total variance.

Table A.3. G study results for VAS in Experiment 3 with a nested design *(Utterance : Speaker) × Listener* by using G Theory.

Effect	DF	SS	MS	var	Percent
Speaker	22	232521.8	10569.17	326.8846	53.1
Listener	4	6450.145	1612.536	10.1703	1.7
Speaker:Utterance	115	82061.47	713.578	110.705	18
Speaker:Listener	88	18394.59	209.0294	8.162901	1.3
Residual	460	73622.87	160.0497	160.0502	26

Note. DF – the degrees of freedom in the source; SS – the sum of squares due to the source; MS – the mean sum of squares due to the source; var – the variance components estimate; Percent – percentage of the total variance.

Appendix B (Chapter 4)

Table B.1 shows the detailed correlation results between each pair of phoneme-level measures in the whole set of the Sentence Experiment (N=36) and the Word Experiment (N=18). The results for the subset of the Sentence Experiment (N=18) are not shown here because they were similar to those in the whole set. The implementation of the correlation table below was conducted by using the *Hmisc* (Harrell, 2021) R package.

Table B.1. Correlations between each pair of phoneme-level intelligibility measures. The upper triangle (with light green shading) represents the results of the whole set of the Sentence Experiment (N=36). The lower triangle represents the results of the Word Experiment (N=18).

Correlation		PhonD	AcP		AcP for the place of articulation						AcP for the manner of articulation					AcP for frontness			AcP for openness		
			all	cons-onants	vowels	labial	alveolar	post-alveolar	dorsal	glottal	plosive	fricative	nasal	trill	approx-imant	front	central	back	open	middle	closed
PhonD		**0**	-0.94	-0.94	-0.92	-0.84	-0.93	-0.91	-0.86	-0.85	-0.91	-0.91	-0.90	-0.91	-0.88	-0.91	-0.85	-0.93	-0.91	-0.92	-0.87
AcP	**all**	-0.95	**0**	1.00	0.99	0.94	0.99	0.93	0.95	0.87	0.98	0.98	0.99	0.92	0.90	0.99	0.95	0.97	0.94	0.98	0.92
	conso-nants	-0.93	0.99	**0**	0.98	0.95	0.99	0.94	0.95	0.88	0.99	0.97	0.98	0.94	0.90	0.98	0.94	0.97	0.93	0.97	0.91
	vowels	-0.95	0.96	0.92	**0**	0.93	0.98	0.88	0.94	0.85	0.96	0.97	0.99	0.87	0.90	0.99	0.95	0.98	0.95	0.99	0.94
AcP for place of articulation	**labial**	-0.90	0.95	0.96	0.89	**0**	0.92	0.90	0.88	0.75	0.91	0.94	0.96	0.89	0.82	0.92	0.88	0.91	0.84	0.91	0.88
	alveolar	-0.91	0.98	0.99	0.89	0.93	**0**	0.91	0.93	0.85	0.98	0.98	0.98	0.91	0.89	0.98	0.93	0.96	0.93	0.97	0.91
	post-alveolar	-0.75	0.77	0.77	0.73	0.70	0.76	**0**	0.87	0.83	0.92	0.88	0.90	1.00	0.85	0.88	0.84	0.89	0.85	0.87	0.81
	dorsal	-0.87	0.92	0.91	0.89	0.83	0.90	0.64	**0**	0.91	0.95	0.91	0.93	0.87	0.85	0.94	0.96	0.92	0.89	0.94	0.88
	glottal	-0.03	0.05	0.01	0.14	0.09	-0.03	-0.36	0.09	**0**	0.90	0.80	0.82	0.83	0.81	0.87	0.87	0.82	0.83	0.86	0.75
AcP for manner of articulation	**plosive**	-0.87	0.96	0.98	0.87	0.95	0.97	0.77	0.86	0.01	**0**	0.95	0.96	0.91	0.88	0.97	0.94	0.94	0.92	0.96	0.87
	fricative	-0.81	0.95	0.94	0.92	0.89	0.93	0.75	0.83	0.11	0.91	**0**	0.97	0.87	0.84	0.96	0.92	0.95	0.89	0.96	0.93
	nasal	-0.76	0.85	0.86	0.77	0.84	0.85	0.45	0.90	0.09	0.83	0.73	**0**	0.90	0.88	0.98	0.93	0.97	0.91	0.98	0.93
	trill	-0.74	0.78	0.78	0.73	0.70	0.76	1.00	0.65	-0.33	0.77	0.75	0.46	**0**	0.84	0.87	0.82	0.87	0.83	0.86	0.80
	approx-imant	-0.85	0.89	0.91	0.80	0.87	0.91	0.68	0.86	-0.05	0.86	0.76	0.79	0.69	**0**	0.88	0.85	0.94	0.97	0.88	0.78
AcP for frontness	**front**	-0.88	0.91	0.88	0.95	0.80	0.86	0.75	0.82	0.09	0.84	0.91	0.71	0.74	0.72	**0**	0.95	0.97	0.93	0.99	0.91
	central	-0.83	0.78	0.74	0.84	0.66	0.73	0.53	0.79	-0.13	0.66	0.79	0.64	0.55	0.63	0.77	**0**	0.94	0.90	0.95	0.89
	back	-0.92	0.94	0.91	0.96	0.93	0.87	0.73	0.82	0.13	0.87	0.86	0.76	0.73	0.83	0.85	0.70	**0**	0.97	0.97	0.92
AcP for openness	**open**	-0.71	0.60	0.57	0.66	0.47	0.57	0.62	0.50	0.05	0.49	0.66	0.39	0.63	0.40	0.70	0.63	0.59	**0**	0.94	0.84
	middle	-0.91	0.96	0.95	0.94	0.94	0.92	0.68	0.93	0.12	0.93	0.89	0.89	0.68	0.85	0.85	0.73	0.93	0.49	**0**	0.92
	closed	-0.78	0.79	0.73	0.90	0.71	0.68	0.59	0.71	0.22	0.67	0.77	0.53	0.59	0.65	0.87	0.79	0.82	0.49	0.73	**0**

Appendix C (Chapter 4)

Since our phoneme-level measures had generally very strong correlations as shown in Chapter 4, we also explored their performance in the two classification tasks using Random Forest (RF), which is more robust to correlated independent variables than SVM. The same three groups of measures were used: (1) all 20 phoneme-level intelligibility measures, (2) two higher-level measures, and (3) combining all 22 measures. To implement RF, datasets for the three experiments were randomly split into training (70% of the data) and validation sets (30% of the data). The number of trees was set to 100, and the number of variables randomly sampled at each stage was 4. The implementation of RF was conducted by using the *randomForest* (Liaw & Wiener, 2002) R package. The results are shown in Table C.1.

Table C.1. Accuracy scores for speaker type and severity level classifications using three general feature groups in a Random Forest implementation of the three experiment datasets. Train = training speaker set; Valid = Validation speaker set.

ML algorithm	Dependent variable (output)	Independent variables (input)	Sentence Experiment (N=36)	Sentence Experiment (N=18)	Word Experiment (N=18)
			Accuracy (Train/Valid)	Accuracy (Train/Valid)	Accuracy (Train/Valid)
Random Forest	speaker type	phoneme	1/1	1/0.83	1/0.83
		high	1/1	1/0.83	1/0.83
		all	1/1	1/0.83	1/0.83
	severity level	phoneme	0.96/0.72	1/0.33	1/0.5
		high	1/0.54	1/0.33	1/0.5
		all	1/0.72	1/0.33	1/0.67

Note. "**phoneme**" – 20 phoneme-level intelligibility measures; "**high**" – 2 higher-level intelligibility measures; "**all**" – all 22 measures.

In general, we observed results with similar trends to the SVM analyses. We observed, as it is expected, higher accuracy values in speaker type classification than in severity level classification. In general, around half of the speakers were classified correctly in both classification tasks. However, the accuracy of the validation set in the subset of the Sentence Experiment was slightly lower than those in the other two sets of experiments for the severity level classification.

Further, our phoneme-level measures performed comparably well in the whole set of the Sentence Experiment and the Word Experiment, where only half of the speakers in the Sentence Experiment were involved. This, again, suggests an advantage of using word lists in clinical practice and research.

Appendix D (Chapter 4)

For severity level classification, we further applied multinomial regression to study the relationship between phoneme-level measures (as separate predictors) and SevL (the dependent variable, four categories). The outcome was evaluated by using the Nagelkerke R^2. Following this, we calculated the percentage of speakers correctly classified into the four levels of SevL in the multinomial regression analysis. The implementation of the analyses was conducted by using the *stats* (R Core Team, 2020), *nnet* (Venables & Ripley, 2002), and *DescTools* (Signorell et al., 2021) packages in R version 4.0.2.

From the results shown in Figure D.1, it is clear that more than half of the phoneme-level measures showed higher correlations and percentage values in the Word Experiment than in both sets of the Sentence Experiment. In particular, AcP for post-alveolar and trill classes showed very strong correlations (>0.80) with the SevL in the Word Experiment. Also, many more phoneme-level measures resulted in percentage values larger than 60% in the Word Experiment compared to those in both sets of the Sentence Experiment.

