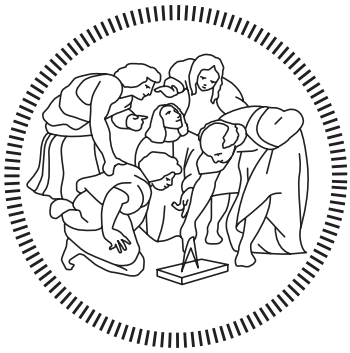
POLITECNICO DI MILANO

School of Industrial and Information Engineering

Master of Science in Biomedical Engineering



Design, implementation, and pilot testing of a language-independent speech intelligibility test

Supervisor :

Prof. Riccardo Barbieri

Co-Supervisors:

Prof. Toon van Waterschoot

Dr. Alessia Paglialonga

Master thesis by:

Giulia Rocco

Matr. 858902

Academic year 2016-2017

Alla mia famiglia e ad Antoine

Abstract

Over 5% of the world's population, i.e. 466 million people, is affected by disabling hearing loss. This number is expected to increase by 2050 to 1 out of 10 people, hence it is fundamental to promote hearing screening programs and useful tools for quick and reliable early hearing-impairment detection. Speech-in-noise tests are the most adequate indicators of the effective ability to understand voice in realistic listening conditions. Nevertheless, they are usually addressed to specific language speakers and not suitable for uncontrolled environmental conditions, so limited in use. This work aims at designing a multiple choice language-independent speech-in-noise test for domestic usage. The characterization of the new dataset with vowel-consonant-vowel (VCV) words is based on a comparative inter-language analysis of intelligibility prediction measures (STOI) and subjective listening scores from previous tests. The test implementation includes a calibration procedure, in which the subject self-adjusts the volume in response to a given stimulus, and an environmental monitoring strategy, by periodically recording the external noise and adapting the test noise to maintain a given SNR. English was chosen as trade-off between small variability and response dynamics. Feasibility and test-retest reliability were assessed by a pilot experiment with 11 normal hearing subjects, among whom 10 non-native English listeners. Subjects' speech reception thresholds, i.e. SNR corresponding to 79,4% of correct responses, were homogeneous and reliable, and no learning effect was found. The results are promising and suggest further investigations in this direction.

Keywords : speech-in-noise, STOI, VCV, psychoacoustics, hearing screening

Sommario

Oltre il 5% della popolazione mondiale, ovvero 466 milioni di persone, e' affetta da perdita uditiva invalidante. Questa cifra e' destinata a crescere nel 2050 fino ad una persona su dieci, per cui la promozione di programmi di screening uditivo e strumenti utili ad effettuare precocemente diagnosi veloci ed affidabili acquista un ruolo fondamentale. I test di "speech-in-noise" (letteralmente "parlato nel rumore") risultano gli indicatori piu' adeguati della effettiva abilita' di comprensione del discorso in condizioni di ascolto realistico. Tuttavia, tali test sono di solito indirizzati a soggetti che parlano una specifica lingua e non sono adatti ad essere testati in ambienti non controllati, per cui limitati nell'uso. Questo lavoro e' finalizzato allo sviluppo di un test di speech in noise a scelta multipla indipendente dalla lingua, adatto ad essere usato in ambiente domestico e in generale non controllato. La caratterizzazione del materiale del nuovo dataset con parole costituite da vocale-consonante-vocale (VCV) e' basata su un'analisi comparativa tra diverse lingue riguardo a misure oggettive di intelligibilita' (STOI) e risultati da precedenti test soggettivi. L'implementazione del test include una procedura di calibrazione, in cui il soggetto auto-regola il volume in risposta ad un dato stimolo, ed una strategia di monitoraggio ambientale, tramite la registrazione periodica del rumore esterno col fine di adattare il rumore da usare nel test e mantenere un dato SNR. L'inglese e' stato scelto come lingua da utilizzare nel test, poiché rappresenta il trade-off tratra bassa variabilita' e ampia dinamica dell'intelligibilita'. Fattibilita' e accuratezza sono state valutate tramite uno studio pilota con 11 soggetti senza disabilita' uditiva, di cui 10 aventi madrelingua differente dall'Inglese. Le soglie di percezione (SRT, speech reception thresholds) ottenute dai soggetti, ovvero il valore di SNR corrispondente al 79,4% di risposte corrette, sono omogenee e ripetibili, e nessun effetto di learning e' stato osservato. I risultati sono promettenti e suggeriscono ulteriori incoraggiano ulteriori approfondimenti e sviluppi in questa direzione.

Keywords : speech-in-noise, STOI, VCV, psicoacustica, screening uditivo

Acknowledgement

I would like to thank all the persons closely or remotely involved into this project.

Especially my supervisors: Prof. Riccardo Barbieri, who believed in this work from the beginning; Dr. Alessia Paglialonga, for her precious advices and extreme availability; and finally Prof. Toon van Waterschoot, for his heartwarming welcome in his team and for having supervised the project with enthusiasm.

A special thanks goes to Giuliano and Randi for giving me always fast and useful feedback, besides insights into the "research life".

A heartfelt thanks to all my friends, those who crossed my path during university, those who were my lovely housemates, those who have been supporting me for a lifetime.

To my family and Antoine: it is difficult to express how grateful I am to you for everything you did for me day-to-day. I just say that if I never felt alone and continued on my way, it's thanks to you.

Contents

List of Figures

List of Tables

List of Symbols

γ	Chance performance
λ	Lapsing rate
$\mathcal{N}$	Normal distribution
μ	Mean of the normal distribution
Φ	Real experimental psychometric curve
ρ	Pearson correlation coefficient
ρ_{spear}	Spearman correlation coefficient
σ	Standard deviation of the normal distribution
φ	Scaled psychometric curve
$P_{disturb}$	Power of the recorded environmental disturb
P_{noise}	Power of the test noise
P_{signal}	Power of the signal (VCV)

List of Abbreviations

1U3D 1-up/3-down.

2AFC Two-alternative forced-choice.

3AFC Three-alternative forced-choice.

ARHL Age-related hearing loss.

GUI Graphic user interface.

HI Hearing-impaired.

IPA International Phonetic Alphabet.

ISO International Organization for Standardization.

LTASS Long-term average speech spectrum.

n.d. Not defined.

NH Normal hearing.

NHT National hearing test.

NIHL Noise-induced hearing loss.

PTA Pure tone audiometry.

RMSE Root mean square of the prediction error.

SNHL Sensorineural hearing loss.

SNR Signal to Noise Ratio.

SRT Speech Reception Threshold.

SSN Speech shaped noise.

STOI Short-time objective intelligibility measure.

SUN Speech Understanding in Noise.

VCV Vowel-consonant-vowel.

Chapter 1

Introduction

1.1 Hearing loss in adults

Hearing loss is the complete or partial loss of the ability to hear from one or both ears [NIH - National Library of Medicine, 1993], when some form of hearing impairment affects the auditory system.

It is commonly identified by an elevated threshold of 25 dB from 0,5 to 4 kHz, and referred as disabling when the hearing loss is greater than 40 dB in the best hearing ear in adults, and 30 dB in the best hearing ear in children [WHO, 2018].

Over 5% of the world's population, i.e. 466 million people, has disabling hearing loss (432 million adults and 34 million children). The majority of people with disabling hearing loss live in low and middle income countries. Due to the population aging, it is estimated that by 2050 over 900 million people, which means one out of every ten people, will have disabling hearing loss [WHO, 2018].

A large variety of problems can impact the auditory system and give rise to hearing loss. Therefore, a classification is needed to discriminate between the different types of hearing loss. The standard classification [Moore et al., 2010] differentiates between conductive, sensorineural, and mixed loss. Problems in the outer ear or the middle ear cause conductive hearing loss; impairment of the inner ear or higher levels of the auditory pathway is referred to as sensorineural hearing loss (SNHL); mixed hearing loss is a combination of the two.

Hearing loss can arise at birth or soon after because of congenital causes such as genetic factors or certain complications during pregnancy and childbirth, or at any age by acquired causes. Among those, particular relevance is attributed to aging, excessive noise (including occupational noise such as that from machinery and explosion), recreational exposure to loud

sounds (such as that from use of personal audio devices at high volumes and for prolonged periods of time and regular attendance at concerts, nightclubs, bars and sporting events).

Both age-related (ARHL) and noise-induced hearing loss (NIHL) are identified by damage of cochlear hair cells. In particular, the first hair cells that usually undergo the damaging process are the ones on the base of the basilar membrane, which are responsible for the high frequency sounds transmission. This is the reason why these two forms of SNHL are commonly associated to high-frequency hearing loss. The presence of such hearing loss results in complications of one's communication abilities with others, giving rise to feelings of loneliness, isolation, and frustration [Nachtegaal et al., 2009, Arlinger, 2003].

Because of the gradual development of hearing loss, people with mild high-frequency hearing loss are often unaware of their impairment until the disability reaches a certain degree [Vogel et al., 2007]. It has been reported [Davis et al., 2007] that the majority of hearing aids or prostheses users live along with their hearing disability for more than ten years, before seeking for help, causing also the disease progress.

However, despite the large incidence of these diseases and their severe consequences on quality of life, these problems are still under-diagnosed. Since hearing damage is irreversible, it is essential to recognize it as soon as possible, so as to undertake precautionary measures to prevent more impairing, permanent, hearing damage [Meyer-Bisch, 1996].

In this context, it is extremely important to raise awareness about early detection of hearing impairments in adults and to promote initiatives and hearing screening programs [Grandori et al., 2009, Liu et al., 2011], providing also suitable and affordable instruments.

1.2 Speech in noise tests

One of the most common complaints of hearing-impaired adults, even at early stage, is the difficulty in understanding speech in noise [Humes, 2013], since such challenging situations test the ability to recognize fast speech transients and consonants [Laplante-Lévesque et al., 2011].

Although the pure-tone audiometry (PTA) is the most diffused hearing screening test, this is not actually suitable to accurately predict the extent to which one is able to recognize speech in noise. Therefore, it does not constitute a direct measure of the real hearing disability [American Speech Language Hearing Association and others, 1997]. It has been largely demonstrated that same audiograms for different adults correspond to a really wide variability of performance in voice recognition.

Therefore, hearing tests based on speech-in-noise listening and recognition are the most adequate indicators of the effective ability of understanding voice in realistic listening conditions [Killion and Niquette, 2000].

In this scenario, performing a speech-in-noise hearing test can be extremely useful in the early detection and prevention of hearing loss. Moreover, an easily administered hearing screening test can both raise awareness of possible hearing problems and stimulates people to seek audiological help by giving feedback of individual hearing status [Smits et al., 2006].

In the recent years, speech-in-noise tests largely diffused in clinical practice [Strom, 2003]. Several relevant attempts of computer-based speech-in-noise intelligibility tests were made and will be listed below.

The Quick SIN (speech-in-noise) test [Killion et al., 2004] is a faster and more accurate version of the original SIN test developed by Etymotic Research [Etymotic Research, 2018].

It consists of a series of IEEE sentences presented in a background of four-talker babble, with fixed level of the sentence and varying level of the noise. Since the noise levels vary automatically, it is suitable for clinical use. Keywords recognition is scored in each sentence and one point is given for each keyword repeated correctly. The hearing loss in SNR, intended as the increase in signal-to-noise ratio required by a listener to obtain 50% correct words, can be determined in both ears in about two minutes.

A widespread speech-in-noise self-test for the Dutch language is the National Hearing Test (NHT) [Smits et al., 2004]. It consists in the presentation of digit triplets in noise, thus, it is easily administered by telephone. Afterwards, an internet version of this test was created [Smits et al., 2006]. According to the SRT resulting from the test, a recommendation for follow-up might appear.

Starting from NHT, an internet-based speech-in-noise test aimed at evaluating in particular higher frequencies recognition, was developed, Earcheck [Albrecht et al., 2005].

An additional improvement was made to be specifically applicable in commercial enterprises and monitor the hearing ability of employees in noisy occupations, Occupational Earcheck [Ellis et al., 2006]. It has a similar procedure, but it was designed to have better precision by increasing the number of stimuli and by consecutive monaural testing of both ears.

In general, many other speech-in-noise tests were implemented and validated for specific languages [Paglialonga et al., 2014, Ozimek et al., 2009, Nielsen and Dau, 2009].

The results have been promising, but there are still some limitations which prevent the

use of these tests in large scale and in uncontrolled environment.

All the tests datasets consist of short sentences or words, based on the language phonetic and/or lexical characteristics; thus, they are fully language-dependent and cannot be adapted to other languages, unless a new dataset definition, characterization and recording procedure is done.

The use of meaningless words, such as vowel-consonant-vowel stimuli, is independent on language lexical characteristics and meaning understanding, thus on subjects' education and literacy [Cooke et al., 2010]. In addition, tests could be easily adapted to different languages [Paglialonga et al., 2014].

Moreover, tests are usually conducted in controlled laboratory environments, in which no disturbances could affect the subject's performance and bias the results. Even though some of the existing tests were already thought to be executed in non clinical settings [Smits et al., 2006, Paglialonga et al., 2014], the development of strategies allowing home-based testing needs to be further investigated and validated. This would largely increase the test usability, and thus the subjects undergoing hearing screening.

1.3 Research outline

This work is aimed at overcoming one of the two major constraints of the previous tests, i.e. the language dependence, and at further exploring the effect of an uncontrolled testing environment, or at least initiating the process towards this, by implementing a new language-independent speech-in-noise test suitable for domestic testing and verifying its feasibility with a pilot experiment.

This decision is supported by the study of Cooke et al. [Cooke et al., 2010], in which they observed a large degree of language independence in speech perception. Their findings suggest that many aspects of intervocalic consonant identification are largely independent of the native language of the talker. In particular, strong between-language similarity was seen in the ranking of the proportion of information transmitted for phonetic features, and the way this value changes in different noise conditions.

Furthermore, earlier studies on speech-in-noise testing in a living room environment, using either headphones [Ozimek et al., 2009] or loudspeakers [Culling et al., 2005], yielded promising results, which were highly comparable to those obtained in laboratory conditions. However, these testing environments were simulated and not real, hence the need of further

investigations and environment monitoring strategies.

In Chapter 2, all the test design criteria are discussed and motivated. In particular, the characterization of the new dataset is defined, based on a comparative inter-language analysis of intelligibility prediction measures and subjective listening scores. A calibration procedure, a simple monitoring strategy for eventual external noises, and a pilot experiment are described.

In Chapter 3, the results concerning the intelligibility measures comparison are shown together with other figures of merit and statistical tests. The outcome of the pilot study is also analyzed through quantitative measures.

In Chapter 4, the choice of the dataset based on the previous results and the pilot results are illustrated and commented.

In Chapter 5, findings from this work are summarized and suggestions for further research are drawn.

Chapter 2

Material and methods

In this chapter all the choices concerning the dataset, the test design and implementation are discussed and motivated. In particular, the intelligibility comparison between the real data, which come from subjective listening tests in previous studies, and the predicted intelligibility values in the current work is investigated across the different languages and the dataset items.

The outcome of this analysis leads to the choice of a particular language for the new test with the purpose of being as much language-independent as possible. In addition, it provided deep insights into the benefits of intelligibility prediction besides differences and similarities in the recognition task among the languages taken into account.

After the test was implemented, a pilot experiment was conducted in order to evaluate test feasibility, reliability and time and to propose any kind of optimization or modifications for the test improvement from any perspective.

2.1 Task and dataset

According to the purpose of speech in noise test, the subject is requested to hear a given signal, which is corrupted by a specific noise, and to recognize it. The SNR and the presented signal can vary during the test execution depending on the test design criteria.

Several ways to communicate the occurred recognition from the subject have been used in the literature. In particular, automatization attempts both in typed and spoken response have been made. Typed word and sentence correction algorithms have been proved to be robust against spelling errors and large variety of mistypes and to perform even better than human operator [Francart et al., 2009]; an automatic speech recognizer developed for SRT measurements has been evaluated in comparison with the performance of the same task by

an audiologist and in view of the positive results and high accuracy it has been proposed as a viable tool for these kinds of measurements [Deprez et al., 2013].