Regarding the percentage values for different levels of SevL, several phoneme-level measures succeed in classifying more than 50% of speakers correctly at all four levels of SevL. These measures were AcP for nasal consonants and front vowels in the whole Sentence Experiment, AcP for post-alveolar and plosive consonants in the subset of the Sentence Experiment, and AcP for fricative consonants in the Word Experiment. Furthermore, the Word Experiment showed better results in the percentage of correctly classified speakers for three SevL levels (i.e., healthy, moderate and severe) compared to those for both sets of the Sentence Experiment. The results for the mild level of SevL were surprisingly worse in the Word Experiment compared to those for both sets of the Sentence Experiment. This might be due to the mixed results between mild and moderate levels in the Word Experiment, as shown in the boxplots in Figure 4.2 and the statistical significance results in Table 4.6. Overall, these findings showed that phoneme-level measures are better when using word lists compared to using meaningful sentences according to their relatively higher correlations with severity levels of dysarthria and better performance in classifying speakers.

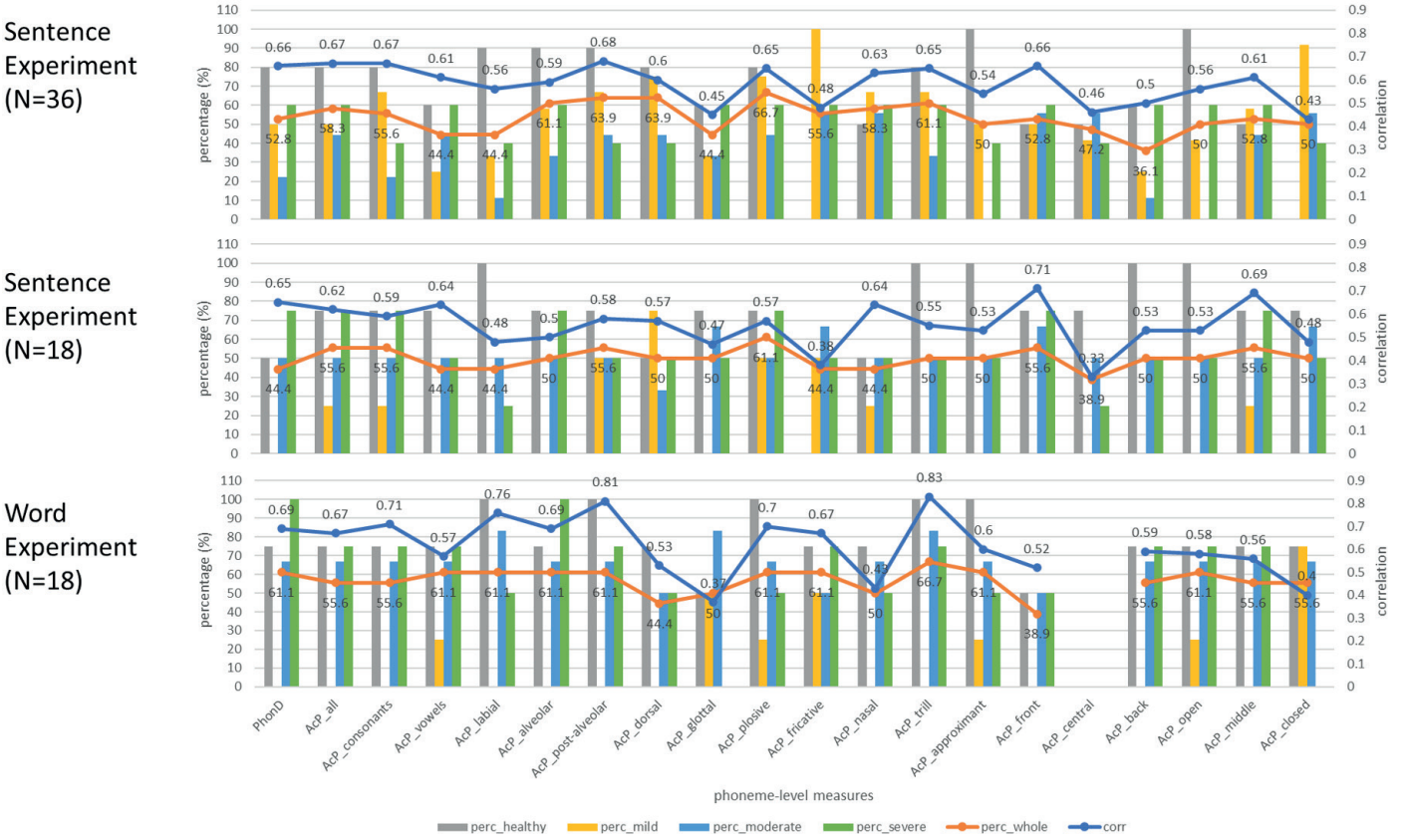


Figure D.1. Correlations (Nagelkerke R^2) of phoneme-level measures with SevL and the percentage of correctly classified speakers for speakers at four SevL levels. The blue line with markers connects the correlation results. The red line with markers connects the percentages of correctly classified speakers. Bars with different colors represent the percentages for different levels of SevL. **Best viewed in color.**

Appendix E (Chapter 5)

So far, how different acoustic features are correlated to different intelligibility measures has not been well addressed. Thus, in addition to the measures studied in Chapter 5, we further considered acoustic features related to Vowel Space Area (VSA). We studied these features in relation to three different measures of speech intelligibility in three datasets.

The three datasets were TORGO (Rudzicz et al., 2012), COPAS (Middag, 2012), and IS2016 (Ganzeboom et al., 2016) in English, Flemish Dutch, and Netherlandic Dutch, respectively. In the TORGO dataset, the subjective intelligibility measure was collected for dysarthric speakers through Frenchay Dysarthria Assessment (FDA) which has nine levels ranging from 'a' to 'e'. This nine-level range was transformed to a range of '1' to '9' for further statistical analyses. Further, healthy speakers' intelligibility scores were assigned to be fully intelligible, i.e., '1'. In the COPAS dataset, the intelligibility measure was Phoneme Intelligibility, which was calculated by the percentage of correctly transcribed target phonemes using the wordlists in the DIA task, for both healthy and dysarthric speakers. In the IS2016 dataset, the intelligibility measure was collected for dysarthric speakers and was calculated by the averaged ratings at utterance level using VAS ranging from 0 to 100. Healthy speakers were assigned to be fully intelligible, 100 in this case. An overview of the distributions of speakers regarding gender can be found in Table E.1. The boxplots of intelligibility measures for the two types of speakers (dysarthric and healthy speaker) in the three datasets can be found in Figure E.1.

Table E.1. The distributions of speakers in the three datasets.

Number of speakers in the datasets	Dysarthric speaker			Healthy speaker			In total
	Male	Female	Total	Male	Female	Total	
TORGO	5	3	8	4	3	7	15
IS2016	7	0	7	4	1	5	12
COPAS	29	20	49	33	48	81	130

Taking these three datasets was due to their different advantages and disadvantages. In detail, the COPAS dataset contains a large number of speakers compared to the other two sets, but only has a subjective intelligibility measure at the fine-grained phoneme level.

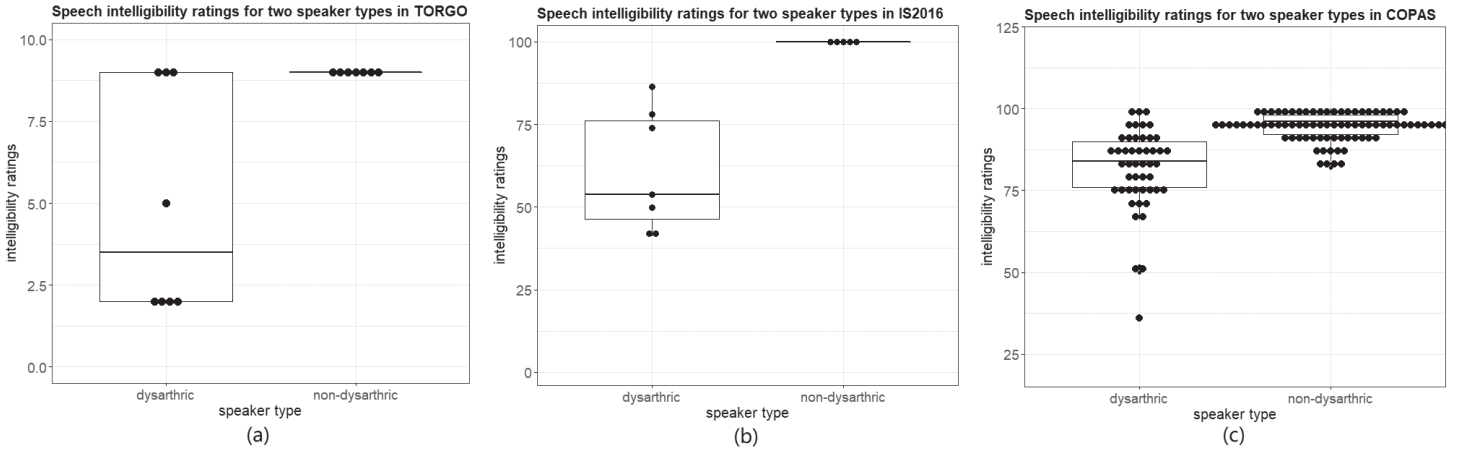


Figure E.1. Boxplots of intelligibility scores for two speaker types in (a) TORGO, (b) IS2016 and (c) COPAS.