Despite the positive feedback of different automatic approaches, the use of multiple-choice task is adopted in this work for several reasons. First of all, multiple-choice tasks are easy to implement in a user-operated and automated procedure, which could be, as in this case, a simple graphic user interface in which the subject selects the desired response [Paglialonga et al., 2014, Albrecht et al., 2005]. Moreover, the simple interaction that is required - i.e. clicking on the screen - contributes to reducing significantly the unease and anxiety in the subjects during the test and the negative influence of cognitive decline, slowed temporal processing and added memory demands on speech recognition, especially for older adults [Gordon-Salant, 2005].

The adopted multiple-choice task is the three-alternative forced-choice (3AFC), in which the subject is obliged to select an answer among three mutually exclusive alternatives for each question. In psychophysical tasks, this method performs better than the popular and widespread two-alternative forced-choice (2AFC) in terms of threshold estimation stability under different types of adaptive procedures [Leek, 2001]. [Kollmeier et al., 1988] reported that until proven otherwise the 3AFC in combination with other test settings is the most efficient method. These settings will be further discussed in Section 2.7.

The chosen stimuli to use in the test recognition task are meaningless vowel-consonant-vowel (VCV) words. The use of VCVs in clinical tests has been largely diffusing for years, especially in hearing disease-related applications and early diagnosis [Dorman et al., 1990, Paglialonga et al., 2014, Loizou et al., 2000, Baer et al., 2002]. This is supported by a considerable number of advantages:

- As they are meaningless words, they are independent on the lexicon and semantics, and limit the influence of cognitive-auditory interactions and the involvement of higher level processing centers linked to them, which are known to contribute to phoneme perception [Nittrouer and Boothroyd, 1990];
- They are suitable for hearing screening, given the fact that adults experience hearing difficulties at early stage with high frequencies and consonants recognition in noise [Laplante-Lévesque et al., 2011];
- Intervocalic consonant identification revealed to be largely independent on the subjects' education, literacy and native language [Cooke et al., 2010], further supporting the use

of a language-independent approach in this study;

– VCVs have good test-retest reliability and are appropriate for use in investigations which are based on administration of identical items under multiple experimental conditions [Dubno and Dirks, 1982].

In the current study, the choice of the VCVs alternatives to present at each question of the 3AFC task is determined following a maximal opposition criterion [Gierut, 1989]: the two wrong options differ from the effectively reproduced audio sample in manner, voicing and place of articulation, referring to the International Phonetic Alphabet (IPA) [Association, 1999].

As in the majority of the tests, the consonant stimuli are presented in the /a/ vowel context throughout the test; this vowel has been demonstrated to provide at times better intelligibility for the consonants [Donaldson and Kreft, 2006].

2.2 Comparative language analysis

All speech intelligibility tests up to now developed have been designed specifically for one language. Therefore, the first step towards a language-independent test is the selection of an appropriate set of stimuli spoken in a determined language, which could bring advantages in terms of intelligibility scores and not deteriorate performance of non native listeners.

In this regard we found it useful to analyze corpora of VCVs words recorded for the implementation of a speech intelligibility test in six different languages - the Speech Understanding in Noise (SUN) [Paglialonga et al., 2014, Vaez et al., 2014]. The test procedure has been designed for Italian and then, adapted to the other languages; the dataset, instead, has been tailored to the specific language with different recordings of several phonemes relevant for the language. Thus, differences among corpora were both in number and type of phonemes.

Despite that, VCV utterances were recorded with the same identical procedure for all the languages, i.e. as single exemplars in a sound-treated room by a professional language native male speaker who was instructed to pronounce the VCVs with no prosodic accent, with the stress on the first vowel and with constant pitch across the list. Stimuli were recorded in a professional recording studio using a Neumann TLM 103 microphone, a SSL S4000 64-channels mixer, Motu HD 192 A/D converters (44,1 kHz, 16 bit), and GENELEC 1025A control room monitor. The level of VCV recordings was digitally equalized across the set to meet the equal speech level requirement as set in the ISO Standard for Speech Audiometry [ISO 8253-3:1996(E), 1996].

The analysis of the recorded set languages, i.e. English, French, German, Italian, Portuguese and Spanish, is relevant since they are in the top 20 among the languages with at least 50 million first-language speakers [Simons and Fennig, 2017]. Languages which do not make use of Latin alphabet were not taken into account for the ambiguities that could have arisen in the phoneme-grapheme correspondence.

The recorded VCVs were reviewed in order to find a subset of phonemes which are common to all the languages and which show a large variety of speech features (voice, manner and place). As a result, 12 stimuli common to all the languages were selected (Table 2.1).

Stimuli					
aba	ada	afa	aga	aka	ala
ama	ana	apa	ara	asa	ata

TABLE 2.1. Chosen VCVs for the new test

Even though the intervocalic consonant recognition is largely independent of the subjects first language, there is an intrinsic linguistic perceptual factor which does not allow subjects to determine whether a language is more intelligible than another one. Hence the idea to use an objective intelligibility measure. Since subjective testing is laborious, time-consuming, and expensive, the diffusion and the investigation of automated, repeatable, fast, and cost-effective objective intelligibility monitoring tools is gaining ground both for normal hearing (NH) and hearing-impaired (HI) subjects who make use of assistive listening device.

The short-time objective intelligibility measure (STOI) is used. It is an intrusive measure based on a correlation coefficient between the temporal envelopes of the time-aligned reference (in our case, the VCV) and processed speech signal (VCV plus noise) in short-time overlapped segments of 386 ms [Taal et al., 2011]. This results in better correlation with speech intelligibility in listening tests compared to five other reference objective intelligibility models, which are all promising candidates for intelligibility prediction of noisy speech: Dau Auditory Model [Dau et al., 1996], Coherence Speech-Intelligibility Index [Kates and Arehart, 2005], Normalized Covariance Based Speech Transmission Index [Goldsworthy and Greenberg, 2004], Frequency -Weighted Segmental SNR [Hu and Loizou, 2008], and Normalized Subband Envelope Correlation [Boldt and Ellis, 2009]. since they rely on the global statistics over longer segments.

2.3 Noise processing

In order to perform STOI computation, choices about the noise specifications for the test were defined. In a large number of speech in noise tests, the noise is the result of different filtering techniques applied to steady-state unmodulated white noise. For instance, in SUN test a FIR filter with magnitude response equal to the long term average speech spectrum (LTASS) for the specific language of the test was used [Paglialonga et al., 2014, Byrne et al., 1994]; in other works [Cooke et al., 2010, Leensen et al., 2011] the LTASS of the concatenation of all test words was used. Both choices do not fit in our evaluation: the use of a specific language spectrum to evaluate all the languages intelligibility would be in contrast with a language-independent approach and would bias the result; equally, it would be not coherent to choose a priori some words to concatenate before knowing they would be effectively part of the dataset of the final test.

Byrne et al.[Byrne et al., 1994] conducted a study about comparison among LTASS of 12 languages, including several non-European ones. They found out that LTASS was similar for all languages despite some statistically significant differences. However, such differences were small and not always consistent for male and female samples of the same language. Therefore, a "universal" LTASS is suggested as being applicable, across languages, for many purposes including use in hearing aid prescription procedures and Articulation Index, which is a good intelligibility predictor diffused before STOI [Kryter, 1962]. In view of this, the international LTASS is chosen for our purpose and shown in Figure 2.1.

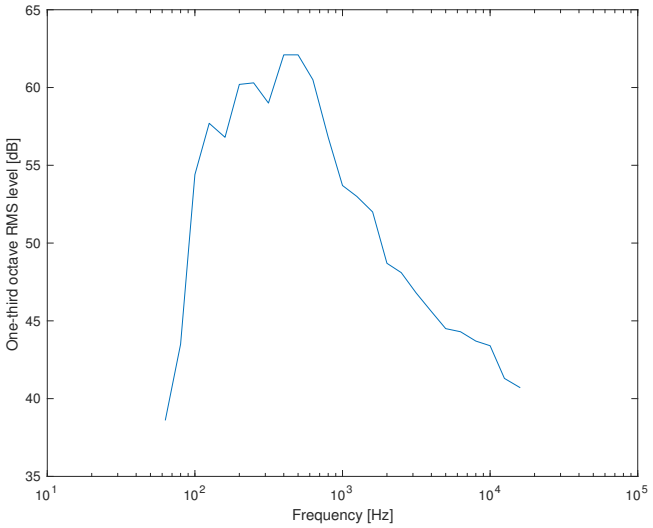


FIGURE 2.1. International LTASS proposed by [Byrne et al., 1994]

The complete noise processing chain is represented in Figure 2.2. The result of the FIR filtering, based on LTASS, is a speech shaped noise (SSN), which is subsequently scaled in order to match the level of the speech material. At this point, further noise processing techniques have been suggested, analysed and compared to each other in literature. The use of a low pass filter with a cut-off frequency of 1,4 kHz and with a steep roll-off slope (100 dB/octave) plus the addition of a noise floor after filtering, which consisted of the speech-shaped noise attenuated by 15 dB, has been chosen [Leensen et al., 2011, Leensen and Dreschler, 2013]. The low-pass noise is able to most finely differentiate than other processed noise (e.g. high-pass, modulated, etc.) between the normal hearing and the mild hearing-impaired subjects; the noise floor is added in order to mask potential ambient noise levels.

Noise has not the same duration as the VCV, but it sets at specific times before and after the spoken word. Noise onset time has been proved to have an impact on the test performance: in particular, carrying out evaluations with onset time multiple of 57 ms, top performances for all the languages have been obtained at 228 ms [Cooke et al., 2010]. Then, this is the choice for the onset time, while 500 ms is the one for the offset [Versfeld et al., 2000].

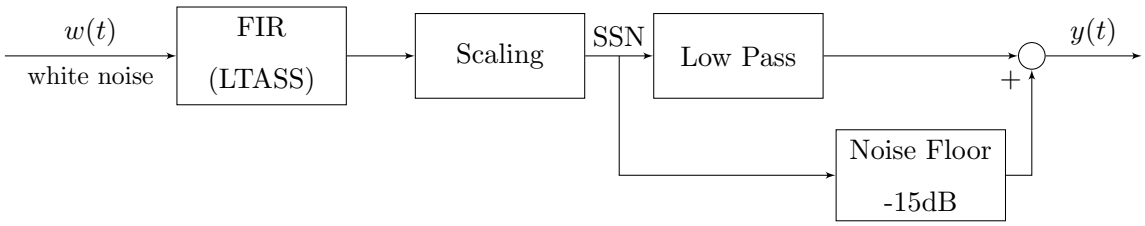


FIGURE 2.2. Noise processing

2.4 STOI variability sources

STOI calculation is made as a function of the SNR in a range from -12 to +6 dB, which is shown to cover the entire dynamics for nearly all the stimuli in the experimental data, and thus will be with great probability part of the examined range in the test for all the subjects.

Some preliminary calculations showed that the intrinsic stochasticity of the noise, due to its white nature, could modify the STOI value for the same signal. In order to evaluate the entity of this variability, STOI calculation simulations were run with noises generated by processing 100 different Gaussian noise realizations. The calculation were done for all stimuli of all languages, for a total of 72 times (12 VCVs x 6 languages).

The estimation of VCV intelligibility through these simulations will lead to the identification of the preferred language for the test. The obtained values are analyzed by averaging and evaluating the STOI variance both across languages and across VCVs.

For each VCV stimulus of the test, the average across languages is meant to show differences in intelligibility among the consonants, as we should expect, given the different speech features and frequency which characterize all of them.

For each language, instead, the average across VCVs is crucial to determine the suitability for the test and an eventual partition of the VCVs in homogeneous classes for intelligibility. While using adaptive procedures to assess speech intelligibility, it is extremely important that the speech stimuli have equal difficulty in noise in order to produce accurate and consistent results. Thus, stimuli level should be equalized to ensure perceptual homogeneity. This process is usually done by word-specific intelligibility functions drawn by pre-tests on purpose performed (e.g. [Leensen et al., 2011]). Since this is a time-consuming process, we tried to extract this information from the STOI curves graphical grouping and thresholds (see Section 2.7).

Statistical analysis of STOI mean and variance across VCVs is done. The evaluation criteria are summarized in Table 2.2. Regarding the STOI mean, it would be appropriate to have high values not to take the risk to increase the task difficulty for non-native language listeners, but noticeable variability across the dataset, which would denote differences among the VCVs, allowing the level equalization. Regarding the STOI variance, every statistics should result in small values so as to show robustness to noise stochasticity.

Parameter	Statistics	Ideal value	Why
STOI mean	Mean	High	High intelligibility with noise, not to increase difficulty of the task
	Variance	Noticeable	Groups of VCVs classes with different SNR for the test
STOI variance	Mean	Low	Robust to intrinsic stochasticity of the noise
	Variance		

Table 2.2. Ideal evaluation criteria of statistics across VCVs

In order to ensure good dynamic features, the range Δy is measured as the difference between the highest and lowest STOI mean value. The dynamic range is an important measure of the sensitivity of the signal to the noise; then, it should be adequately large to prove noticeable changes in performance.

2.5 Objective vs subjective intelligibility measures

Besides the dataset development, language-specific listening tests were performed for all the analyzed languages, except Spanish for practical issues [Paglialonga et al., 2014, Vaez et al., 2014]. In order to assess the accuracy of objective intelligibility prediction, STOI values are compared by different means to the results collected by using the considered VCV datasets in this study.

To evaluate the performance of the objective measures, several performance criteria were used, taking into account the insights provided by previous studies [Falk et al., 2015].

First of all, the linear relationships between the predicted intelligibility values and the subjective scores are measured through Pearson correlation (ρ). Subsequently, the objective metrics are assessed in terms of ranking capabilities by the Spearman rank correlation, (ρ_{spear}): the computation is very similar to ρ, but with the original data values replaced by their ranks. The combination of these two measures highlights the necessity of a nonlinear monotonic mapping between the objective metric scale and the subjective rating scale.

Such mapping is done by using a logistic function [Taal et al., 2011, Xia et al., 2012]:

$$f(x) = \frac{1}{1 + e^{(ax+b)}} \tag{2.1}$$

where x is the objective intelligibility score, a and b are the parameters that are tuned with a non-linear least square procedure. After the logistic fitting, the performance of the objective metric is evaluated by calculating the RMS of the prediction error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{S}\sum_{i}(s_i - f(x_i))^2} \tag{2.2}$$

in which s_i indicates the subjective intelligibility rating and S the total number of scores taken into account for the fitting. Low values of RMSE correspond to high ability of the objective measure to predict intelligibility scores in the considered language.

Besides the above-mentioned correlation-based measures, other two relevant parameters are extracted from both objective and subjective intelligibility curves: the slope and the Speech Reception Threshold (SRT). These two measures are particularly instructive about the dynamic feature of the response and require a specific fitting of the experimental data. This procedure will be explained in details in the next section (Section 2.6).

2.6 Data fitting for slope and SRT extraction

As in our case, the interest in fitting and analyzing the psychometric function underlying a perceptual process, i.e. the mathematical relationship between the probability of correct responding $p_c(x)$ to the physical variable x under study, is typically finalized to the extraction of slope and threshold values. The rationale behind this calculation lies in the fact that the steepness of the function represents a measure of reliability of the sensory performance and for this reason could verify a valuable diagnostic power of psychophysical tasks [Chauhan et al., 1993]; the empirical comparison among different thresholds could outline as well subjects' classification upon the performance.