For acoustic features, we considered VSA-related and the global features studied in Chapter 5. The global features are considered because VSA has been found to not fully predict intelligibility (Turner et al., 1995). To compute the VSA-related features, we used Praat-based VowelTriangle (Van Son et a., 2018). These features are mainly the VSA areas, the distances between corner vowels (/i/,/a/, and /u/) and the centroid of the area, and mean frequencies of the first two formant frequencies (f1 and f2) of the corner vowels. The f1 and f2 were used to calculate the other two relevant features: Vowel Articulation Index (VAI) and Formant Centralization Ratio (FCR). Global features were related to fundamental frequency, intensity and formants. Details of the acoustic measurements can be found in Table E.2.

Table E.2. Details of acoustic features.

VowelTriangle	**Global features**
Area2 (mean + 2 standard deviation)	pitchMin (minimal pitch)
Area1 (mean + 1 standard deviation)	pitchMax (maximal pitch)
i.dist (distance between /i/ and centroid)	pitchMean (mean of pitch)
a.dist (distance between /a/ and centroid)	pitchStd (standard deviation of pitch)
u.dist (distance between /u/ and centroid)	pitchVar (the mean of the absolute values of the pitch slope)
VTL (Vocal Tract Length)	IntMin (minimal intensity)
Intensity	IntMax (maximal intensity)
Slope (of F0)	IntMean (mean of intensity)
meanf1_a	IntStd (standard deviation of intensity)
meanf1_i	f1
meanf1_u	f2
meanf2_a	f3
meanf2_i	f4
meanf2_u	Center_gravity
VAI / FCR (based on f1/f2 of vowels)	

We explored the Pearson correlation coefficients between the subjective intelligibility measures and each acoustic feature. The results can be found

in Figure E.2. We also applied stepwise multiple linear regression (SMLR) models to select acoustic features to predict the subjective intelligibility measures. The results of SMLR models are shown in Table E.3.

Table E.3. The outcomes of the final selected models of stepwise multiple linear regression (SMLR).

Database	Selected features	Adjusted R-squared score
TORGO	a.dist + pitchVar + IntStd	0.773
IS2016	pitchVar + pitchMax	0.789
COPAS	i.dist + meanf2_u	0.149

As shown in Figure E.2, the correlations between the global features and intelligibility in TORGO and IS2016 were generally stronger than those in COPAS. One of the explanations is that TORGO and IS2016 adopted a rather global measure of intelligibility, while in COPAS, intelligibility was measured at the phoneme level. Regarding VSA-related features, the distance between the corner vowel /u/ and the centroid of VSA was positively correlated with the intelligibility measures in both IS2016 and COPAS, but an opposite pattern was found in TORGO. This may be due to the different languages used in these datasets. Speech to be evaluated in IS2016 and COPAS was in Dutch, but that in TORGO was in English. Although VSA was positively correlated with the intelligibility measures in IS2016 and COPAS, the distances between the corner vowels and the centroid of VSA played a more important role than the size of VSA as the distance features were more frequently selected by the SMLR models to predict the intelligibility measures, especially in TORGO and COPAS. Pitch variability was negatively correlated with the intelligibility measures in TORGO and IS2016, indicating that increasing pitch variability leads to decreasing intelligibility. This could be explained as a deficit in the mechanisms of pitch control. The lack of a correlation in COPAS between pitch variability and the intelligibility measure may be because the intelligibility measure in COPAS was at the phoneme level, indicating that the granularity level of intelligibility measures has an impact on their relation with acoustic features.

In conclusion, vowel space and pitch variability appear to be related to intelligibility measures at different granularity levels. The positions

of corner vowels could be potential predictors of intelligibility ratings. Such studies are made very difficult by the fact that different datasets and different studies use different measures of intelligibility. It would be helpful to agree on subjective measures of intelligibility for some criteria in terms of granularity levels to facilitate future research. Important insights from this comprehensive study may also be useful in analyzing speech intelligibility in the area of second language learning, with the potential for overlap and generalization.

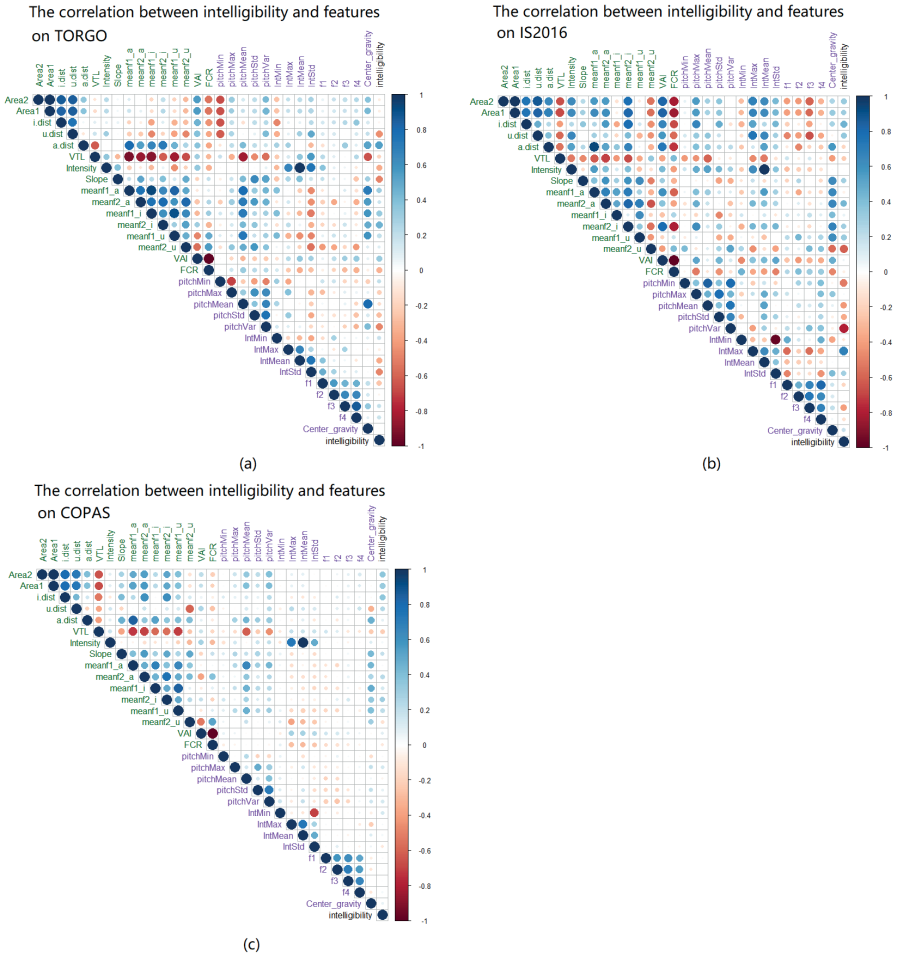


Figure E.2. Correlation plots between the subjective intelligibility measures (black) and acoustic features (VSA-related features – green, and global features – purple) on (a) TORGO, (b) IS2016 and (c) COPAS. **Best viewed in color.**

Personal data:

I collected and/or processed the following personal data: email address, gender, and experience with dysarthric speech. The email address was collected because the participants were invited as expert listeners to participate in our online listening experiments and thus received our invitations via email. The gender data was collected to conduct potential studies about its effect on intelligibility assessments. The information about their experience with dysarthric speech was used to demonstrate whether the participants were expert listeners, in contrast to naïve listeners who may result in different intelligibility measures.

Further, I used existing personal data from published datasets: speech recordings. I ensured that the speech recordings I used in my project were from speakers who had given their consent.

I ensured that I did not collect more personal data than necessary for achieving the goals of my research project. I am going to retain the personal data that was needed to answer my research questions (gender and experience with dysarthric speech) for the following 10 years.

Informed consent:

I received approval for my research project from the Ethics Assessment Committee Humanities of the Faculty of Arts and the Faculty of Philosophy, Theology and Religious Studies at Radboud University with reference number Let/MvB19U.514400 for application 2019-3197 on 11.12.2019. The listening experiments were conducted under this approval.

Since I worked with human participant data, I needed an informed consent procedure. As the participants participated in the online experiments, they were informed online before the experiments and gave their consent by clicking on a button. Such a procedure followed the informed consent procedure established by the Ethics Assessment Committee Humanities at Radboud University.

Protection of participants' privacy:

I protected the privacy of my participants by anonymizing or pseudonymizing their personal data. I did this in the following manner: all the personal data I collected were stored in a separate file; each participant was pseudonymized by using a key, which consisted of a letter and two digits, to refer to participants in the files and datasets containing their responses (intelligibility assessments). The keys were retraceable to the participants in order to comply with the Dutch law on the protection of personal data (i.e., Wet Bescherming Persoonsdata). The file holding the personal data (the key linking to participant names) was registered in accordance with the university's safety regulations for sensitive and critical data and is kept for a minimum of 10 years.

Data storage in the context of scientific integrity:

When I was gathering personal data off-campus (online), I used a secure VPN connection to transfer the data to the University's network drive. After my research, the data has been stored in Radboud Data Repository. This storage location meets the legal and ethical requirements. Safe and secure storage is guaranteed by the IT security and safety protocols. I organized my project's folder according to my institute's RDM protocol.

I followed the policy of my institute and ensured to archive the research data associated with my publication (including raw data, metadata and documentation) in a closed collection (i.e. Data Acquisition Collection) at the Radboud Data Repository (https://data.ru.nl) for a minimum of 10 years.

Further, the audio recordings to be assessed for their intelligibility were selected from the existing published dataset COPAS, and these selected audio files were also saved under the same location. The reference to the COPAS dataset is: Middag, C. (2012). Automatic analysis of pathological speech [Doctoral dissertation, Ghent University]. Ghent University. http://hdl.handle.net/1854/LU-3007443. The URL for the COPAS dataset is: https://taalmaterialen.ivdnt.org/download/tstc-corpus-pathologische-en-normale-spraak-copas/

Giving access to the data:

This project is funded by the Horizon 2020 research programme of the European Commission. There are no real requirements for the sharing of data, but the funder generally encourages sharing and re-use of research data. I followed the policy of the CLS institute to make my research data accessible via the Radboud Data Repository (https://data.ru.nl). The data is available under restricted access since I do not have permission from my participants to share the data publicly. The link to the data in the Data Acquisition Collection at the Radboud Data Repository is: https://data.ru.nl/collections/ru/cls/developing_valid_measurement_procedure_of_pathological_speech_intelligibility_dac_365 .

In connection with this, we allow interested consortium partners access to the pseudonymized data available for re-use (whenever permission is granted by the participants). Pseudonymized data (i.e., experimental data files containing intelligibility assessments through scales and orthographic transcriptions, as well as processed intelligibility measures at different granularity levels) may be shared with external researchers if access has been granted to them by the managers of the data collections (i.e., W. Xue and H. van den Heuvel).

ENGLISH SUMMARY

What distinguishes us humans from other living organisms is our ability to use language and thus communicate more effectively and freely. Language can be conveyed by speech, writing, and sign. As speech is a powerful tool in daily communications, having impairments in speech can affect human communication due to a failure in message delivery. People with dysarthria suffer from speech impairments due to neurological diseases (e.g., parkinsonism and amyotrophic lateral sclerosis) or injuries (e.g., traumatic brain injury and thrombotic/embolic stroke). Dysarthria can cause a loss of control over the muscles used for speech, resulting in disorders in speech strength, speed, range, steadiness, and tone (Duffy, 2013, p. 3). It can reduce their intelligibility, leading to difficulties in communication. As a consequence, they may lose contact with others and eventually become isolated from social life and society. These consequences severely affect their quality of life. To alleviate such speech impairments and social impacts, speech therapy has been shown to be useful. For measuring the effectiveness of therapeutical treatments and monitoring developments, e.g., through pre- and post-therapy evaluations, it is necessary to have a clear definition and a robust operationalization of speech intelligibility.

In this dissertation, intelligibility is defined in line with Hustad as "how well a speaker's acoustic signal can be accurately recovered by a listener". This definition implies that measuring speech intelligibility requires the participation of human listeners, and this procedure is therefore considered to be subjective. A typical implementation of *subjective procedures* is to conduct listening experiments in which a group of listeners are asked to assess speech intelligibility of speakers with speech impairments. The assessment of intelligibility can be performed with different *measurement methods*, i.e., scalar judgments and item identifications, for speech of different *speech materials*. Speech intelligibility can be evaluated at different *granularity levels* with respect to the units to be studied, such as graphemes (letters), phonemes, syllables, words, and sentences. Listeners recruited to participate in such experiments can either be expert listeners, such as speech-language therapists, or naïve listeners, such as college students. Many studies have shown that these procedures can produce reliable measures and they have been widely used in research and clinical practice. However, these studies are limited. First, the effects that different factors

in subjective procedures could have on measures of speech intelligibility have not been extensively analyzed so far. In particular, the comparison involving orthographic transcription between speech materials has been limited by the use of a typical form of transcription that allows only existing words. Furthermore, commonly used statistical analyses for reliability examination cannot handle all relevant factors in a procedure and different experimental designs. Also, the validity of speech intelligibility measures, a key question in research, has rarely been examined in the field of dysarthric speech.

In addition to subjective procedures, many studies have explored the possibility of using objective procedures to measure speech intelligibility where involving human listeners is not essentially required. One objective procedure focuses on studying acoustic features of dysarthric speech. The other procedure employs more sophisticated machine learning (ML) models such as automatic speech recognition (ASR) systems. However, these studies about objective procedures have several limitations. First, studies focusing on acoustic features normally investigate the relation between acoustic features with only one specific intelligibility measure. Thus, it is worthwhile to extend previous research to different intelligibility measures since they may be influenced by different implementations of the factors. This more comprehensive exploration could also help to understand how such different measures can be used to develop easy-to-use tools in clinical practice. Second, the outcomes of studies employing ML-based models are not easy for speech-language pathologists to interpret, not to mention being used for diagnosis. Furthermore, although these studies showed high correlations with subjective measures of speech intelligibility, they require large amounts of labeled data for training models, whereas low-resource is actually one of the pain points in assessing dysarthric speech.

The goal of this dissertation is to gain insights and to establish guidelines to develop valid procedures for measuring the intelligibility of pathological speech. To that end, both subjective and objective procedures were evaluated. Regarding subjective procedures, this dissertation focuses on comprehensively studying the effects of the four factors (i.e., *speech materials*, *measurement methods*, *granularity levels*, and *listener characteristics*) as well as the reliability and validity of intelligibility measures based on

the investigations in three listening experiments. These three listening experiments covered different speech materials, measurement methods, and granularity levels of intelligibility measures. Specifically, these experiments employed three types of speech materials varying in length, morphosyntactic complexity, and semantic predictability. The intelligibility measures were collected by both categories (i.e., scalar judgments and item identifications) of measurement methods - by Visual Analogue Scales (VAS) and orthographic transcriptions, respectively. For orthographic transcriptions, a novel form of transcription that allows pseudowords was proposed and compared with the typical form of transcription. Various intelligibility measures were extracted at different granularity levels, i.e., utterance, word, and subword (grapheme and phoneme). Five expert listeners were recruited to give assessments of speech intelligibility on speakers with varying severity levels of dysarthria, dysarthria type, gender, age, etc. The results of the five expert listeners in Chapters 2 through 4 were indirectly compared to eleven naïve listeners in Chapter 5, to study the effects of listener experience. Chapter 2 presents a comprehensive analysis of eight measures in the three listening experiments. Chapter 3 further studies two measures at utterance and word level, and focused on the reliability issues by applying Generalizability Theory, which has rarely been used in the field of speech intelligibility and pathology but can handle all relevant factors in experiment designs. Moreover, the usability of our novel pseudoword-allowing form of transcription was examined in depth. Chapter 4 expands the study of two types of phoneme-level measures and explored the possibility of using them to classify speakers.

For the investigation of objective procedures, this dissertation focuses on acoustic correlates of intelligibility and on addressing the low-resource problem in a pluricentric language, Dutch in this case, when ASR models are used. Specifically, Chapter 5 studies a small set of features that are related to pitch, intensity, and formant frequencies. The features are extracted from both dysarthric and healthy speech, and a stepwise logistic regression model is applied to select relevant features to classify dysarthric and healthy speech. Based on the outcomes of the regression model, we calculate an acoustic-phonetic probability index and study its relation with subjective measures of intelligibility at the utterance and word level.

Chapter 6 studies a larger acoustic feature set – eGeMAPS, including features related to e.g., frequency, amplitude, and spectrum, and its relation with a phoneme-level measure, i.e., Phoneme Intelligibility, in two types of speech materials. A set of temporal features is also considered to explore whether the relation between acoustic features and subjective intelligibility measures is material-dependent. Chapter 7 evaluates the contribution of resources from the dominant variety (Netherlandic Dutch) to improving the ASR models on the non-dominant variety (Flemish Dutch) in terms of predicting subjective measures of intelligibility and, for the first time, generating human-comparable transcriptions. The aim of studying the possibility of generating human-comparable transcriptions is to explore whether ASR models can, on the one hand, fully replace the role of human listeners in the assessment of intelligibility and, on the other hand, maintain the deviations of dysarthric speech so that therapists can further evaluate and use them for diagnosis.

The results for subjective procedures showed clearly that all four factors (i.e., *speech materials, granularity levels, measurement methods,* and *listener characteristics*) have an impact on the measure of intelligibility. Specifically, for *speech materials*, the intelligibility measures generally increase when the degrees of semantic predictability increase. For *granularity levels*, different intelligibility measures can be used interchangeably when averaged per speaker but not when averaged per utterance. In particular, the scalar judgments through VAS are more reliable and robust in different speech materials compared to transcription-based, word-level measures. Phoneme-level measures are generally reliable and valid, indicating a successful reduction in human effort in deriving these measures in a programmatic manner. For *measurement methods*, our novel pseudoword-allowing form of transcription is a valuable tool for obtaining reliable measures and for reducing the impact of contextual cues. For *listener characteristics*, expert listeners seem to provide more reliable intelligibility measures than naïve listeners. In addition, the newly applied Generalizability Theory is presented as a valuable method for studying the reliability of intelligibility measures since it can accommodate all relevant factors in experiment designs. In order to obtain reliable measures, scalar judgments require three samples per speaker in combination with four listeners irrespective of speech

materials, but transcription-based, word-level measures require only two samples and two listeners in word lists.

The results of objective procedures showed the scalar judgments from human listeners and the acoustic-phonetic probability index seemed to complement each other in classifying dysarthric and healthy speakers. Furthermore, the eGeMAPS feature set seems to be effective for predicting Phoneme Intelligibility in dysarthric speech but not effective for healthy speech. The relation between acoustic features and intelligibility measures seems to be material-dependent, and intelligibility measures at different granularity levels are associated with different acoustic features. The results for how to address the low-resource problem of ASR models in the pluricentric context of Dutch demonstrated that using dysarthric speech resources from the dominant variety of Dutch can benefit the dysarthric speech from the non-dominant variety in terms of assessing intelligibility and generating human-comparable transcriptions.