Based on common practice, a two-parameter function in the range (0,1) is fit to the data. Let's consider an adaptive procedure (Section 2.7) in which the stimulus level is the random variable $\mathbf{X}_n$ and the subject's response $\mathbf{Z}_n$ is the random variable at trial n. The value of $\mathbf{Z}_n$ will be 0 if the answer is incorrect and 1 if it is correct. According to the definition of the psychometric function, we have:

$$\mathrm{P}[\mathbf{Z}_n = 1|\mathbf{X}_n] = \Phi(\mathbf{X}_n) \tag{2.3}$$

$$\mathrm{P}[\mathbf{Z}_n = 0|\mathbf{X}_n] = 1 - \Phi(\mathbf{X}_n) \tag{2.4}$$

This means that the the score of correct identified stimuli at any level is binomially distributed. If extreme values approaching the asymptotes are avoided, the distribution is with good approximation a Gaussian [Lyregaard, 1997]. Hence, the psychometric function can be modeled by a cumulative normal function:

$$\mathcal{N}_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{(t-\mu)^2}{2\sigma^2}} dt \tag{2.5}$$

with mean μ and standard deviation σ defining SRT (displacement along x-axis) and slope (by inverse relationship).

In order to estimate μ and σ, the cumulative function range (0,1) should be modified and adapted according to real experimental performance, taking into account a lower and an upper bound which are respectively described by two parameters, γ and λ. This model of the curve Φ is shown in Figure 2.3.

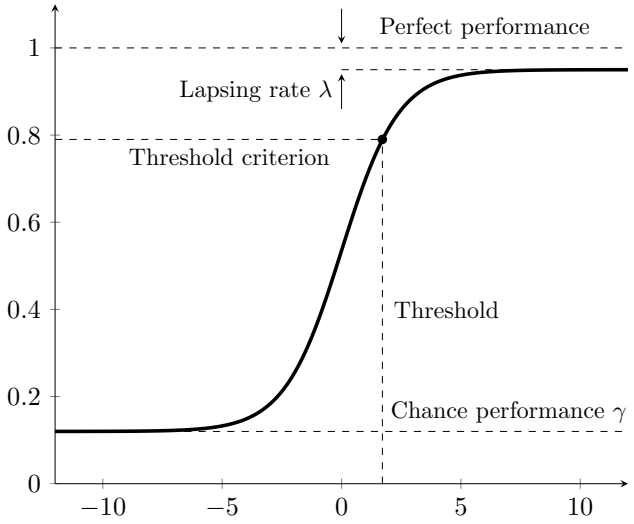


FIGURE 2.3. Model of the psychometric function for detection task

The parameter γ identifies the lower bound of the curve and can be interpreted as the base rate of performance in absence of signal (chance performance). In a n-AFC task it is usually fixed at the reciprocal of the number of alternatives per trial, e.g. 0,33 for 3AFC. The parameter λ defines the upper bound of the curve (1 - λ) and it is the rate at which the subject lapses, i.e. responding incorrectly regardless of stimulus intensity; it is theoretically equal to 0. Despite the theoretical assumptions about γ and λ values, we chose to attribute them the actual values corresponding to the minimum and the maximum points of the experimental curve Φ, without inferring anything a priori about the subject's performance. Nevertheless, the values themselves are of secondary scientific interest, since they arise from the stimulus-independent mechanisms of guessing and lapsing [Wichmann and Hill, 2001].

In view of the above, the psychometric curve can be expressed as follows:

$$\Phi(x) = \gamma + (1 - \lambda - \gamma)\varphi(x) \tag{2.6}$$

where $\Phi(x)$ is the real experimental psychometric curve, ranging from γ to 1 - λ, while φ is

the psychometric curve scaled between 0 and 1.

From Equation 2.6, the inverse expression can be derived:

$$\varphi(x) = \frac{\Phi(x) - \gamma}{1 - \lambda - \gamma} \tag{2.7}$$

Since $\varphi(x)$ lies within 0 and 1, it is the function whose parameters will be estimated.

Thus, the fitting procedure can be summarized in three steps:

1. scaling of the real psychometric function $\Phi(x)$ (with range from γ to 1 - λ) to the range from 0 to 1, resulting in $\varphi(x)$ (see Equation 2.6);

2. fitting the function $\varphi(x)$ with a cumulative normal model, obtaining a fitting function $\hat{\varphi}(x)$;

3. scaling back $\hat{\varphi}(x)$ within γ and 1 - λ, obtaining $\hat{\Phi}(x)$, which is the fitting function approximating $\Phi(x)$ (see Equation 2.7).

Step 2. is done by using one of the classical procedure employed to analyze data gathered in a psychophysical experiment: transforming the data from the range (0,1) to the range $(-\infty, +\infty)$. Such a transformation is finalized to the adoption of a linear model for the parameters estimation. Commonly used transformations are the logistic, the probit and the complementary log-log transformation, which correspond respectively to the logistic, the normal and the Gumbel distribution [Treutwein, 1995]. In our case, given the choice of the cumulative normal distribution for $\varphi(x)$, the probit analysis, which is also the best known in psychophysics, is performed [Finney and Tattersfield, 1952].

The probability data $\varphi(x)$, analytically expressed in Equation 2.8, are transformed into standardized z-values through the inverse on the cumulative normal function (Equation 2.9):

$$\varphi(x) = p = \mathcal{N}_{\mu,\sigma}(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{\frac{y^2}{2}} dy \tag{2.8}$$

$$z(p) = \mathcal{N}^{-1}_{\mu,\sigma}(p) = \mathcal{N}^{-1}_{\mu,\sigma}(\varphi(x)) \tag{2.9}$$

Once the z-values at the different stimulus levels are calculated, parameters c_0 and c_1 are estimated by means of linear regression:

$$z(x) = c_0 + c_1 x \tag{2.10}$$

By expliciting the standardization for $z(x)$ (Equation 2.11), the parameters μ and σ are derived (Equation 2.12):

$$c_0 + c_1 x = \frac{x - \mu}{\sigma} \tag{2.11}$$

$$\mu = -\frac{c_0}{c_1} \qquad \sigma = \frac{1}{c_1} \tag{2.12}$$

It is important to underline that when the stimuli are not equally distributed for every level, the regression has to be weighted: the weights depend on the number of trials at a given stimulus level and the probability of a correct answer. Neither the objective metric nor the subjective scores regressions need to be weighted, since the same amount of stimuli per level was taken into account in the simulations or presented in the listening tests.

As [Treutwein, 1995] reports, it should not be neglected that all transformations, including probit one, are undefined for certainties, i.e. probability 1 or 0, and it is not strictly possible to incorporate these data points in the linear regression. According to this statement, these points were not included in the regression in case of occurrence. In addition, in order to minimize bias in the slope estimation, subjective data were fitted by using a value of λ slightly higher than zero, i.e. 10^{-4} [Wichmann and Hill, 2001, Strasburger, 2001].

Once the fitting is done, slope and threshold can be calculated.
The slope (%/dB) is obtained by the following formula, suggested by [Strasburger, 2001], which returns the value of the maximum slope on the cumulative normal function, i.e. in the point of inflection:

$$\text{Slope} = \frac{1 - \gamma - \lambda}{\sigma\sqrt{2\pi}} \tag{2.13}$$

The threshold (dB) corresponds to the SNR value whose level of performance is equal to $79,4\%$, in accordance with the 3AFC test [Leek, 2001] and can be derived by the following expression:

$$\text{Thr } = \mathcal{N}_{\mu,\sigma}^{-1}\left(\frac{0,0794 - \gamma}{1 - \lambda - \gamma}\right) \tag{2.14}$$

Values of slopes and thresholds from objective and subjective results were, then, compared by parametric/nonparametric paired tests, after the data distribution is verified.

2.7 Adaptive procedure

Psychometric procedures are ways of testing the observer in order to gain information about the psychometric function. The main distinction is between *fixed-levels* and *adaptive* techniques.

Fixed-levels procedures deliver a set of stimuli whose values span the entire range of sensation, from imperceptible to considerably perceptible, and are predefined before the experiment. Each stimulus level is presented a n number of times in a randomized sequence and the response data are obtained by the calculating the proportion of correctly identified stimuli per level among the n presented [Wichmann and Hill, 2001, Treutwein, 1995]. The n number of stimuli is usually the same for the levels, so that the curve sampling is uniformly distributed in space and experimental time.

In contrast, adaptive procedures determine the signal level on each trial by the subject's responses on previous trials. In a formal way, the stimulus level presented at a trial n is considered as a stationary stochastic process [Treutwein, 1995]. This results in the optimization of the placement of trials along the stimulus levels, i.e. greater sampling in the mid-range of the curve, and decrease of the experiment duration. The rule which determines the stimuli and the levels to present aims at converging to a desired target performance on the curve $\Phi(x)$. The value of such target performance is dependent on the rule itself and the task [Leek, 2001]. Unlike fixed-levels methods, the threshold is the point corresponding to the target level performance and it does not require the fully psychometric function sampling. Moreover, adaptive procedures can monitor the effect of learning and lapses [Leek, 2001]. For the mentioned advantages, an adaptive procedure is chosen for the new test.

Among the different types of the existing adaptive procedures, a *staircase* is chosen [Paglialonga, 2009, Paglialonga et al., 2011]. It is one of the most commonly employed approaches [Levitt, 1971] for its simplicity: the tracking algorithm does not rely on any assumption about the underlying model and the convergence is lead to the target performance by non parametric statistical models.

Staircase procedures are characterized by a set of variables, which can influence the algorithm convergence with more or less extent. The settings chosen for the new test are listed

and described in details in the followings:

- *Starting value*: the first stimulus level is 8 dB SNR, so as to allow perfect consonant recognition at the beginning of the test. The choice to present highly above-threshold values, thus easily recognizable, is made for the sake of subject's familiarization with the test [Green et al., 1989].

- *Tracking algorithm*: the transformed up-down method is chosen [Levitt, 1971]. In particular, the rule for governing the changes in the level of the signal is the 1-up/3-down (1U3D), i.e. the SNR is increased after any incorrect response and decreased after three correct responses in a row. The advantages of transformed updown methods are many, including simplicity, high efficiency, robustness, smallsample reliability, and relative freedom from restrictive assumptions. While the simple up-down procedure is designed primarily to place observations in the region of 50% correct response,the transformed up-down is well suited for estimating points other than 50%. Indeed, the convergence of 1U3D staircase occurs at 79,4% [Levitt, 1971], that is the SNR at which the probability to obtain a level decrease is equal to the probability to an increase. Moreover, it has been demonstrated by both mathematical models and listening tests that the combination of 1U3D with 3AFC tasks is the most efficient method [Kollmeier et al., 1988].

- *Step size*: it is constant and fixed at 2 dB for both upward and downwards runs. This choice was driven by the fact that in 1U3D staircase large step sizes might cause a downward shift of the target [García-Pérez, 1998]. In addition, it has been reported that bias in the slope estimation decreases with holding step size constant at 2 dB [Leek et al., 1992].

- *Stopping rule*: the staircase terminates after the SNR track has passed through 20 reversals, i.e. when a decrease in signal level is followed by an increase (lower turning point) or an increase in signal level is followed by a decrease (upper turning point). Staircases with less than 20 reversals are subjected to significant statistical bias and low precision of threshold estimates [García-Pérez, 1998].

- *Threshold estimation method*: the target point is estimated by the average of the SNR values at the midpoints of the ascending runs. The first three reversals are discarded in order to reduce estimation bias. Although this method appear simple, previous studies showed that is robust, efficient and accurate [Levitt, 1971].

2.8 Control for dichotic presentation

After all the design criteria were defined, the test was implemented taking into account environmental factors. As pointed out in Chapter 1, this new test is meant to be independent of external variables as much as possible. Thus, after a strategy to overcome the language specificity, the second step is to allow testing in uncontrolled and domestic settings and to grant much larger tool viability with respect to the numerous constraints of the previous tests, widening its potential user population considerably.

Earlier studies on speech-in-noise testing in a living room environment yielded SRT results that were highly similar to those obtained under headphones in laboratory conditions [Ozimek et al., 2009], even when loudspeakers were used [Culling et al., 2005]. However, other studies indicate that a different set of reference values is needed when speech-in-noise tests are performed using loudspeakers instead of headphones [Smits et al., 2006]. Although the use of loudspeaker would lose the potential to test each ear separately, it will noticeably increase the potential users due to their availability as standard PC equipment. In addition, the use of loudspeakers permits to assess hearing capacities of people with hearing aids. Consequently, in accordance with these considerations, the test will be performed by using loudspeakers. The subject is only recommended not to be too far from the loudspeaker in order to maximize the direct-to-reverberant ratio and not in a reverberant room, like a bathroom or kitchen [Culling et al., 2005].

In order to limit somehow the potential deterioration induced by non-ideal and unfavourable environmental conditions, a simple monitoring strategy between each trial and the following one is implemented. It consists of short recordings 0,5 seconds long of the environmental noise by the computer microphone. These recordings are realized starting from the moment in which the subject confirms a given answer for a trial presentation and the next stimulus presentation. Such an expedient is finalized at adjusting the level of the noise to superimpose to the next to be presented VCV, accounting for eventual disturbances, with the following calculation:

$$\mathrm{SNR} = 10 \log_{10} \frac{P_{signal}}{P_{noise} + P_{disturb}} \tag{2.15}$$

$$P_{noise} = \frac{P_{signal}}{10^{(0{,}1 \cdot \mathrm{SNR})}} - P_{disturb} \tag{2.16}$$

where P_{signal}, P_{noise}, $P_{disturb}$ are respectively the power of the signal (VCV), the test noise

and recorded environmental disturb, assuming that the last two sources are uncorrelated. Then, given the SNR value for a certain stimulus at a certain trial, the noise to add to the VCV would decrease depending on the environmental noise without inducing any change in the desired SNR presentation.

2.9 Graphic user interface

The test was implemented on Matlab (MATLAB Release 2017b, The MathWorks, Inc., Natick, Massachusetts, United States).

Here follows a screenshot of the graphic user interface (GUI) by which the test can be executed:

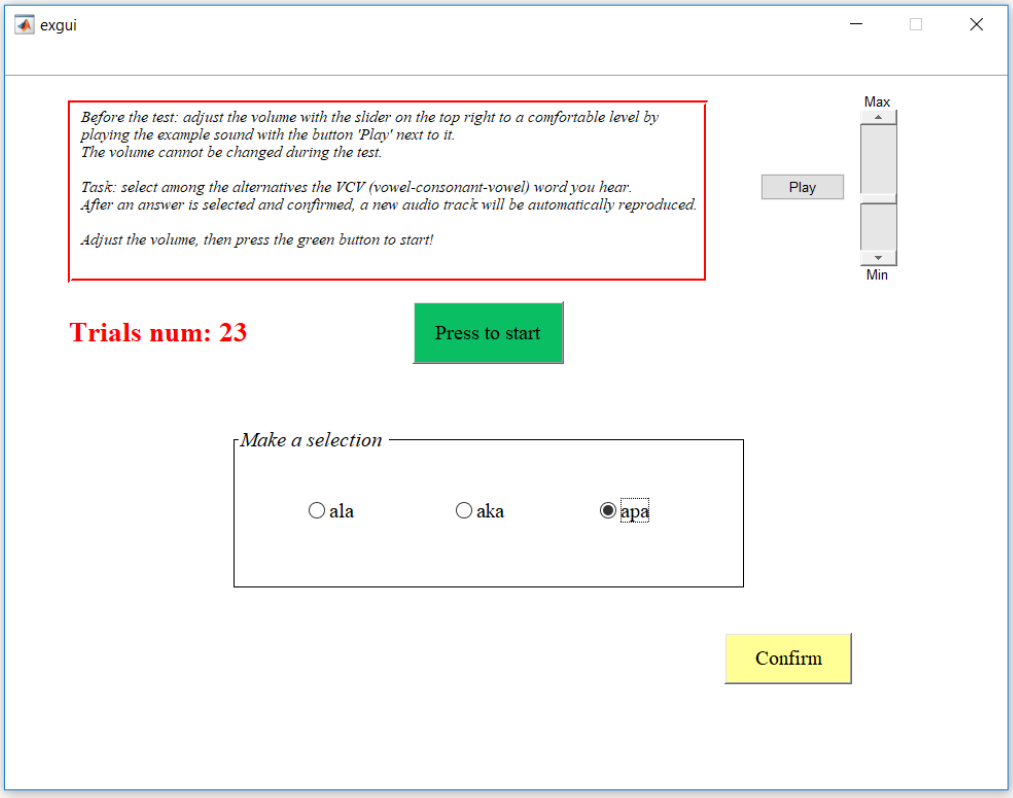


FIGURE 2.4. GUI during test

A small set of instructions is specified to the subject on the top left corner. First thing to do before starting the test is a simple calibration procedure: the subject is asked to adjust the volume at a comfortable level by using the slider on the top right corner and playing the example sound with the button close to it. The example sound is the word /asa/ without

the noise, since from the examined subjective results is the one with the highest intelligibility scores even at low SNR.

The test starts when the subject press the green button. Every time the audio sample is played, the subject has to select an answer among the three given alternatives and to confirm it, pressing the yellow button. After this is pressed, the new audio track will be automatically reproduced. Confirmation is asked so that the subject is willing to proceed the test and ready to perceive the next stimulus.

Each stimulus can be played only once, so no possibility of multiple hearing is given. This choice has been dictated by the fact that allowing to hear the stimulus twice or more times would have lead to bias in the SRT estimation (making it lower) and increased the test duration.

After the 20 reversals are reached, the test terminates with a message on the screen (Figure 2.5).

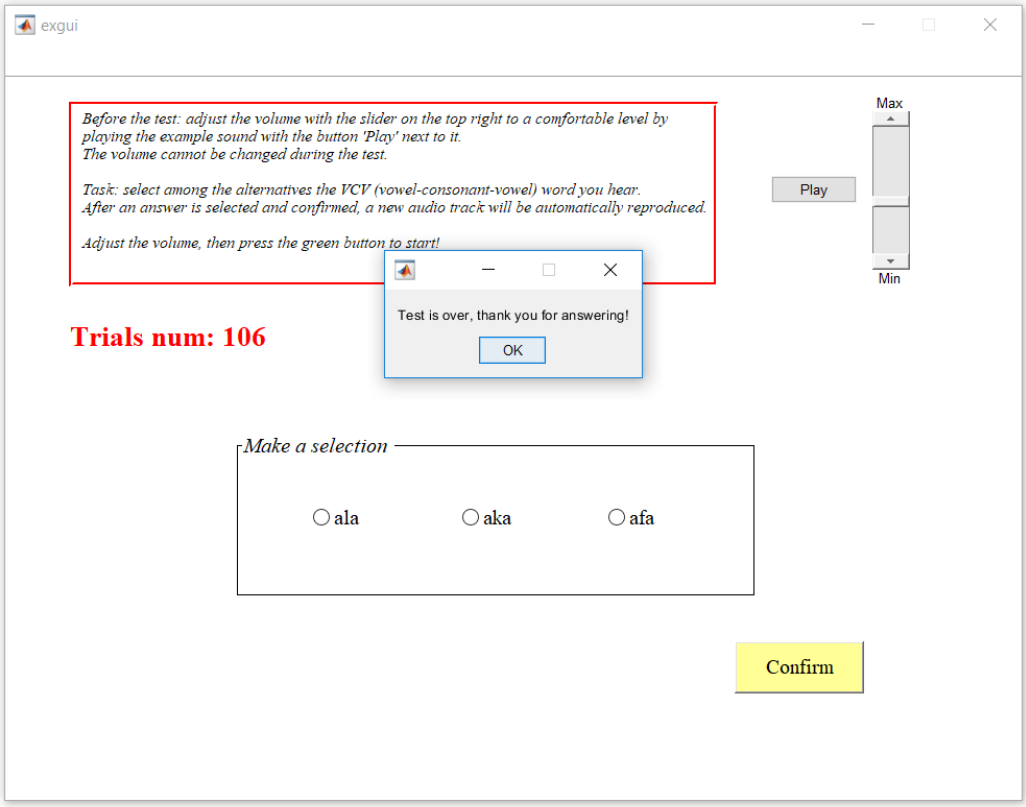


FIGURE 2.5. GUI with message box indicating the end of the test

2.10 Pilot experiment

A pilot experiment was conducted in order to evaluate feasibility, time, cost, and eventually improve the test weak points.

11 voluntary subjects (9 male and 2 female, between 26 and 30 years old) were tested twice in order to assess test-retest reliability and the presence of a learning effect. All the subjects were normal hearing, i.e. otologically normal according to the standard [ISO 7029:2017, 2017]. All participants gave written informed consent and ethical approval for the study was obtained from the Medical Ethics Committee UZ/KU Leuven.

All the subjects performed the test in working environments or offices with the same laptop, since Matlab installation and the dataset audio files were required to execute the test.

The results were analyzed in terms of SRT differences between the two trials, intra-individual standard deviation (SD_{intra}) and non-parametric statistical test. Since the subjects have hearing thresholds in the same ramge, the results were expected to be homogeneous, i.e. in a small range and with low variability.

Chapter 3

Results

The results can be divided into two thematic areas:

- a core part dedicated to the characterization of the speech material for the new test, based on the comparison between objective intelligibility metric (STOI) and subjective tests scores [Paglialonga et al., 2014], presented from Section 3.1 to 3.4;
- a final part related to the outcome of the pilot experiment, analyzing the test general performance, reliability and feasibility (Section 3.5).

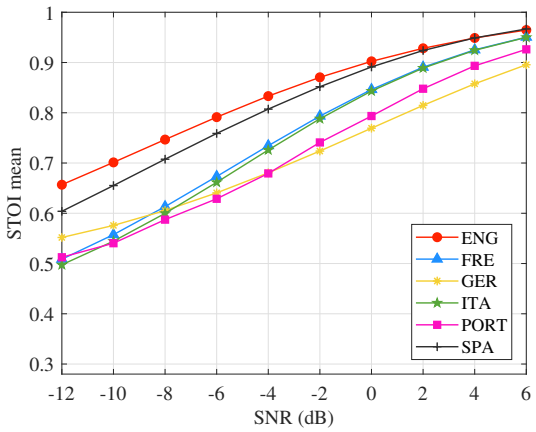
3.1 STOI simulations results

The values obtained by the simulations with 100 different realizations of Gaussian noise are plotted for each VCV in a different graph in function on the SNR; the languages are differentiated by color. The mean of the calculated values for each sample are represented in Figure 3.1. The variance among the simulations is depicted in Figure 3.2. If needed, extended numerical version can be found in Appendix A.

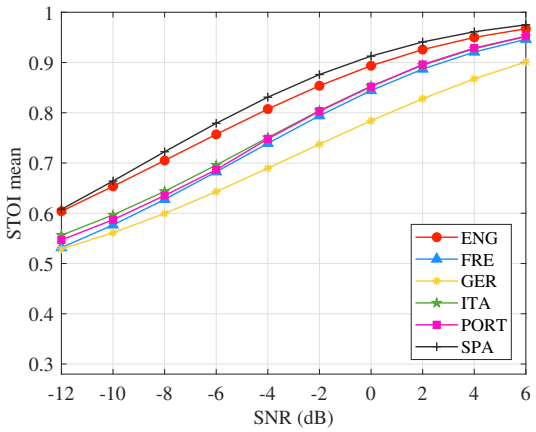
In general, it can be noticed that the relationship between STOI and SNR is monotonous: as expected, the intelligibility score increases with the increase of the SNR and it has a sigmoid-like shape. The greatest differences in intelligibility among the languages can be more easily detected for the lowest SNRs, while the curves tend to converge to a very high level of intelligibility (around 0,9).

The variance, instead, generally decreases by increasing the SNR, but the relationship is not strictly monotonous: in particular, the variance dynamics for the Portuguese is extremely irregular.

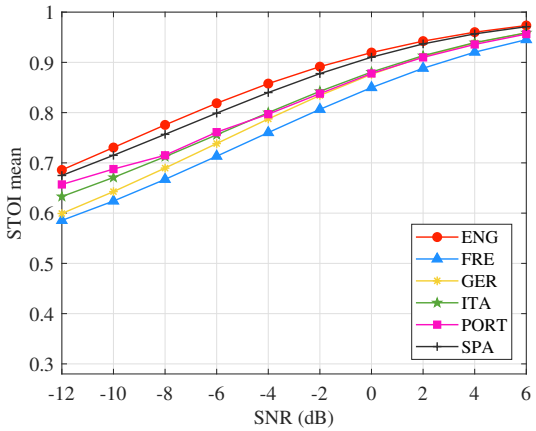
These results were analyzed separately with respect to the different VCVs and languages.



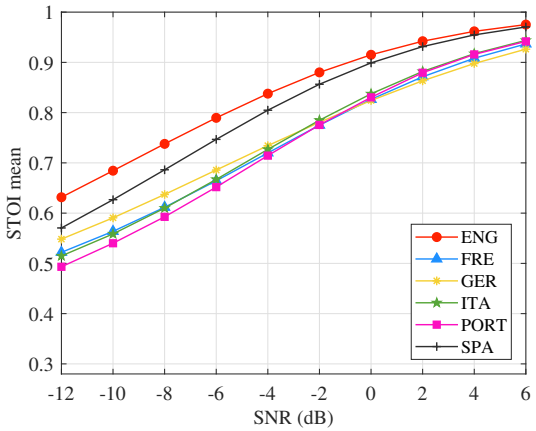
(A) ABA



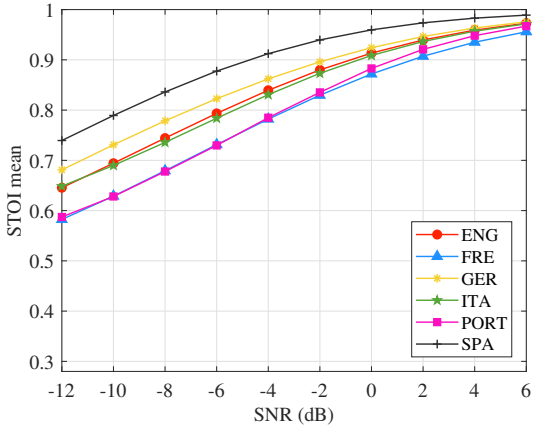
(B) ADA



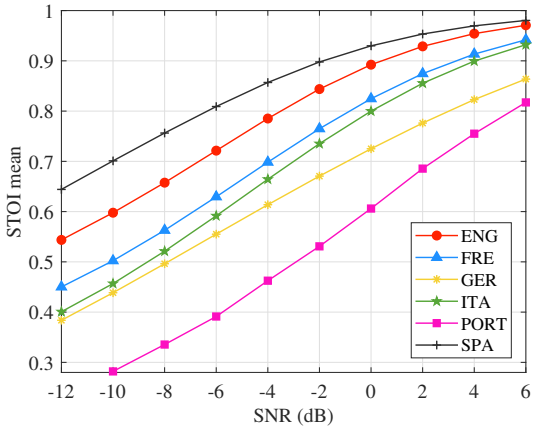
(C) AFA



(D) AGA

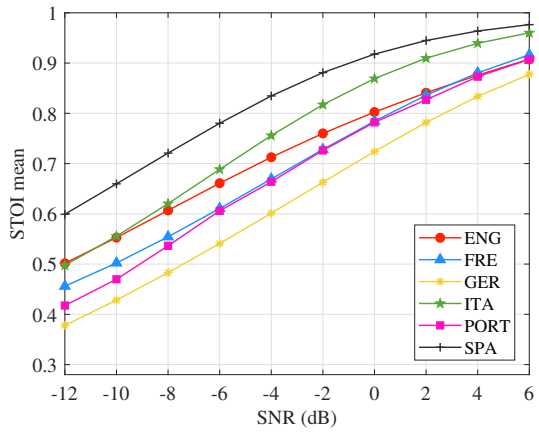


(E) AKA

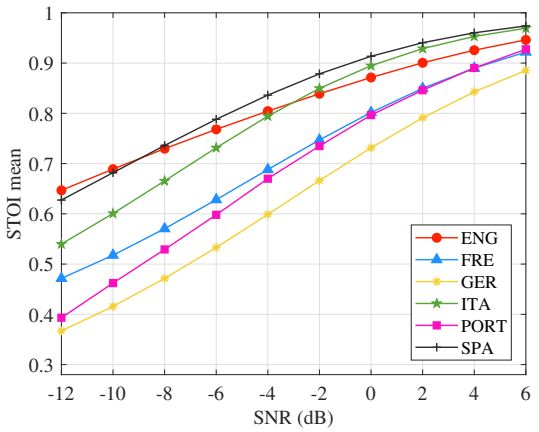


(F) ALA

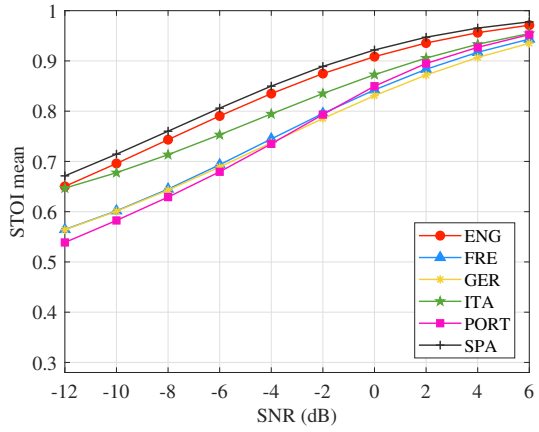
FIGURE 3.1. STOI mean for each VCV calculated on 100 Gaussian noise realizations