Taken together, the research in this dissertation provides insights and guidelines for developing valid procedures for measuring the intelligibility of pathological speech, which could be helpful for clinical practice and research.

NEDERLANDS SAMENVATTING

Wat mensen onderscheidt van andere levende organismen is ons vermogen om taal te gebruiken en daardoor effectiever en vrijer te communiceren. Taal kan worden overgebracht door middel van spraak, schrift en gebaren. Aangezien spraak een krachtig instrument is in de dagelijkse communicatie, kunnen spraakstoornissen de menselijke communicatie negatief beïnvloeden doordat de boodschap niet wordt overgebracht. Mensen met dysartrie lijden aan een spraakstoornis als gevolg van neurologische aandoeningen (bijv. Parkinsonisme en amyotrofe laterale sclerose) of letsel (bv. Traumatisch hersenletsel en trombotische/embolische beroerte). Dysartrie kan leiden tot verlies van controle over de spieren die gebruikt worden voor het spreken, wat resulteert in stoornissen in spraakkracht, snelheid, bereik, vastheid en toon (Duffy, 2013, p. 3). Dit kan leiden tot een verminderde spraakverstaanbaarheid, wat communicatieproblemen tot gevolg kan hebben. Als gevolg hiervan kunnen zij het contact met anderen verliezen en uiteindelijk geïsoleerd raken van het sociale leven en de maatschappij. Deze gevolgen tasten hun kwaliteit van leven ernstig aan. Om dergelijke spraakstoornissen en hun sociale gevolgen te reduceren is logopedie nuttig gebleken. Om de effectiviteit van therapeutische behandelingen te meten en ontwikkelingen te volgen, bijvoorbeeld door evaluaties voor en na de therapie, is een duidelijke definitie en een robuuste operationalisering van spraakverstaanbaarheid nodig.

In dit proefschrift wordt spraakverstaanbaarheid in navolging van Hustad gedefinieerd als “hoe goed het akoestische signaal van een spreker nauwkeurig kan worden gereconstrueerd door een luisteraar”. Deze definitie impliceert dat het meten van spraakverstaanbaarheid de deelname van menselijke luisteraars vereist, en deze procedure wordt daarom als subjectief beschouwd. Een gangbare toepassing van *subjectieve procedures* is het uitvoeren van luisterexperimenten waarbij een groep luisteraars wordt gevraagd de spraakverstaanbaarheid van sprekers met spraakstoornissen te beoordelen. De beoordeling van de verstaanbaarheid kan worden uitgevoerd met verschillende *meetmethoden*, d.w.z. subjectieve beoordelingen op een schaal en itemidentificaties, voor verschillende soorten *spraakmaterialen*. Spraakverstaanbaarheid kan worden beoordeeld op verschillende *granulariteitsniveaus* met betrekking tot de te bestuderen eenheden, zoals grafemen (letters), fonemen, lettergrepen, woorden en zinnen. Luisteraars

die voor dergelijke experimenten worden geworven, kunnen deskundige luisteraars zijn, zoals logopedisten, of naïeve luisteraars, zoals studenten. Vele studies hebben aangetoond dat deze procedures betrouwbare metingen kunnen opleveren en ze zijn op grote schaal gebruikt in onderzoek en klinische praktijk. Deze studies zijn echter beperkt. Ten eerste is tot nu toe niet uitgebreid geanalyseerd hoe verschillende factoren in subjectieve procedures de metingen van spraakverstaanbaarheid kunnen beïnvloeden. In het bijzonder is de vergelijking met orthografische transcriptie tussen spraakmaterialen beperkt door het gebruik van een typische vorm van transcriptie die alleen bestaande woorden toestaat. Bovendien kunnen veelgebruikte statistische analyses voor betrouwbaarheidsonderzoek niet alle relevante factoren in een procedure en verschillende experimentele designs combineren. Verder is de validiteit van spraakverstaanbaarheidsmetingen nauwelijks onderzocht op het gebied van dysarthrische spraak.

Naast subjectieve procedures hebben veel studies de mogelijkheid onderzocht om *objectieve procedures* te gebruiken om spraakverstaanbaarheid te meten, waarbij het betrekken van menselijke luisteraars niet essentieel is. Deze procedures richten zich op het bestuderen van akoestische kenmerken van dysartrische spraak. Andere procedures maken gebruik van geavanceerde automatische spraakherkenningssystemen (ASR). Objectieve procedures hebben echter verschillende beperkingen. Ten eerste onderzoeken studies die zich richten op akoestische kenmerken vaak de relatie tussen akoestische kenmerken en slechts één specifieke maat voor verstaanbaarheid. Het is dus de moeite waard om het onderzoek uit te breiden naar verschillende verstaanbaarheidsmaten. Dit onderzoek zou ook kunnen helpen om te begrijpen hoe zulke verschillende maten kunnen worden gebruikt om praktische instrumenten voor de klinische praktijk te ontwikkelen. Ten tweede zijn de uitkomsten van studies met ASR niet gemakkelijk te interpreteren voor logopedisten, en nog moeilijker te gebruiken voor diagnostische doeleinden. Bovendien, hoewel deze studies hoge correlaties vertoonden met subjectieve maten van spraakverstaanbaarheid, vereisen ze grote hoeveelheden gelabelde gegevens voor het trainen van modellen, terwijl het gebrek aan gegevens juist een van de pijnpunten is bij het beoordelen van dysartrische spraak.

Het doel van dit proefschrift is inzichten te verwerven om richtlijnen op te stellen voor het ontwikkelen van valide procedures voor het

meten van spraakverstaanbaarheid van pathologische spraak. Daartoe werden zowel subjectieve als objectieve procedures geëvalueerd. Voor het onderzoek van subjectieve procedures richt dit proefschrift zich op de betrouwbaarheid en validiteit van verstaanbaarheidsmetingen op basis van drie luisterexperimenten. Deze drie luisterexperimenten hadden betrekking op verschillende spraakmaterialen, meetmethoden en granulariteitsniveaus van verstaanbaarheidsmetingen. Deze experimenten maakten gebruik van drie soorten spraakmateriaal, variërend in lengte, morfosyntactische complexiteit en semantische voorspelbaarheid. De verstaanbaarheidsmetingen werden verzameld met twee meetmethoden, nl. Visual Analogue Scales (VAS) en orthografische transcripties. Voor orthografische transcripties werd een nieuwe vorm van transcriptie toegepast die pseudowoorden toestaat naast de traditionele vorm van transcriptie op grond van bestaande woorden. Er werden verschillende maten voor verstaanbaarheid geëxtraheerd op verschillende granulariteitsniveaus, d.w.z. uiting, woord en subwoord (grafeem en foneem). Vijf deskundige luisteraars beoordeelden de spraakverstaanbaarheid van sprekers met een verschillende ernst van dysartrie, type dysartrie, geslacht, leeftijd, enz. De resultaten van de vijf deskundige luisteraars worden in hoofdstuk 5 vergeleken met die van elf naïeve luisteraars, om de effecten van luisterervaring te bestuderen. Hoofdstuk 2 bevat een uitgebreide analyse van acht metingen in de drie luisterexperimenten. Hoofdstuk 3 bestudeert verder twee metingen op uiting- en woordniveau, en richt zich op de betrouwbaarheidsproblemen door toepassing van de Generalizability Theory. Bovendien wordt de bruikbaarheid van onze nieuwe pseudowoord-transcriptievorm grondig onderzocht. Hoofdstuk 4 breidt de studie van twee soorten metingen op foneemniveau uit en onderzoekt de mogelijkheid ze te gebruiken om sprekers te classificeren.

Voor het onderzoek naar objectieve procedures richt dit proefschrift zich op akoestische correlaten van verstaanbaarheid en op het aanpakken van het probleem van de geringe omvang van gegevens in een pluricentrische taal, het Nederlands in dit geval, wanneer ASR-modellen worden gebruikt. Specifiek bestudeert hoofdstuk 5 een kleine verzameling van kenmerken die gerelateerd zijn aan toonhoogte, intensiteit en formantfrequenties. De kenmerken worden geëxtraheerd uit zowel dysartrische als gezonde

spraak, en een stepwise logistic regression models wordt toegepast om relevante kenmerken te selecteren om dysartrische en gezonde spraak te classificeren. Op basis van de uitkomsten van het regressiemodel berekenen we een akoestisch-fonetische waarschijnlijkheidsindex en bestuderen we de relatie met subjectieve maten van verstaanbaarheid op het niveau van de uiting en het woord. Hoofdstuk 6 bestudeert een grotere set akoestische kenmerken - eGeMAPS, waaronder kenmerken met betrekking tot bijvoorbeeld frequentie, amplitude en spectrum, en de relatie daarvan met een maat op foneemniveau, namelijk foneemverstaanbaarheid, in twee soorten spraakmateriaal. Een reeks temporele kenmerken wordt ook bekeken om na te gaan of de relatie tussen akoestische kenmerken en subjectieve verstaanbaarheidsmaten materiaalafhankelijk is. Hoofdstuk 7 evalueert de bijdrage van bronnen van de dominante variëteit (Nederlands Nederlands) aan het verbeteren van de ASR-modellen op de niet-dominante variëteit (Vlaams Nederlands) in termen van het voorspellen van subjectieve maten van verstaanbaarheid en, voor het eerst keer, het genereren van menselijk vergelijkbare transcripties. Het doel van het onderzoek naar de mogelijkheid van het genereren van mensvergelijkbare transcripties is na te gaan of ASR-modellen enerzijds de rol van menselijke luisteraars bij de beoordeling van de verstaanbaarheid kunnen overnemen en, aan de andere kant, de afwijkingen kunnen behouden zodat therapeuten deze verder kunnen evalueren en gebruiken voor diagnose.

Uit de resultaten voor de subjectieve procedures blijkt duidelijk dat alle vier de factoren (d.w.z. *spraakmateriaal*, *granulariteitsniveaus*, *meetmethoden* en *luisteraarkenmerken*) van invloed zijn op de mate van verstaanbaarheid. In het bijzonder voor *spraakmaterialen* geldt dat de mate van verstaanbaarheid in het algemeen toeneemt wanneer de mate van semantische voorspelbaarheid toeneemt. Voor de *granulariteitsniveaus* kunnen verschillende maten van verstaanbaarheid door elkaar worden gebruikt wanneer het gemiddelde per spreker wordt genomen, maar niet wanneer het gemiddelde per uiting wordt genomen. Met name de subjectieve beoordelingen via VAS zijn betrouwbaarder en robuuster in verschillende spraakmaterialen in vergelijking met op transcriptie gebaseerde metingen op woordniveau. Maatregelen op foneemniveau zijn over het algemeen betrouwbaar en valide. Voor *meetmethoden* blijkt de pseudowoord-vorm van

transcriptie een waardevol hulpmiddel voor het verkrijgen van betrouwbare metingen en voor het verminderen van de invloed van contextuele cues. Voor *luisteraarkenmerken* blijken deskundige luisteraars betrouwbaardere verstaanbaarheidsmetingen op te leveren dan naïeve luisteraars. Verder blijkt Generalizability Theory een waardevolle insteek te zijn voor het bestuderen van de betrouwbaarheid van verstaanbaarheidsmetingen, aangezien deze alle relevante experimentele factoren kan opnemen. Om betrouwbare metingen te verkrijgen, zijn subjectieve beoordelingen van drie items per spreker in combinatie met vier luisteraars, ongeacht het spraakmateriaal vereist, maar transcriptiegebaseerde metingen op woordniveau vereisen slechts twee items en twee luisteraars in woordenlijsten.

De resultaten van de objectieve procedures toonden aan dat de subjectieve beoordelingen van menselijke luisteraars en de akoestisch-fonetische waarschijnlijkheidsindex elkaar lijken aan te vullen bij het classificeren van dysartrische en gezonde sprekers. Bovendien lijkt de eGeMAPS kenmerkenset effectief te zijn in het voorspellen van foneemverstaanbaarheid in dysartrische spraak, maar niet voor gezonde spraak. De relatie tussen akoestische kenmerken en verstaanbaarheidsmetingen lijkt materiaalafhankelijk, en verstaanbaarheidsmetingen op verschillende granulariteitsniveaus worden geassocieerd met verschillende akoestische kenmerken. De resultaten voor de aanpak van de geringe beschikbaarheid van databronnen voor ASR-modellen in de pluricentrische context van het Nederlands tonen aan dat het gebruik van dysarthrische spraakbronnen van de dominante variëteit van het Nederlands, de dysarthrische spraak van de niet-dominante variëteit ten goede kan komen wat betreft het beoordelen van de verstaanbaarheid en het genereren van transcripties.

Alles bij elkaar biedt het onderzoek in dit proefschrift inzichten en richtlijnen voor het ontwikkelen van valide procedures voor het meten van spraakverstaanbaarheid van pathologische spraak, die nuttig kunnen zijn voor de klinische praktijk en voor onderzoek.

CHINESE SUMMARY
中文摘要

人类区别于其他生物的地方在于其有能力使用语言，从而更有效和自由地进行交流。语言可以通过语言、书写和手势来传达。语言和语音是日常交流的有力工具，如果语音存在病理如有构音障碍，会因信息传递失败而影响人们的交流。构音障碍常常是由神经系统疾病（如帕金森症和肌萎缩侧索硬化症）或脑部受伤（如脑外伤和血栓性/栓塞性中风）导致的。构音障碍会导致用于说话的肌肉失去控制，从而导致说话的力度、速度、范围、稳定性和音调失调（Duffy，2013，p.3）。这些问题会导致有病理语音的人的语音的可懂度的降低，从而影响他们的交流。语音存在病理的人可能会因此逐渐与他人失去联系，最终与社会脱节。这些后果会严重地影响他们的生活质量。言语治疗已被证明在缓解这种病理语音的相关影响上是有用的。为了衡量言语治疗的效果和监测构音障碍的发展，例如通过比较治疗前后的效果，有必要对语音可懂度（speech intelligibility；又称语言可懂度、言语可懂度、语音清晰度等）进行明确的定义和有力的操作。

在这篇论文中，我们采用Hustad对可懂度的定义，即 "说话人的声音信号能被听者准确地恢复的程度"。这个定义直接表明测量语音可懂度需要人类听众的参与。我们认为这种有人类听众参与测量程序是依托于主观感知的。这种*主观测量程序*的典型实施是听觉实验，即要求一组听众评估有语言障碍的说话人的语音可懂度。可懂度的评估可以通过不同的*测量方法*，即标度判断和项目识别，对不同*语音材料*的语音进行评估。语音可懂度可以在不同的*颗粒度水平*上对所要研究的单位进行评估，如在字形（字母）、音素、音节、单词和句子水平上进行评估。被招募参与此类实验的*听众*既可以是专家级听众，如语言治疗师，也可以是没有任何经验的天真听众，如大学生。许多研究表明，这些主观测量程序可以产生可靠的测量结果，并已经在研究和临床实践中被广泛使用。然而，现有采用主观测量程序的研究有几点不足。首先，到目前为止，主观测量程序中的不同因素是如何影响语音可懂度还没有得到广泛和系统的分析。特别地，在与语音材料相关联的正字法转录（orthographic transcription）的比较上，现有的一种典型的转录形式只允许有意义的单词，从而使得使用正字法转录来比较不同语音材料上测得的语音可懂度有局限性。此外，常用的可靠性检查的统计分析不能处理主观测量程序中的所有相关因素和不同的实验设计。另一方面，语音可懂度测量的有效性作为研究中的一个关键问题，在病理语音领域很少被研究。

除了主观测量程序外，许多研究都探讨了在基本上不需要人类听众参与的情况下，使用*客观测量程序*来评估语音可懂度的可能性。客观测量程序大体上可分为两类。一类依托于病理语音的声学特征，另一类则采用了更复杂的机器学习（Machine Learning, ML）模型，如自动语音识别（Automatic Speech Recognition, ASR）系统。 然而，这些关于客观测量程序的研究有几个限制。 首先，专注于声学特征的研究通常只调查声学特征与一种特定的主观可理解性测量之间的关系。 因为通过不同的主观测量程序得到的可懂度测量可能会受到不同因素的影响而不同，因此，我们认为值得将之前关于客观测量程序的研究扩展到声学特征与不同的主观可懂度测量上。

这种更全面的探索也可以帮助了解如何在临床实践中利用这种不同的措施来开发易于使用的工具。其次，采用基于机器学习模型的客观测量程序所得到研究结果对语言病理学家来说并不容易解释，更不用说用这些结果来帮助病理学家进行诊断了。此外，尽管采用这些模型得到的结果显示出与主观语音可懂度有很高的相关性，但由于机器学习模型需要大量的标记数据来训练，构音障碍语音的低资源问题是其评估语音可懂度的痛点之一。

这篇论文的目的是为了深入了解病理语音的可懂度，并为制定有效的程序来测量病理语音的可懂度制定准则。为此，我对主观和客观测量程序都进行了评估。对于主观测量程序的调查，本论文的重点是基于三个听力实验所收集的可懂度测量的可靠性和有效性来全面研究四个主观测量程序的因素的影响。这三个听力实验涵盖了不同的语音材料、测量方法和可懂度测量的颗粒度水平。具体来说，这些实验采用了三种不同长度、不同形态句法复杂性和不同语义可预测性的语音材料。可懂度测量是通过两类测量方法收集的，即视觉模拟量表（Visual Analogue Scale, VAS）和正字法转录。对于正字法转录，我们提出了一种允许伪词（pseudowords）的新型转录形式，并将其与典型的转录形式进行比较。在不同的颗粒度水平，即语篇、单词和子词（字母和音素），上都提取了可懂度测量。我们招募了五位专家听众，对具有不同严重程度的构音障碍、不同构音障碍类型、不同性别、不同年龄等的说话者进行语音可懂度评估。五位专家听众的结果与第五章中使用的天真听众的结果进行了比较，以研究听众经验的影响。第二章对三个听觉实验中的八个可懂度测量指标进行了综合分析。第三章进一步研究了语篇和单词层面的两个可懂度测量指标，并通过应用可推广性理论（Generalizability theory）重点研究了其可靠性问题。此外，我们还深入研究了我们新颖的允许伪词的转录形式的可用性。第四章扩大了对两种的音素水平的可懂度测量指标的研究，并探讨了用它们来对说话人进行分类的可能性。