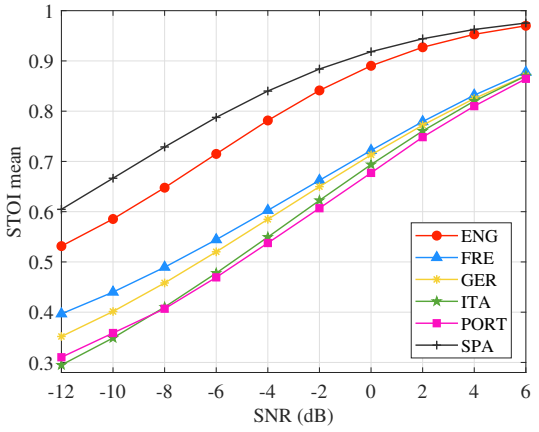
(G) AMA



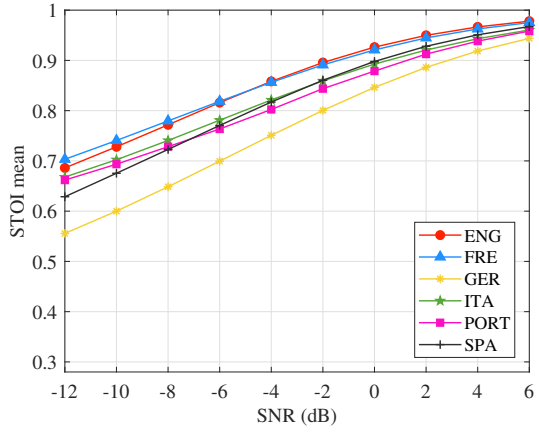
(H) ANA



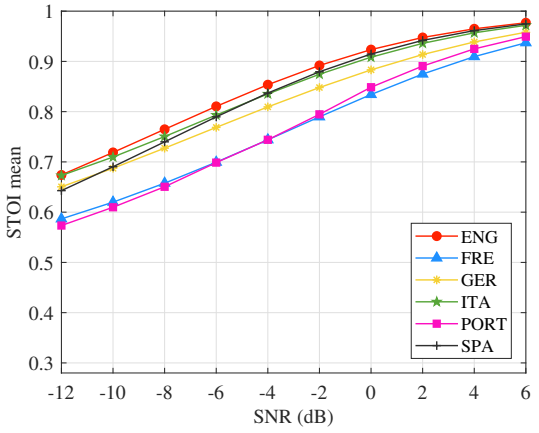
(I) APA



(J) ARA

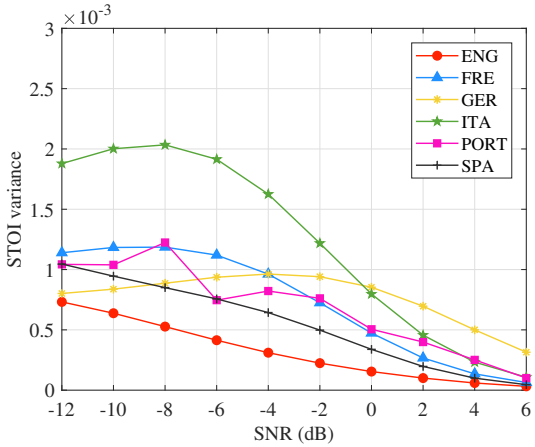


(K) ASA

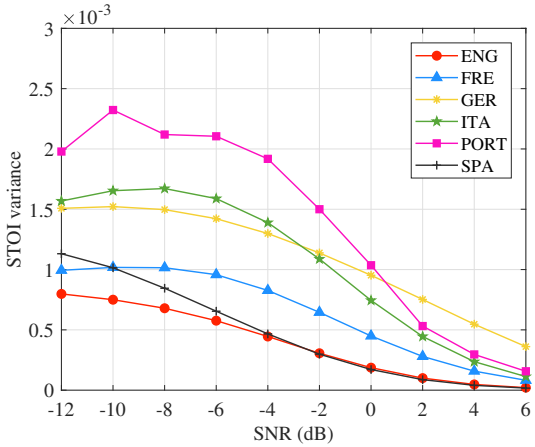


(L) ATA

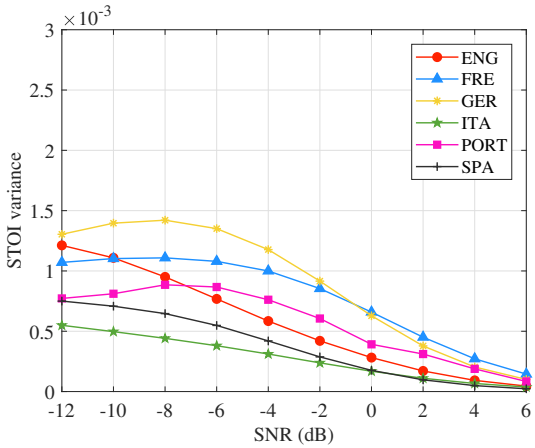
FIGURE 3.1. STOI mean for each VCV calculated on 100 Gaussian noise realizations



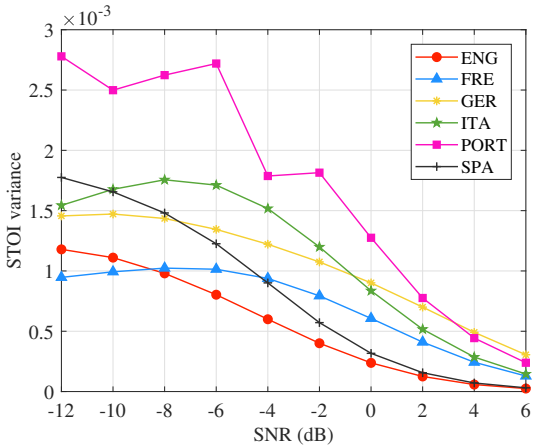
(A) ABA



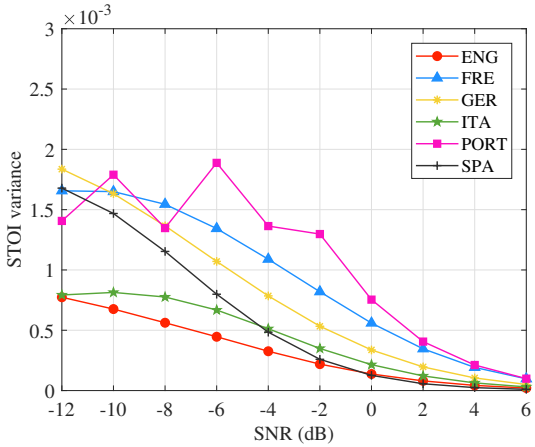
(B) ADA



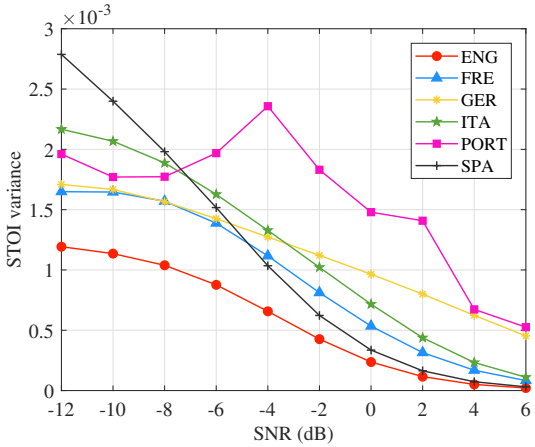
(C) AFA



(D) AGA

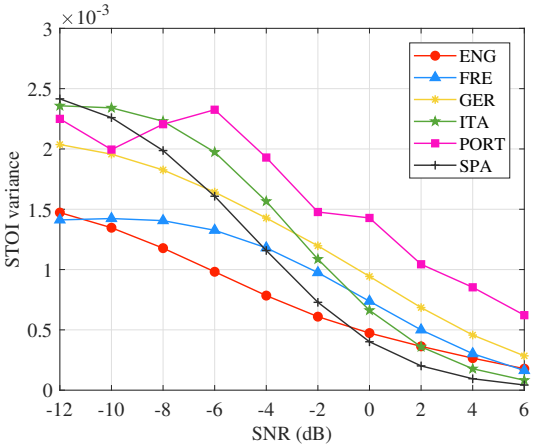


(E) AKA

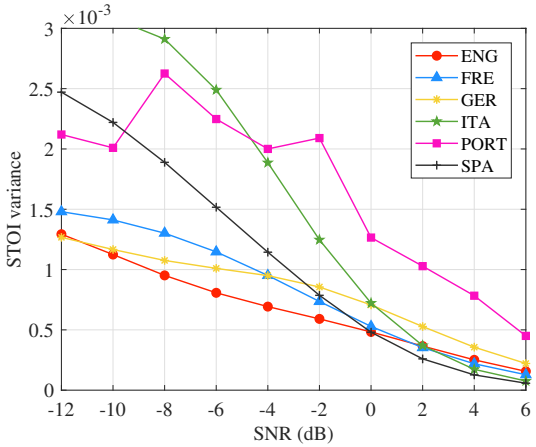


(F) ALA

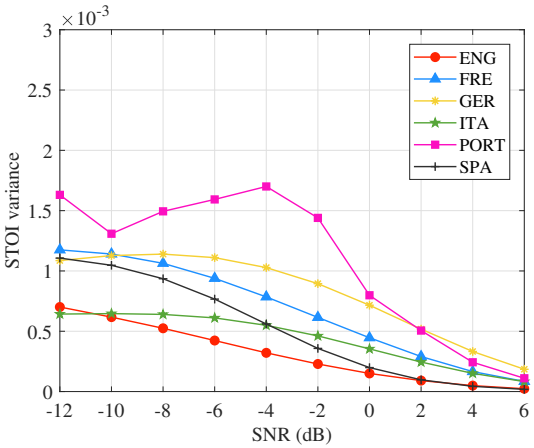
FIGURE 3.2. STOI variance for each VCV calculated on 100 Gaussian noise realizations



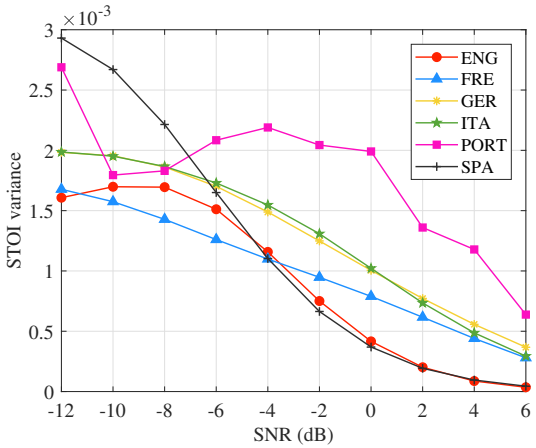
(G) AMA



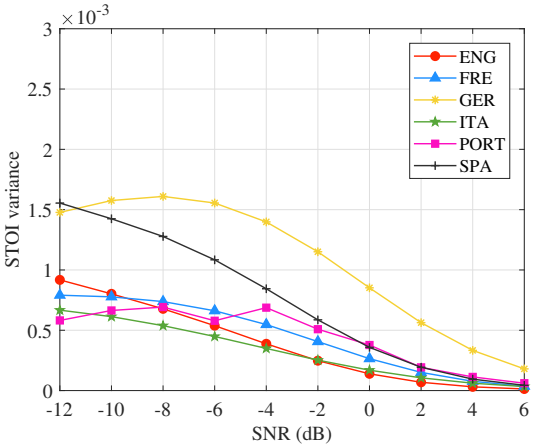
(H) ANA



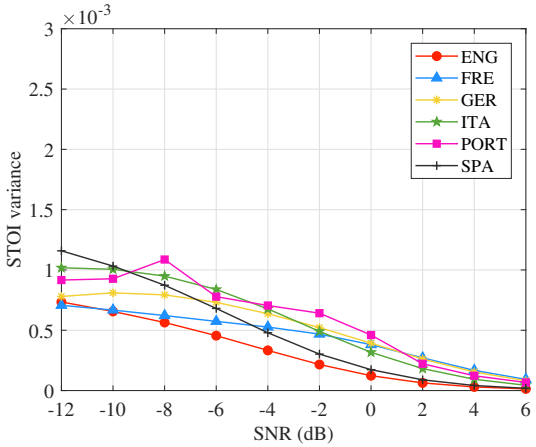
(I) APA



(J) ARA



(K) ASA



(L) ATA

FIGURE 3.2. STOI variance for each VCV calculated on 100 Gaussian noise realizations

3.1.1 STOI evaluation across languages

STOI mean and variance values were averaged across languages, so that each VCV has a single intelligibility value per SNR level. The average was not calculated also over the SNR, because it would have removed the differences in the low part of the range, which are not negligible in the evaluation. The average and the variance of the STOI mean are reported in Table 3.1 and 3.2, with the respective standard deviation (s.d.) and standard error (s.e.).

The intelligibility scores for the highest SNR are comparable for all the VCVs. In contrast, there is a marked difference for the lowest SNR. In particular, the liquid and nasal consonants (/l/,/m/,/n/,/r/) show a lower value of mean (Table 3.1) and a higher value of variance (Table 3.2), indicating that there are noticeable differences among the languages for those phonemes.

This findings are confirmed by the analysis of the STOI variance: in Table 3.3 it is shown that, globally, the nasal and the liquid consonants are the most affected by the noise stochasticity (highest mean variability) at low SNR.

Furthermore, if the above mentioned consonants exhibit in general low robustness to noise for all the languages, from Table 3.4 it emerges that intelligibility of velar consonants (/g/,/k/) is differently influenced by noise across languages.

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
ABA	0,56	0,6	0,64	0,69	0,74	0,79	0,84	0,88	0,92	0,94
ADA	0,56	0,61	0,66	0,71	0,76	0,81	0,86	0,9	0,93	0,95
AFA	0,64	0,68	0,72	0,76	0,81	0,85	0,89	0,92	0,94	0,96
AGA	0,55	0,59	0,65	0,7	0,76	0,81	0,86	0,89	0,93	0,95
AKA	0,65	0,69	0,74	0,79	0,84	0,88	0,91	0,94	0,96	0,97
ALA	0,45	0,5	0,55	0,62	0,68	0,74	0,8	0,85	0,89	0,92
AMA	0,47	0,53	0,59	0,65	0,71	0,76	0,81	0,86	0,89	0,92
ANA	0,51	0,56	0,62	0,67	0,73	0,79	0,83	0,88	0,91	0,94
APA	0,61	0,65	0,69	0,74	0,78	0,83	0,87	0,91	0,93	0,96
ARA	0,41	0,47	0,52	0,59	0,65	0,71	0,77	0,82	0,87	0,9
ASA	0,65	0,69	0,73	0,77	0,82	0,86	0,89	0,92	0,95	0,96
ATA	0,63	0,67	0,72	0,76	0,8	0,85	0,89	0,92	0,94	0,96
s.d.	0,082	0,077	0,071	0,065	0,057	0,050	0,042	0,035	0,027	0,020
s.e.	0,024	0,022	0,021	0,019	0,017	0,014	0,012	0,010	0,008	0,006

TABLE 3.1. STOI mean averaged across languages

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
ABA	4,01	4,44	4,44	4,45	4,12	3,39	2,74	1,94	1,25	0,73
ADA	1,23	1,79	2,28	2,59	2,58	2,37	1,99	1,52	1,06	0,66
AFA	1,66	1,69	1,65	1,5	1,28	0,96	0,64	0,39	0,21	0,11
AGA	2,45	2,87	3,09	3,02	2,68	2,19	1,65	1,1	0,68	0,38
AKA	3,49	3,81	3,66	3,18	2,41	1,66	0,97	0,52	0,26	0,12
ALA	18,55	20,61	20,79	20,68	18,93	16,97	13,83	9,98	6,74	4,11
AMA	5,94	6,53	6,78	6,79	6,63	5,9	4,79	3,55	2,3	1,37
ANA	13,78	13,06	12,06	10,58	8,7	6,78	4,88	3,27	1,98	1,1
APA	3,18	3,24	3,25	3,08	2,63	2,01	1,38	0,87	0,51	0,26
ARA	15,85	16,93	17,88	17,74	16,53	14,33	11,29	7,9	4,99	2,8
ASA	2,79	2,5	2,2	1,89	1,57	1,21	0,86	0,54	0,31	0,16
ATA	1,88	2,15	2,39	2,41	2,35	1,97	1,36	0,88	0,5	0,25
s.d	6,14	6,51	6,57	6,48	5,96	5,32	4,33	3,12	2,08	1,24
s.e.	1,77	1,88	1,90	1,87	1,72	1,54	1,25	0,90	0,60	0,36

TABLE 3.2. Variance of STOI mean averaged across languages (values scaled by 10^{-3})

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
ABA	1,11	1,11	1,12	0,98	0,89	0,73	0,52	0,35	0,21	0,11
ADA	1,33	1,38	1,3	1,22	1,06	0,83	0,59	0,37	0,22	0,13
AFA	0,94	0,94	0,91	0,83	0,71	0,55	0,38	0,25	0,15	0,07
AGA	1,61	1,57	1,55	1,47	1,16	0,98	0,7	0,45	0,27	0,15
AKA	1,36	1,34	1,12	1,04	0,76	0,58	0,35	0,2	0,11	0,05
ALA	1,91	1,78	1,64	1,47	1,29	0,97	0,71	0,54	0,3	0,2
AMA	1,99	1,89	1,81	1,64	1,34	1,01	0,77	0,53	0,36	0,23
ANA	1,95	1,84	1,79	1,54	1,27	1,05	0,7	0,48	0,32	0,18
APA	1,06	0,98	0,97	0,91	0,82	0,67	0,44	0,29	0,16	0,08
ARA	2,15	1,94	1,82	1,66	1,43	1,16	0,93	0,65	0,47	0,28
ATA	1	0,98	0,92	0,81	0,7	0,53	0,36	0,21	0,12	0,06
ASA	0,89	0,85	0,81	0,68	0,56	0,44	0,31	0,18	0,1	0,05
s.d	0,461	0,410	0,387	0,354	0,295	0,240	0,199	0,153	0,115	0,075
s.e.	0,133	0,118	0,112	0,102	0,085	0,069	0,058	0,044	0,033	0,022

TABLE 3.3. STOI variance averaged across languages (values scaled by 10^{-3})

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
ABA	1,68	2,26	2,66	2,64	1,9	1,19	0,71	0,45	0,25	0,11
ADA	1,89	3,29	3,04	3,53	3,37	2,4	1,42	0,68	0,36	0,16
AFA	0,88	1,06	1,18	1,24	1,13	0,82	0,47	0,22	0,08	0,02
AGA	4,1	2,87	3,63	4,7	1,91	2,58	1,52	0,73	0,33	0,12
AKA	2,17	2,24	1,45	2,72	1,6	1,72	0,64	0,21	0,06	0,01
ALA	2,92	1,82	1,13	1,27	3,27	2,41	2,1	2,41	0,76	0,5
AMA	1,98	1,74	1,85	2,23	1,55	1	1,4	0,91	0,74	0,44
ANA	5,21	5,74	6,83	4,75	2,93	3,08	0,89	0,79	0,58	0,21
APA	1,29	0,81	1,24	1,71	2,42	1,96	0,71	0,35	0,12	0,04
ARA	2,95	1,49	0,66	0,74	1,77	2,54	3,48	1,87	1,59	0,5
ATA	1,74	1,72	1,79	1,82	1,5	1,12	0,67	0,32	0,12	0,04
ASA	0,32	0,27	0,39	0,2	0,2	0,24	0,18	0,08	0,03	0,01
s.d.	1,363	1,418	1,758	1,452	0,914	0,874	0,901	0,707	0,449	0,193
s.e.	0,394	0,409	0,507	0,419	0,264	0,252	0,260	0,204	0,130	0,056

TABLE 3.4. Variance of STOI variance averaged across languages (values scaled by 10^{-7})

3.1.2 STOI evaluation across VCVs

STOI mean and variance values were averaged across VCVs, so that each language has a single intelligibility value per SNR. As in the previous section, the average was not calculated over the SNR, either, not to remove the differences in the low part of the range, which in this case are crucial for the choice of the 'optimum'. Then, all the following observations will refer only to the lowest SNR.

As reported in Table 3.5, English and Spanish present the highest intelligibility values. This is in accordance with the specifications in Table 2.2: it is important not to have significant degradation in intelligibility with low SNR not to risk to increase the difficulty of the task for non-native speakers.

The variability of the STOI mean, displayed in Table 3.6, indicates whether the different items of the dataset are well discriminated by the intelligibility measure for a specific language. It is reasonable that languages with lower mean intelligibility scores have a better discrimination throughout the dataset. However, it is interesting to notice that, despite the positive mean, results do not have a within-language negligible variability, either, compared

to Spanish in Table 3.6.

English is the language which globally shows higher robustness to noise stochasticity (Table 3.7). Although the small order of magnitude of the variability of the STOI variance, it is necessary to take this into account, because a value of $1,73 \times 10^{-3}$ (Spanish), means an eventual fluctuation from the mean intelligibility of almost 0,05 upward and downward (i.e. 0,1 in total).

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	0,62	0,67	0,72	0,77	0,82	0,86	0,9	0,93	0,95	0,96
French	0,53	0,57	0,62	0,67	0,73	0,78	0,83	0,87	0,91	0,94
German	0,51	0,56	0,6	0,65	0,7	0,75	0,8	0,85	0,88	0,92
Italian	0,55	0,59	0,64	0,7	0,75	0,81	0,85	0,89	0,93	0,95
Portuguese	0,5	0,54	0,59	0,64	0,69	0,75	0,81	0,85	0,9	0,93
Spanish	0,63	0,69	0,74	0,79	0,84	0,88	0,92	0,94	0,96	0,97
s.d	0,058	0,062	0,063	0,063	0,060	0,054	0,047	0,039	0,030	0,022
s.e.	0,017	0,018	0,018	0,018	0,017	0,016	0,014	0,011	0,009	0,006

TABLE 3.5. STOI mean averaged across VCVs

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	3,94	3,51	2,93	2,31	1,78	1,41	1,14	0,88	0,63	0,39
French	6,71	6,18	5,5	4,71	3,9	3,1	2,33	1,62	1,03	0,58
German	13,11	12,24	11,16	9,86	8,37	6,76	5,15	3,66	2,42	1,48
Italian	13,56	11,95	10,22	8,45	6,71	5,06	3,58	2,34	1,39	0,76
Portuguese	16,69	15,67	14,03	12,67	10,72	9,13	7,29	5,16	3,38	1,98
Spanish	2,01	1,7	1,38	1,07	0,78	0,52	0,31	0,17	0,08	0,04
s.d	5,932	5,546	5,016	4,546	3,885	3,288	2,606	1,848	1,215	0,724
s.e.	1,712	1,601	1,448	1,312	1,121	0,949	0,752	0,533	0,351	0,209

TABLE 3.6. Variance of STOI mean averaged across VCVs (values are scaled by 10^{-3})

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	1,05	0,97	0,86	0,72	0,55	0,39	0,25	0,15	0,09	0,05
French	1,22	1,22	1,17	1,07	0,92	0,73	0,54	0,35	0,21	0,12
German	1,44	1,43	1,37	1,28	1,14	0,97	0,77	0,57	0,39	0,24
Italian	1,52	1,53	1,48	1,33	1,1	0,83	0,56	0,34	0,19	0,1
Portuguese	1,68	1,58	1,66	1,66	1,52	1,33	0,98	0,68	0,45	0,26
Spanish	1,73	1,57	1,34	1,07	0,77	0,5	0,29	0,15	0,07	0,03
s.d.	0,263	0,242	0,274	0,316	0,335	0,341	0,280	0,216	0,154	0,098
s.e.	0,076	0,070	0,079	0,091	0,097	0,098	0,081	0,062	0,045	0,028

TABLE 3.7. STOI variance averaged across VCVs (values scaled by 10^{-3})

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	0,99	1,14	1,21	1,02	0,64	0,33	0,18	0,11	0,07	0,03
French	1,17	1,08	0,9	0,67	0,45	0,3	0,22	0,15	0,09	0,04
German	1,73	1,44	1,16	0,91	0,71	0,58	0,5	0,41	0,29	0,16
Italian	6,36	6,61	6,18	4,94	3,3	1,84	0,88	0,37	0,15	0,05
Portuguese	5,46	3,73	4,17	5,32	3,87	3,3	2,67	1,86	1,19	0,52
Spanish	5,56	4,4	3	1,74	0,85	0,36	0,13	0,04	0,01	0,01
s.d.	2,494	2,240	2,106	2,124	1,525	1,220	0,976	0,686	0,447	0,197
s.e.	0,720	0,647	0,608	0,613	0,440	0,352	0,282	0,198	0,129	0,057

TABLE 3.8. Variance of STOI variance averaged across VCVs (values scaled by 10^{-7})

	Mean average	Mean variability	Variance average	Variance variability
P-value	0,27	$< 0{,}0001$	0,0785	0,0044

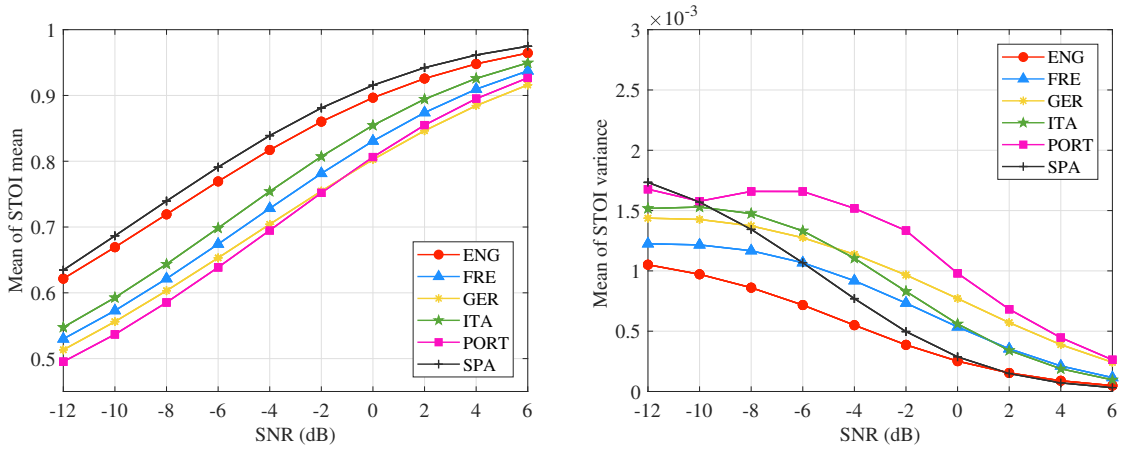
TABLE 3.9. P-value from Kruskal-Wallis test among the STOI averaged values for all languages (see Tables 3.5 – 3.8)

In Table 3.8, it can be observed that English, French and German have the highest within-dataset stability with respect to the noise, that means that all the VCVs are equally influenced to the stochasticity of the noise. Therefore, the previous results are more reliable.

The occurrence of statistical differences among languages was evaluated by statistical non-

parametric tests. Since the values do not come from a normal distribution (verification done by using one-sample Kolmogorov-Smirnov test), inter-language differences were evaluated through the Kruskal-Wallis test. P-values are reported in Table 3.9. While no differences were found among the mean averages, great differences ($p < 0{,}0001$) were found among the mean variability, meaning that STOI values span a larger range for some languages, thus better discriminating and characterizing the intelligibility of different stimuli. The situation is analogous for variance average and variability: languages show different robustness to noise stochasticity across the dataset.

All the consideration presented so far, concerning the inter-languages differences, can be summarized by the plots in Figure 3.3. By looking at the mean values on the left, languages can be visually split into two bands: English and Spanish perform better than the others, so that there is a gap between these two and the other ones. Regarding the variance, English is the language which is less noise-dependent in terms of intelligibility.



(A) Global mean (B) Global variance

FIGURE 3.3. Averaged STOI mean and variance for all languages in function on the SNR

3.2 Correlations between objective and subjective measures

In order to have an extensive outline on the correlation of the two measures, both Pearson and Spearman correlation coefficients were used.

The missing data in the Portuguese column are due to the fact that /m/, /n/ were not included in the subjective listening test [Vaez et al., 2014]. In general, it can be seen that there is a high level of linear correlation among the results ($> 0,8$, in most cases), see

Table 3.10. In particular, English and Portuguese are the languages yielding the highest correlations.

Some peculiarities are observed regarding the stimulus /aSa/: the correlation coefficient are negative for the German and not defined (n.d.) for Italian. In the first case, this is due to the fact that the recognition scores reach already 100% in one of the lowest SNR and slightly decrease after, so that the trend is waving-like. In the second case, correlation cannot be defined since the scores are constantly 100%.

Spearman correlation coefficients confirm the findings from the previous correlation measure (Table 3.11). In particular, the results relative to the two 'outliers' are coherent with previous ones. The globally improved correlation coefficients with respect to Pearson ones reveal the existence of a monotonous relationship, but not necessarily linear.

This justify the need of a non-linear mapping between the objective and the subjective scales with a logistic function (Section 2.5). For every language a scatterplot was created (Figure 3.4): the objectively predicted scores (on the x-axis) are plotted against the subjective intelligibility ratings (on the y-axis) with the respective mapping function (dashed curves).

ρ	English	French	German	Italian	Portuguese
ABA	0,9	0,83	0,92	0,8	0,94
ADA	0,91	0,66	0,88	0,94	0,83
AFA	0,96	0,95	0,9	0,96	0,89
AGA	0,88	0,55	0,89	0,87	0,85
AKA	0,83	0,72	0,91	0,99	0,91
ALA	0,98	0,83	0,8	0,7	0,85
AMA	0,74	0,56	0,63	0,69	/
ANA	0,75	0,59	0,67	0,88	/
APA	0,94	0,87	0,98	0,83	0,86
ARA	0,91	0,78	0,79	0,86	0,78
ASA	0,92	0,55	-0,65	n.d.	0,75
ATA	0,78	0,79	0,96	0,97	0,85

TABLE 3.10. Pearson correlation coefficients between objective and subjective measures

ρ_{spear}	English	French	German	Italian	Portuguese
ABA	0,9	0,95	0,93	0,76	0,98
ADA	0,88	0,76	0,93	0,98	0,93
AFA	0,84	0,97	0,85	0,98	0,97
AGA	0,74	0,18	0,95	0,87	0,90
AKA	0,81	0,58	0,92	0,98	0,76
ALA	0,96	0,88	0,72	0,77	0,71
AMA	0,51	0,41	0,28	0,73	/
ANA	0,53	0,48	0,34	0,89	/
APA	0,84	0,93	0,97	0,81	0,94
ARA	0,95	0,65	0,65	0,97	0,96
ASA	0,8	0,01	-0,71	n.d.	0,57
ATA	0,78	0,91	0,98	1	0,97

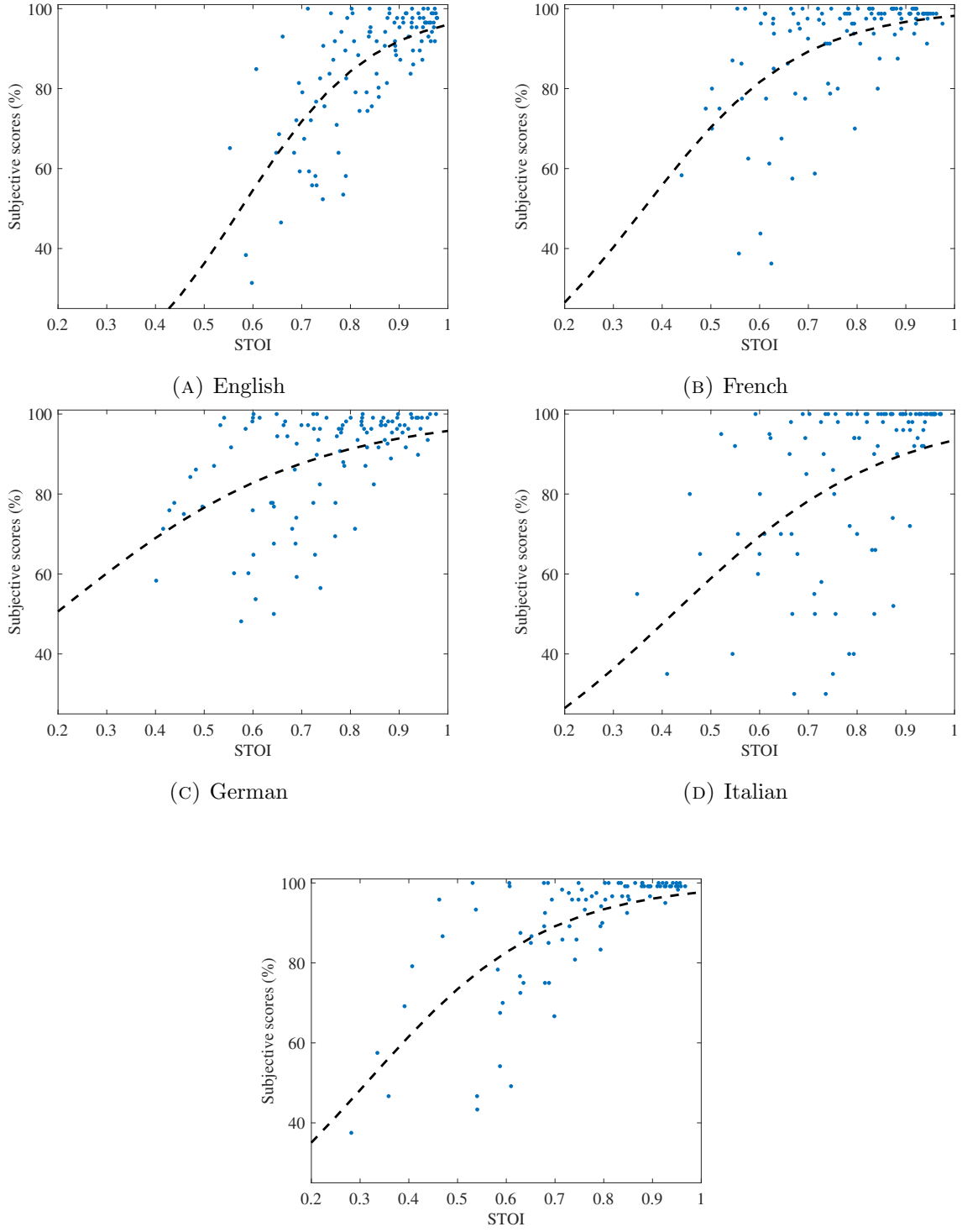
TABLE 3.11. Spearman correlation coefficients between objective and subjective measures

As can be seen in the different scatterplots, the English has the steepest mapping function, suggesting that a change in the objective intelligibility measure is followed by a nearly proportional change in the subjective score. For some of the other languages (e.g. Portuguese), the data points could be grouped in two different areas above and below the fitting function, as a sort of hysteresis phenomenon. This can be explained by the fact that intelligibility prediction yields accurate measures only for certain kinds of stimuli.