对于客观测量程序的调查，本论文着重于可理解性的声学特征和探讨在使用ASR模型的程序中如何解决多中心语言的低资源问题，并以荷兰语作为例子。具体来说，第五章研究了一个与音高、强度和频率有关的小的声学特征集。这些特征从病理和健康的语音中提取出来，并应用逐步逻辑回归模型（stepwise logistic regression model）来逐步选择相关的特征来对病理和健康的语音进行分类。基于回归模型的结果，我们计算出声学-语音概率指数（Acoustic-phonetic probability index, API），并研究其与语篇和单词层面的主观可懂度的关系。第六章研究了一个更大的声学特征集eGeMAPS，其包括与频率、振幅和频谱等有关的特征。这些特征是从两类语音材料中被提取的，并研究其与一个音素级测量，即音素可懂度（Phoneme Intelligibility，PI）的关系。我们同时考虑了一些长时特征（temporal features），以探索声学特征和主观可懂度测量之间的关系是否与材料有关。第七章评估了作为多中心语言的荷兰语的主导品种即尼德兰荷兰语（Netherlandic Dutch）的资源对改善非主导品种即弗拉

芒荷兰语（Flemish Dutch）的ASR模型在预测可懂度方面的贡献。并且，我们首次探讨了使用ASR产生的转录与从人类听众处收集的转录是否相当，以探索ASR模型是否能够一方面完全取代人类听众在评估可理解性方面的作用，另一方面保持对病理语音问题的真实呈现，以便治疗师能够进一步评估和使用它们进行诊断。

主观测量程序的结果清楚地表明，所有四个因素（即*语音材料*、*颗粒度水平*、*测量方法*和*听众特征*）都可对可懂度的测量产生影响。具体来说，从*语音材料*上来说，当语义可预测性的程度增加时，可懂度测量值也会增加。从*颗粒度水平*上来说，当在每个说话人上取平均时，不同的可懂度测量可以互换使用，但却不能在对每个语篇取平均时互换使用。特别是，在不同的语音材料中，与基于转录的、词级的测量相比，基于VAS的标度判断得到的可懂度更加可靠和稳健。音素级别的可懂度测量总体上是可靠和有效的，表明提取这些测量所采用的程序化的方式成功地减少了人工成本。从*测量方法*上来说，我们新颖的允许伪词的转录形式是一个有价值的工具，可以获得可靠的可懂度测量，并能减少上下文线索对可懂度测量的影响。从*听众特征*上来说，专家听众似乎比天真听众提供了更可靠的可懂度测量。此外，新近应用的Generalizability theory被认为是研究可懂度测量的可靠性的一个有价值的方法，因为它可以综合考虑实验设计中的所有相关因素。我们基本所收集的数据并应用Generalizability theory得到如下推论：基于VAS的标度判断不管在何种语音材料上都需要每个说话人三个样本结合四个听众以获得可靠的测量结果，但基于转录的、词级的可懂度测量在单词表这种语音材料上只需要两个样本和两个听众就足够了。

客观测量程序的结果显示，来自人类听众的标度判断和声学-语音概率指数在对有病理语音的说话人和健康的说话人进行分类时似乎可以相互补充。此外，eGeMAPS特征集似乎对预测病理语音的说话人的音素可懂度即PI很有效，但对健康的说话人却无效。此外，声学特征和可懂度测量之间的关系似乎是与语音材料相关的，且不同颗粒度水平的可懂度测量与不同的声学特征有关。关于如何在荷兰语的多中心背景下解决ASR模型的低资源问题的结果表明，在评估可懂度和生成与人类转录相当的转录方面，使用荷兰语主导品种的病理语音资源可以帮助提升其在非主导品种的病理语音上的表现。

综上所述，本论文的研究为开发测量构音障碍语音可懂度的有效程序提供了见解和指导，这对临床实践和研究都有帮助。

ACKNOWLEDGEMENTS

After finalizing my manuscript, I finally have the time to sit in front of my pandemic-induced home office moving from Nijmegen to Saarbrücken, go through the past five years in my head, and thank those who have helped me along the way and made it pleasant and memorable.

I would like to first and foremost express my deepest gratitude to my supervisors: Helmer Strik, Roeland van Hout, and Catia Cucchiarini, for their constant patience, guidance, personal attention, wisdom, and support throughout my doctoral training program. You three together form a perfect supervision team that I cannot ask for more. Helmer, thank you for picking me as a candidate for your TAPAS project. I still remember the first phone call from you asking about what happened to me in my Master's and saying that it was not my problem. Your second phone call delivering me the final decision made me cry tears of joy that I was finally recognized by someone. The question I asked at the end of your phone call: how you supervise students, has been answered vividly in the past years: I have been supported in all possible ways through this journey whenever I needed help. Thank you, Helmer, for your continuous help and support. Roeland, thank you for your endless patience whenever I had either a general question or a statistical one though you major in so many other disciplines! I appreciate very much your valuable advice and your insightful comments on my research, as well as your guidance during the past years. Your enthusiasm and continuous passion for research are infectious to me, wherever and whenever you are, including in your office with its mountains of books and on your boat swaying somewhere in the Netherlands. Catia, thank you for your invaluable advice on my research, your valuable comments on all my written drafts, which always amaze me and make me learn a lot, and your encouragement at the right moment supporting me going until the very end. I will never forget the afternoon we talked in a small garden near your house. That talk greatly changed my view of the world and how to get along with others. You shared with me your experience and feelings as a female researcher who also went through a similar situation, which is precise for my whole life.

Thank you to my funding agency: EU's H2020-ETN-ITN-MSCA Training Network on Automatic Processing of PAthological Speech (TAPAS) for setting up such an incredible research network that encourages

interdisciplinarity, collaboration, and mobility. Carrying out my doctoral research within the TAPAS project has been a piece of novel and incredible experience for me. I am very grateful to many people in the TAPAS project for organizing and formulating those wonderful events. I would like to thank the senior investigators: Aki, Alberto, Bjorn, Elmar, Gwen, Heidi, Julie, Kris, Marc, Maria, Mathew, and Rob, as well as those from the partner institutions, for all your valuable input and suggestions. Special thanks to Gwen and Marc for being amazing hosts for my secondment at your respective institute, and to their colleagues for participating in my listening experiments. Thank you for arranging my stay there, supporting my research, and especially for taking the time to arrange and help me with my listening experiments as well as participate in them with your colleagues. Thank you, Barbara, for helping arrange our various wonderful meetings. Thanks also to my fellow ESRs from around the world: Abdessalem, Bence, Camilo, Enno, Eugenia, Julian, Timothy, Tomás, Thomas, Sebastião, Shalini, Srikanth, Viviana, Yilin, Zhao, and Zhengjun for creating so much fun and inspirational memories in Eindhoven, in Lisbon, in Toulouse, in Antwerp, and online at Gather Town!

Thank you to the manuscript committee: Astrid, Esther, Fermin, Heidi, and Khiet for taking the time to carefully read my thesis and find improvements. I look forward to seeing (some of) you in person in Nijmegen in March.

Special thanks go to my two lovely paranymphs: Lotte and Yu. Lotte, having you accompany me throughout my PhD journey has been so wonderful. I am always amazed by your enthusiasm for so many different things. You are so mature, despite being so young, and helped me go through some hard times. You are open and friendly, so we can talk about research, hobbies, travelling, funny shirts & socks, TV shows, hiking, and so many other things we can't stop talking about. I used to be a very passive person in a new relationship, but your warmth makes me want to be active, just like how the wind is formed. Thank you for being my paranymph. Yu, we are similar yet different, we together did so much different yet normal things: living together, feeding my cats, watching movies, going to restaurants, shopping, and drinking hot chocolate. Thank you for those small surprises whenever you back to Nijmegen, including the cat-like coaster that is now under my mug.

Working from home for the past three years due to the covid pandemic has made me cherish and miss my easy access to the eighth floor of the Erasmus

building. Thank you goes to my officemates: Mario and Muzakki. Thank you, Mario, for inviting me to lunch on my very first day of work and making me feel welcome so that I could start my work in this new environment so smoothly. Muzakki, thank you for coming to the office and for those random talks, which made me feel not alone. I also enjoyed the fun times, random talks, and food and drinks with Aurélia, Aurora, Chantal, Chen, Claire, Elly, Emily, Ferdy, Figen, Gert-Jan, Jinbiao, Hannah, Hanno, Hongling, Katherine, Saskia, Tashi, Theresa, Thijs, Tim, Xiaoru, Xing, and many among others. Thanks to Aurora for taking me to Lotte's birthday party, which was my first time attending such a party abroad. Thanks to Chen for those random talks, Chinese Tea making, and times with her cute rabbits. Thanks also go to Chantal, Emily, Katherine, and Saskia for their suggestions when I looked for help. Thank you, Elly (and Thijs I guess), for taking care of my cats and I think they love you.

I am grateful to the Centre for Language Studies for providing an open and supportive environment for carrying out my doctoral research. Thank you to the members of the Language and Speech, Learning and Therapy and Speech Production and Comprehension groups for helping to broaden my knowledge of various research topics in linguistics and for giving me the chance to practice my presentation skills in front of a critical but fundamentally supportive crowd. Special thanks to Martha for her constant encouragement and for helping me get in touch with Helmer, leading me to this amazing journey. Thanks to Henk for your help with GDPR and other issues related to my research, and thanks to Louis for sharing your expertise about mixed effects models and building up automatic speech recognitions. Thank you also to the members of the ICT development group at CLST: Erwin, Micha, Thijs, and Wessel. Thank you for always asking me to join your 1 o'clock lunches and those fun chats about Dutch culture, funny facts, and interesting hobbies including puzzles. Special thanks to Micha for helping me build up the online listening experiment platform – dyspinas, from which I learnt how to build (a simple) web application. Thank you, Wessel, for always responding quickly to my questions and requirements and for the fun activities we had during the institute's Day Out.