The RMSE calculated with Equation 2.2 is shown in Table 3.12. Once again the English is the language with best performance in terms of prediction (lowest prediction error), followed by the Portuguese.

	English	French	German	Italian	Portuguese
RMSE	10,45	11,30	11,53	20,18	10,48

TABLE 3.12. RMSE in %



(A) English

(B) French

(C) German

(D) Italian

(E) Portuguese

FIGURE 3.4. Scatter plots of the objectively predicted scores against the subjective intelligibility ratings with the mapping results (dashed curves)

3.3 Slope measure

The slope was measured after fitting the data with the cumulative normal model and the procedure described in Section 2.6, which is largely diffuse in psychoacoustics. Here follows an example of fitted STOI values:

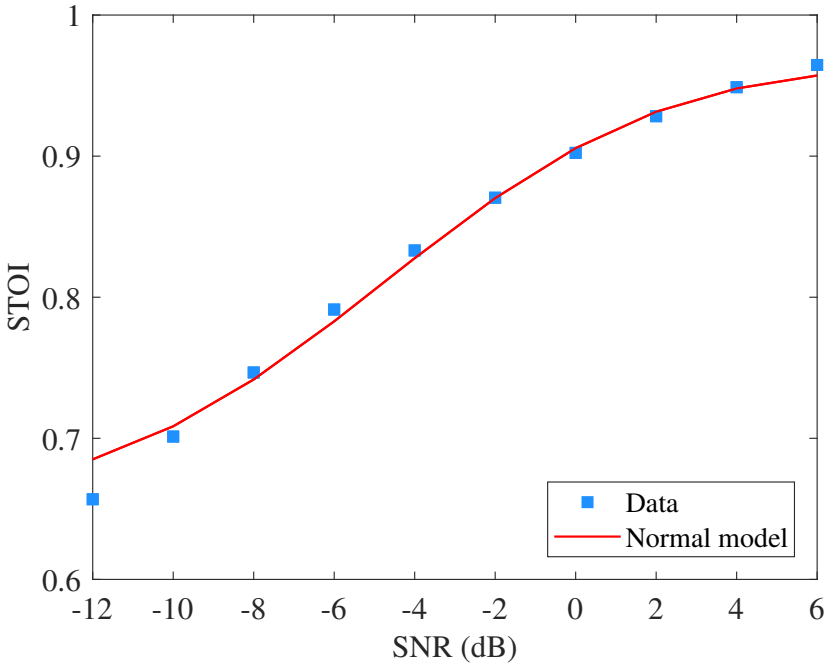


FIGURE 3.5. Averaged STOI variances for all languages in function on the SNR

The model fits well the curve characteristics in the middle of the range by paying with loss of accuracy at the extremes. However, for the sake of the slope calculation it is essential to obtain an accurate fitting around the point of inflection. Then, the approximate fitting at the extremes does not induce any bias in the estimation.

The slope values of the objective intelligibility scores are comparable for all the languages (Table 3.13). It is not surprising that Portuguese has the highest ones, since with the same range of measurement (in terms of SNR), it spans a larger interval of intelligibility values, starting from lower scores and converging above 0,9, thus resulting in a steeper curve.

The results for the subjective scores are in accordance with the correlation coefficients (3.2), especially for the outliers (/aSa/).

All the results were analyzed with statistical tests. After having verified that all the data do not come from a normal distribution by the one-sample Kolmogorov-Smirnov test, non-parametric statistical tests were performed.

	English	French	German	Italian	Portuguese	Spanish
ABA	0,023	0,034	0,025	0,035	0,032	0,027
ADA	0,027	0,031	0,027	0,03	0,032	0,028
AFA	0,021	0,027	0,027	0,024	0,023	0,022
AGA	0,026	0,031	0,027	0,033	0,035	0,031
AKA	0,025	0,028	0,021	0,025	0,030	0,019
ALA	0,034	0,038	0,033	0,04	0,043	0,026
AMA	0,028	0,034	0,036	0,036	0,036	0,029
ANA	0,021	0,034	0,038	0,033	0,038	0,026
APA	0,024	0,029	0,027	0,023	0,032	0,024
ARA	0,035	0,035	0,037	0,042	0,040	0,028
ASA	0,023	0,026	0,023	0,023	0,022	0,026
ATA	0,023	0,021	0,029	0,022	0,029	0,025

TABLE 3.13. Slope values for objective intelligibility measures (dB^{-1})

	English	French	German	Italian	Portuguese
ABA	0,024	0,080	0,062	0,080	0,063
ADA	0,035	0,042	0,046	0,058	0,054
AFA	0,049	0,087	0,055	0,098	0,031
AGA	0,033	0,015	0,048	0,100	0,076
AKA	0,022	0,011	0,014	0,137	0,025
ALA	0,100	0,038	0,019	0,024	0,077
AMA	0.0278	0.0170	0.0166	0.0344	/
ANA	0.0191	0.0229	0.0225	0.0441	/
APA	0,062	0,055	0,046	0,059	0,027
ARA	0,080	0,042	0,042	0,080	0,082
ASA	0,051	0,009	-0,004	0,000	0,004
ATA	0,026	0,040	0,051	0,135	0,067

TABLE 3.14. Slope values for subjective intelligibility measures (dB^{-1})

	English	French	German	Italian	Portuguese
P-value	0,061	0,507	0,471	0,007	0,186

TABLE 3.15. P-value from Wilcoxon rank-sum test on the objective-subjective pairs

First, the variability inter-language were investigated, i.e. among the results obtained from the same type of measure (objective or subjective). Objective measure slopes (Table 3.13) were compared through the Kruskal-Wallis test: significant differences were found between the mean of Portuguese and the mean of English and Spanish (p-value = 0,0147).

Same test was conducted for the subjective measure slopes (Table 3.14): no significant differences were found (p-value = 0,1571).

Finally, intra-language differences, i.e. between objective and subjective values for the same language, were evaluated through the Wilcoxon rank-sum paired test. The p-values can be seen in Table 3.15. No significant differences were found, except for Italian.

3.4 SRT extraction

The SRT values were derived for both objective and subjective measures.

First of all, the objective measures were analyzed by a non parametric statistical test. Since the SRT values do not come from a normal distribution (verification done by using one-sample Kolmogorov-Smirnov test), inter-language differences were evaluated through the Kruskal-Wallis test. Highly significant differences were found among the languages ($p < 0.001$): in particular, English and Spanish are significantly different from the other languages. This finding agrees with results found in Section 3.1: the higher the intelligibility, the lower the SRT.

The SRT extracted from the subjective measures are significantly different in terms of trend (Table 3.17). The undefined values are explained by the fact that the curve is always above 79,4%. Due to the presence of several undefined values and missing data, it would be meaningless and too much stimulus-specific to perform a pairwise comparison with the ones coming from the objective measure.

The value which is marked with an asterisk (/aLa/ for Italian) was manually attributed an SRT value, derived from the graph: since the curve was slightly below 79,4%, the fitting procedure missed the lowest value because of the inaccuracy at the extremes.

	English	French	German	Italian	Portuguese	Spanish
ABA	-5,499	-2,048	0,724	-1,859	-0,377	-4,332
ADA	-4,323	-2,082	0,110	-2,427	-2,348	-5,097
AFA	-6,785	-2,592	-3,624	-4,105	-4,078	-5,908
AGA	-5,483	-1,450	-1,551	-1,759	-1,531	-4,152
AKA	-5,620	-3,388	-6,959	-5,257	-3,545	-10,317
ALA	-3,584	-1,210	2,542	-0,441	7,036	-6,244
AMA	-0,629	0,032	2,240	-2,709	0,203	-5,176
ANA	-4,353	-0,534	1,851	-3,789	-0,273	-5,421
APA	-5,491	-2,125	-1,774	-3,868	-2,170	-6,193
ARA	-3,432	2,338	2,636	3,030	3,601	-5,379
ASA	-6,685	-6,936	-2,331	-5,055	-4,359	-4,766
ATA	-6,396	-1,983	-4,507	-5,637	-2,315	-5,504

TABLE 3.16. SRT extracted from objective measures (dB)

	English	French	German	Italian	Portuguese
ABA	-5,29	-3,72	-1,68	-5,65	-2,13
ADA	-4,15	-7,46	-4,59	-4,39	-4,66
AFA	-3,69	-2,82	-2,35	-2,30	-7,68
AGA	-6,58	-21,48	-5,11	-0,90	-5,29
AKA	n.d.	n.d.	n.d.	-1,91	-10,50
ALA	-1,75	-7,88	-9,96	-12*	-4,80
AMA	-8,83	n.d	-14,18	-11,95	/
ANA	-8,29	-15,58	-9,37	-7,67	/
APA	-2,97	-1,72	-3,12	-4,28	-9,54
ARA	-3,28	-4,69	-4,69	-2,96	-5,54
ASA	-5,04	n.d.	n.d.	n.d.	n.d.
ATA	-4,69	-5,72	-2,91	-1,40	-4,07

TABLE 3.17. SRT extracted from subjective measures (dB)

3.5 Test outcome

In Figure 3.6, the outcome of an adaptive test in one of the subjects who participated in the pilot study is reported as an example: the SNR level is represented in function of the question number. The typical stepwise change in SNR is the reflection of the 1U3D rule of the staircase procedure. The oscillations in the lowest part of the graph with the increase of the trial number are the SNRs relevant for the SRT estimation.

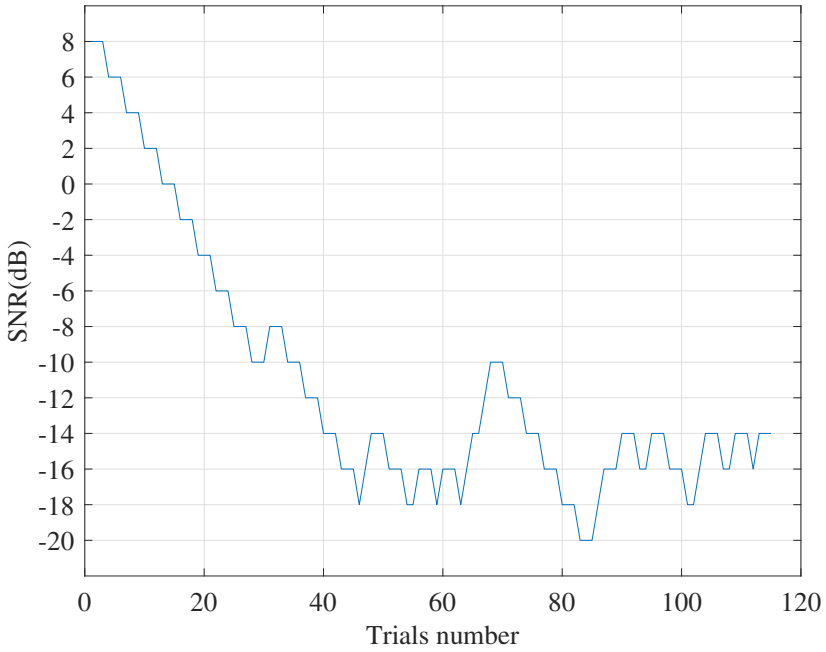


FIGURE 3.6. SNR track of a performed test

The test-retest results are reported in Table 3.18. The SRT means are the same (at the second significant digit) for both test sessions. A measure of reliability is the difference between test and retest results, estimating a possible systematic change in performance due to learning or fatigue. It is shown that the performance in the second test did not noticeably improve with respect on the first test (0,005 dB): indeed, some subjects performed actually better on the first try. Paired-sample non-parametric tests revealed that test-retest differences were not significant ($p = 0,743$). The small intra-subject standard deviations is consistent, since the subjects are all normal hearing.

The mean time for the first test is 588 seconds (9 minutes and 48 seconds), while 566 seconds (9 minutes and 26 seconds) for the second one. The small difference could be attributed to the fact that the first three subjects in the first test did not make use of a mouse, but used the laptop touchpad, increasing the duration of $2 - 3$ minutes.

Test 1 mean (sd)	Test 2 mean (sd)	Mean test-retest difference	SD_{intra}
$-15,12$ $(0,69)$	$-15,12$ $(0,82)$	$0,005$ $(p = 0,743)$	$0,71$

TABLE 3.18. Mean SRT values and test-retest characteristics

Chapter 4

Discussions

4.1 Language choice

The intelligibility prediction analysis, with the use of STOI, and the subsequent detailed comparative study among languages was aimed at the definition of the dataset for the new speech-in-noise test.

Considering both the objective intelligibility measure and the comparison with the subjective intelligibility scores, it was not possible to uniquely define a language which overcomes all the others in terms of performance.

Therefore, the first decision was whether to choose a 'global' or 'local' language, i.e. to use the same language for the entire test or to change the language depending on the VCV, using a stimulus-specific approach. Finally, a global language was chosen and thus a single speaker. The rationale behind it is that listening to the same speaker throughout the test induces minor variability in the subject's auditory experience and perception: voice changes cause significant differences in the implicit memory effect, by which exposure to a stimulus influences a response to a later stimulus [Weingarten et al., 2016, Schacter and Church, 1992]. Moreover, voice changes result in momentary cognitive overload, whose severity depends on the total number of voice changes [Potter, 2000].

Afterward, the choice of the language was done by balancing all the results exposed in Chapter 3.

By looking at the results provided by the average of the simulations across VCVs (Section 3.1.2), a first grasp of the differences in intelligibility can be obtained, particularly on how a certain language intelligibility is influenced by the presence of the noise and how it varies across the entire dataset. It can be noticed that higher intelligibility mean values are

associated to a minor ability of stimulus discrimination by STOI. As a result, SRT values are significantly lower. This trend is reasonable, since higher mean values imply an upward shift of the intelligibility measures range and a reduced within-dataset variability.

In a similar manner, a high stability with respect to noise stochasticity is associated to smaller within-dataset variability. In particular, it is interesting to notice the irregularity of the Portuguese behaviour, which is most of the times not monotonous (Figure 3.2): this is likely caused by peculiar phonetic language characteristics, as already pointed out in other studies regarding the nasal consonants [Reis and Kluge, 2008], which effectively led to the exclusion of these phonemes in the respective SUN test implementation [Beddor, 1983, Vaez et al., 2014].

Non-parametric statistical tests revealed that the differences among the averaged STOI values are not significant, in contrast to the differences in STOI variability, which resulted not statistically negligible. These results globally suggest that differences among languages could actually consist in the ability to characterize the different dataset stimuli in intelligibility in noise.