I am also very grateful for being a member of the Graduate School for the Humanities and a guest at the International Max Planck Research School for Language Sciences. Thank you to Peter and Nicolet, the coordinators at

GSH, for organising such wonderful informative lunch meetings, practical workshops/courses, and fun social gatherings, which truly enriched my PhD life. I also appreciate the high-quality courses offered by IMPRS on a wide range of experimental and statistical methods.

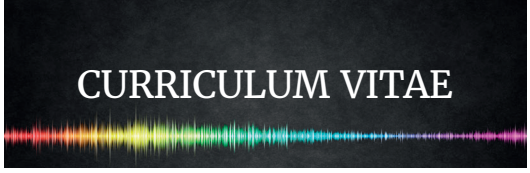
The four years I lived in Nijmegen and the Netherlands have been a great and colorful time for me. I am grateful for the great time cooking, celebrating (Chinese) festivals, travelling and being full of laughter with Chao, Chuyao, Hongtao, Jie, Liu, Manxia, Muqing, Nan, Ran, Ruifei, Ruiqi, Tiemei, Shanshan, Yahui, Yanan, Yancong, Yang, Yao, Yidong, Yuxi, Yuze, Xiangrong, Xue, Zhuoran, Zhe, Zongtian, and many among others. Thank you to my former housemate, Ruifei, for showing me around Nijmegen in my first week there, which helped me get settled in quickly, and let us have a cute cat (the grey one on the cover). Thank you, Zahar and Anum, for inviting us to your home and for making appetizing Pakistani food, which I will never forget. Having cats has also introduced me to an amazing group of cat lovers. I loved the (online) conversations (many about cats) and fun times with Aoxue, Gaonan, Kun, Lucy, Qian, Qin, Qinyun, Shurong, Stella, Xiaojiao, Xiaoyang, Zhe, Zuozuo, and many others. Thanks to Xiaojiao for taking care of my cats when I went on vacation and back home, and for the delicious food she cooked.

Ich möchte mich auch bei meinen ehemaligen Kollegen bedanken, als ich wissenschaftliche Mitarbeiterin in der Hochschule Ruhr West (HRW) war. Und ich bedanke mich auch bei meinen Freunden in Mülheim an der Ruhr in Deutschland. Eure Unterstützung und Hilfe haben mir den Mut gegeben, mein Studium fortzusetzen. Ich danke Prof. Dr.-Ing. Jörg Himmel, Prof. Dr. sc. Lothar U. Kempen, Prof. Dr.-Ing. habil. Kourosh Kolahi, Prof. Dr.-Ing. Frank Kreuder, Prof. Dr.-Ing. Zhichun Lei, Prof. Dr.-Ing. Hartmut Paschen, Prof. Dr.-Ing. Dirk Rüter, Prof. Dr. sc. techn. Klaus Thelen, and Prof. Jiadi Wang für die freundliche Hilfe. Danke Prof. Dr.-Ing. Zhichun Lei für die Vorstellung der Arbeitsmöglichkeit. Herzlichen Dank, Prof. Dr.-Ing. Jörg Himmel und Prof. Dr. sc. techn. Klaus Thelen für ihre Bereitschaft, meine Referenz zu sein. Ihr habt mir geholfen, meine Doktorandenstelle zu erhalten und meine Forschungslaufbahn fortzusetzen. Vielen Dank, Prof. Dr. sc. Lothar U. Kempen, für die Bereitstellung des Sprachproofs. Danke Brigitte für ihre Freundlichkeit, für ihre Hilfe bei der Organisation

und für die interessanten Diskussionen über Sprache. Ich schätze auch die zufälligen Gespräche, die Döner- und Pizza-Mittagessen und die lustigen Aktivitäten mit meinen Kollegen: Admir, Christoph, Christoph (ja, ich hatte zwei Kollegen, die Christoph hießen), Dawei, Dragan, Martin, Tino, und vielen anderen. Mein besonderer Dank geht an Dawei, Cui und ihren Familien, die mir das Gefühl gegeben haben, wie zu Hause zu sein. Dawei, danke, dass du mich vom Flughafen abgeholt hast, mir Wonton gekauft hast, mir geholfen hast, in eine richtige Wohnung ein zu ziehen usw. Du warst die erste Person, die ich in einer neuen Umgebung kennengelernt habe, und du bist immer nett und freundlich. Ich bin dankbar, dass ich dich als Kollege und Freund habe. Ich bin auch dankbar, dass ich viel Spaß mit meiner ehemaligen Mitbewohnerin Kefei und meinen Freunden in Mülheim hatte: Jingyi, Ping, Yinggang, Yuankai und Zhaoguo. Danke Kefei für deine Hilfe und deine Geduld, sich meine Beschwerden anzuhören, und danke deiner Familie für das köstliche hausgemachte chinesische Essen. Vielen Dank auch Ping, die immer so nett und freundlich ist. Die Zeit mit euch war so friedlich. Danke Jinyi, die so freundlich war und mich durch die schwierigen Zeiten begleitet hat.

感谢我在中国的家人、朋友们对我在外求学的支持：在这五年中我无法常常回国陪伴他们，艰难和快乐的时刻我们无法立刻分享。首先我要感谢我的妈妈，她伟大又坚强，是让我非常尊重和敬佩的女性。谢谢妈妈对我求学的支持，对我各种小固执和笨拙关心的包容，以及对我无私的爱。感谢我的小姨，在我求学在外时，对我妈妈和家人的帮助和照顾，让我少了些后顾之忧。我很想念从小带我到大的姥姥，我没能最后见到您，但您常常出现在我的梦里，您从没离开过。感谢我先生的家人：谢谢你们对我们在海外求学的支持和关心，让我时时感谢有了第二个家。

Last but not least, my beloved one, Zhengyu, like a friend and like a teacher, thank you for your constant support and encouragement. We cried in difficult times, we laughed with each other when achieving success. Without you, I would not have come to the very end. I appreciate having you by my side for the past ten years, from Hainan to Tianjin, and then to Nijmegen, together through this journey. I am very much looking forward to our unknown but, I believe, wonderful, future.



Wei Xue was born in Shijiazhuang, Hebei in China in 1992. In 2014, she completed her Bachelor's study in Telecommunication Engineering at Hainan University, China. She obtained a postgraduate recommendation from Hainan University due to her outperformance and continued her study in Information and Communication Engineering at Tianjin University. During her Master's study, she learnt German and worked as a research assistant at Ruhr West Applied University (HRW) in Mülheim an der Ruhr, Germany. In 2017, she obtained her Master's degree and continued working as a research assistant. Later on, she moved to Nijmegen, The Netherlands, and she began her PhD as an early-stage researcher in 2018. Her PhD research was funded by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement project TAPAS. She conducted her research at the Centre for Language Studies of the Radboud University in Nijmegen, where she was a member of the Graduate School of Humanities.

PhD related

Xue, W., van Hout, R., Cucchiarini, C., & Strik, H. (2023). Assessing speech intelligibility of pathological speech in sentences and word lists: the contribution of phoneme-level measures. *Journal of Communication Disorders*, *102*, Article 106301. https://doi.org/10.1016/j.jcomdis.2023.106301

Xue, W., van Hout, R., Cucchiarini, C., & Strik, H. (2021b). Assessing speech intelligibility of pathological speech: test types, ratings and transcription measures. *Clinical Linguistics & Phonetics*. Advance online publication. https://doi.org/10.1080/02699206.2021.2009918

Xue, W., van Hout, R., Boogmans, F., Ganzeboom, M., Cucchiarini, C., & Strik, H. (2021a). Speech intelligibility of dysarthric speech: human scores and acoustic-phonetic features. In *Proceedings of Interspeech 2021*, 2911-2915. https://doi.org/10.21437/Interspeech.2021-1189

Xue, W., Mendoza Ramos, V., Harmsen, W., Cucchiarini, C., van Hout, R. & Strik, H. (2020). Towards a comprehensive assessment of speech intelligibility for pathological speech. In *Proceedings of Interspeech 2020*, *3146-3150*. https://doi.org/10.21437/Interspeech.2020-2693

Xue, W. (2020). Developing valid measurement procedures for speech intelligibility of pathological speech [abstract]. In *ISCA-SAC 6th Doctoral Consortium Abstract book*. 2020 Oct 24. Abstract 10.

Xue, W., Cucchiarini, C., van Hout, R. & Strik, H. (2019). Acoustic correlates of speech intelligibility: the usability of the eGeMAPS feature set for atypical speech. In *Proceedings of SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education*, 48-52. https://doi.org/10.21437/SLaTE.2019-9

Xue, W., Cucchiarini, C., van Hout, R., & Strik, H. (2019, August 27-30). *Effects of acoustic characteristics on dysarthric speech intelligibility*. [Poster presentation]. International Symposium on Monolingual and Bilingual Speech 2019. Chania, Greece. https://www.researchgate.net/publication/337831796_Effects_of_acoustic_characteristics_on_dysarthric_speech_intelligibility

Xue, W., Cucchiarini, C., van Hout, R., & Strik, H. (2021). Measuring Speech Intelligibility of Dysarthric Speech through Automatic Speech Recognition in a Pluricentric Language. Submitted to *Speech Communication*.

Others

Xue, W., Xiao, M., Sun, G., & Xu, F. (2016). A Compact Low-Profile and Quad-Band Antenna with Three Different Shaped Slots. *Progress In Electromagnetics Research C, 70*, pp. 43-51, 2016. https://doi.org/10.2528/PIERC16102704