However, despite the findings obtained with the statistical tests, it is necessary to look at effect size measures, such as standard deviation and standard error, which, unlikely significance test, are not influenced by the sample size [Tomczak and Tomczak, 2014]. Indeed, a standard deviation greater than 0.06 (Table 3.5), denotes the eventuality of a difference in intelligibility perception. Since it is not well established the extent to which changes in STOI produces changes in subjective perception, i.e. the minimum ΔSTOI which produces a noticeable variation in human speech perception, such a difference cannot be neglected, especially if in other studies even smaller STOI differences were taken into account in the evaluations [Kandagatla and Subbaiah, 2018, Taseska and Habets, 2017].

This consideration is further supported by the high correlation between objective and subjective intelligibility scores, which is totally in line with previously made comparisons, extended also to other objective intelligibility measures, in which STOI always outperformed [Taal et al., 2011, Websdale et al., 2015, Xia et al., 2012].

Considering that STOI measures were demonstrated to be dependent on the noise type and processing technique [Jensen and Taal, 2014], English showed the most similar results to previous studies about STOI intelligibility prediction with additive noise, i.e. RMSE $\simeq$ 10.

The relatively low values of slope (Table 3.13 - 3.14) around 5%/dB, compared to the ones normally found in literature for a psychometric curve to be considered reliable (around 8-

10%/dB, e.g. [Leensen and Dreschler, 2013, Strasburger, 2001]), has to be attributed to the fact that those were curves (both STOI and psychometric ones from [Paglialonga et al., 2014]) referred to a single stimulus and not to a whole adaptive test run, comprising all the dataset items, thus a greater dynamics. This is the reason why they result shallower. Anyway, it is important to underline that no statistical differences were found between the slopes from objective and subjective measures, indicating once again the good prediction accuracy of STOI.

To summarize and weight all the exposed considerations, two key factors were taken into account to evaluate the most suitable language for the new test: the language variability and the response dynamics.

The insights provided into these requirements can be summarized, in details, as follows:

- **Language variability**. Small language variability — identified in this study by the STOI variance — ensures higher results stability, especially for low SNR. This is reflected, in turn, in a more reliable estimation of the psychometric curve in the experimental measures. In this regard, it is important to notice that English has the lowest STOI variance for unfavourable SNRs. It is not strictly quantified the extent to which a change in STOI produces a noticeable change in human perception in noise, but in some studies even small STOI differences ($< 5\%$) were not neglected [Kandagatla and Subbaiah, 2018, Taseska and Habets, 2017]. Hence, English lowest variance is not disregarded.

- **Response dynamics**. Since STOI is somehow the "preview" of the psychometric curve, a highly dynamic curve is desired. This implies an valuable noise effect on the intelligibility score, i.e. large curve range. Even though Portuguese yielded the highest slopes values in objective measures, significant differences were neither found among subjective measures slopes nor between subjective and objective ones. As a result, the languages dynamics can be considered comparable.

The trade-off between language variability and response dynamics was identified in the English language since despite the lowest variability, it exhibits also an appreciable dynamics. Moreover, English presents higher STOI values at low SNRs, compared to the other languages, which is a considerable factor not to increase the test difficulty for non-native speakers. A cognitive/linguistic factor in the ability to understand speech in noise has been generally pointed out [Nilsson et al., 1992]; thus, the purpose of our language-independent test is not

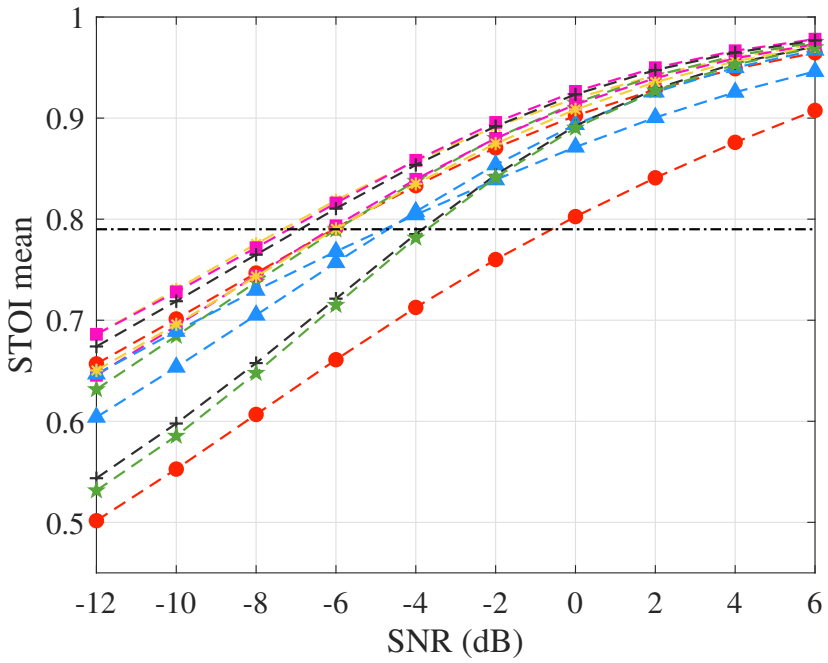


FIGURE 4.1. STOI curves for all the English VCVs and SRT values in the intersection with the dashed line (79,4%)

to enhance it, but rather to obviate and compensate for it.

This choice was supported by the high Pearson correlation coefficients, the lowest RMSE and SRT (Section 3.2 and 3.4), as well as from the widespread diffusion of English as second language.

In order for the speech stimuli to be equally difficult within the context of an adaptive procedure, grouping in homogeneous intelligibility classes had to be done. Then, the SRT of the English STOI measures were extracted to classify the stimuli in more or less intelligible depending on it, as in previous tests (e.g. [Leensen et al., 2011]). An easy-to-interpret visualization of the English stimuli is shown in Figure 4.1.

Since the thresholds range is quite narrow (they are all between -7 and -4 dB, excluding the red line), it is difficult to build classes which discriminates effectively between one stimulus and another one.

Testing an extensive amount of subjects to collect intelligibility scores for each VCV would have been too time-consuming for the sake of this study. Therefore the stimuli used for the VCV test were assumed to be homogeneous in terms of intelligibility, or at least with negligible differences.

4.2 Pilot study outcome

The test was, overall, globally feasible in terms of duration, equipment, results homogeneity and language independence.

The subjects performed the test randomly in two different rooms with uncontrolled environmental conditions both inside and outside. They all familiarized immediately with the GUI and did not encounter any difficulty during the test.

The mean test duration of nearly 10 minutes is in line with other speech in noise adaptive test durations, e.g. the HINT from 5 to 10 minutes [Nilsson et al., 1994]. However, it can be optimized without losing accuracy, as in the QuickSIN test [Killion et al., 2004], where the SRT is estimated in one minute. In some SNR tracks (an example is shown in Figure 3.6), a characteristic peak was observed in the second half of the graph: this could be attributed to a temporary lack of attention from the subject due to the lack of attention from the subject which might be relevant, even for such limited/brief test durations.

The limited number of tested subjects allowed to evaluate test-retest reliability and eventual learning effect. The extremely low mean test-retest difference (Table 3.18) reveals the absence of any learning effect, such that some subjects obtained even lower SRT in the second test session. This finding is not always granted, since in other speech in noise tests, a learning effect was detected, though small [Leensen et al., 2011, Won et al., 2007].

Reliability, meant as consistency of a test results across series of observations, was assessed through the intra-individual standard deviation (SD_{intra}). This is required to be small in order to allow a speech-in-noise test to differentiate between subjects with different degrees of hearing loss. The new test fulfills such a requirement with a SD_{intra} lower than 1 dB; it came to light that it was not only comparable, but even better than NHT (1,20 dB) and OEC (1,26 dB).

Furthermore, the low value of SD_{intra} suggests homogeneity in the SRT results in agreement with the category to which the subjects belong, since they are all normal hearing certified by PTA. Correlation with PTA was not assessed because of the different intrinsic nature of the two tests: this test is supra-threshold and requires more than audibility; moreover, the lack of hearing-impaired subjects would have not ensured correlation reliability.

Due to the lack of testing on a large number of subjects, it is impossible to establish SRT cut-off values to discriminate among the different hearing categories. Nevertheless, a comparison with cut-off ranges from other speech-in-noise tests was done.

Categories	*NHT*	*Earcheck*	*Occupational Earcheck*
Good	$\text{SRT} \leq -5,5$	$\text{SRT} \leq -10$	$\text{SRT} \leq -10$
Moderate	/	$-10 < \text{SRT} \leq -7$	$-10 < \text{SRT} \leq -8$
Insufficient	$-5,5 < \text{SRT} \leq -2,8$	$-7 < \text{SRT} \leq -4$	$-8 < \text{SRT} \leq -6$
Poor	$\text{SRT} > -2,8$	$\text{SRT} > -4$	$-6 < \text{SRT} \leq -4$
Very poor	/	/	$\text{SRT} > -4$

TABLE 4.1. Cut-off values of three different online speech-in-noise screening tests

In Table 4.1, it can be noticed that all the categories ranges are $2-3$ dB large. Assuming that all the tested subjects in this work would fall in the same category, it is reasonable to think that their maximum and minimum values, which are respectively $-16,7$ dB and $-14,2$ dB, could represent with slight variability the extremes of the SRT interval identifying normal hearing people for the new test. The higher threshold values obtained in the new test could be the result of the binaural benefit.

No evidence of language-dependent bias in the SRT estimation was found. As a matter of fact, the only English native speaker who performed the test, obtained slightly higher SRT values.

It is important to highlight that the pilot assessment had proved to be positive in terms of costs, requiring the use of common and largely diffused device, such as laptops.

In this sense, the new test met the expected outcome, providing valuable insight into future adjustments and improvements, which will be described in the next chapter (Section 5.2).

Chapter 5

Conclusions

5.1 Summary

Giving the actual need and demand of fast hearing screening methods for adults, a language-independent speech in noise test was designed with the purpose of detecting early hearing-impairment, especially of noise-induced nature. The test is binaurally presented and can be executed in domestic settings with the use of a loudspeaker.

In order to determine a dataset suitable for the test, a set of VCV previously recorded for other listening tests [Paglialonga et al., 2014, Vaez et al., 2014] for English, French, German, Italian, Portuguese and Spanish was analyzed.

STOI was calculated to measure the intelligibility prediction of each stimulus in a range from -12 to $+6$ dB. Since the noise added to the VCVs was generated by processing a Gaussian noise, the effect of noise stochasticity on intelligibility was examined by running 100 simulations with different Gaussian noise realizations. The results were averaged across VCVs and languages, aiming at gaining knowledge about the predicted intelligibility differences intra-language and inter-language.

A further comparison between objective intelligibility measures (STOI) and subjective scores from listening tests, which were previously carried out with the analyzed dataset, provided an additional mean to evaluate the different items and verify the accuracy of the intelligibility prediction. The measures were compared through correlation indices, non parametric statistical tests on slope and SRT values, after fitting the data with the cumulative normal model.

Since it was not possible to identify an absolute optimal language with outstanding results in all the measured parameters, a trade-off was searched.

Considering the language variability and the response dynamics, English was chosen, because of its smallest STOI variance and the comparable dynamic range with all the other languages. Moreover, it yielded high correlation results with the subjective scores.

The test was implemented based on an adaptive staircase procedure, following 1U3D rule [Leek, 2001, Levitt, 1971]. A fully-automated GUI was created in order to allow an easy-to-use testing.

A simple calibration procedure was included in the GUI: the subject is asked to adjust the loudspeaker volume to a comfortable level by playing an example sound. With the purpose of monitoring the environmental conditions during the test, assuming that the external noise is not stationary, short recordings of the environmental noise were used between one trial and the following one to adjust the test noise level without lowering the presentation SNR.

A pilot experiment was conducted in order to assess feasibility, time, cost, and improve the test design, with special regards on test procedure and duration. 11 voluntary normal hearing subjects were tested twice to evaluate test-retest reliability and the eventual presence of a learning effect.

No learning effect was found and small intra-individual standard deviation was shown. In this sense, the test is reliable and meets the requirements to properly discriminate among people with different degrees of hearing loss.

Due to the limited number of tested subjects and the absence of hearing-impaired people, it is not possible to determine cut-off values to classify hearing loss. However, the subjects' SRT range is compliant to the results obtained in literature [Albrecht et al., 2005].

A mean test duration of nearly 10 minutes is acceptable, but it can be improved to avoid the subject's loss of attention during the performance.

Test outcome is in line with the expectations and promising for future development.

5.2 Further research

The pilot study results suggested several modifications which could improve the test efficiency and accuracy.

First of all, it would be necessary to build homogeneous intelligibility classes for the speech material. The assumption of stimuli with equal intelligibility, based on STOI measures, revealed to be weak: many subjects spontaneously reported to find the recognition of some stimuli more easily even at low SNR, i.e. /aSa/. This is in agreement with the analyzed

subjective scores of the previous tests, where intelligibility classes were effectively identified to equalize the speech material presentation in the test, and actually /aSa/ resulted in nearly 100% recognition in all the trials and languages.

Test duration can be improved by accelerating the convergence to the target point area. As can be seen in Figure 3.6, there is a relatively extensive initial part in the test (around 30 trials) in which no mistakes were committed. Therefore, a variable step size in the procedure could be considered: larger steps in the beginning and smaller ones by progressing with the test. The criterion to switch between the different step sizes could be the occurrence of a first mistake in the response.

In order to ensure a stricter control of the environment, a real time monitoring could be implemented. The calibration procedure can be more structured and accurate, following the examples in QuickSIN [Killion et al., 2004] or other platforms, as APEX [Francart et al., 2008], in which the subject is asked to hear multiple stimuli of different nature and frequency and, consequently, adjust the volume with respect to them.

The insertion of a "I don't know" option in the multiple-choice task could be evaluated in order to reduce the subject's anxiety and not force him/her to give an answer. Naive participants are especially more comfortable with unforced tasks than with forced ones [Kaernbach, 2001]. This could imply a revision of the procedure rule and the target point estimation.

Regarding practical issues, some subjects suggested to introduce the option to perform the test with the keyboard. Further investigations could be conducted also on the influence of a priori instructions on the test results.

Appendix A

Tables

The values obtained by the simulations with 100 different realizations of Gaussian noise are reported in terms of mean and variance for each stimulus in each of the languages (see Section 2.4).

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	0,66	0,7	0,75	0,79	0,83	0,87	0,9	0,93	0,95	0,96
French	0,51	0,56	0,61	0,67	0,73	0,79	0,85	0,89	0,93	0,95
German	0,55	0,58	0,61	0,64	0,68	0,72	0,77	0,81	0,86	0,9
Italian	0,5	0,54	0,6	0,66	0,73	0,79	0,84	0,89	0,92	0,95
Portuguese	0,53	0,56	0,6	0,65	0,7	0,75	0,81	0,86	0,9	0,93
Spanish	0,6	0,66	0,71	0,76	0,81	0,85	0,89	0,92	0,95	0,97

Table A.1. Mean of AbA

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	0,73	0,64	0,53	0,41	0,31	0,22	0,15	0,1	0,06	0,03
French	1,14	1,18	1,19	1,12	0,96	0,72	0,47	0,27	0,14	0,06
German	0,8	0,84	0,89	0,94	0,96	0,94	0,85	0,7	0,5	0,31
Italian	1,88	2	2,03	1,91	1,63	1,22	0,8	0,46	0,23	0,11
Portuguese	1,17	1,53	1,54	1,38	1,47	1,51	1	0,73	0,4	0,25
Spanish	1,05	0,95	0,85	0,76	0,64	0,5	0,34	0,2	0,1	0,05

Table A.2. Variance of AbA (scaled by 10^{-3})

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	0,6	0,65	0,71	0,76	0,81	0,85	0,89	0,93	0,95	0,97
French	0,53	0,58	0,63	0,68	0,74	0,79	0,84	0,89	0,92	0,95
German	0,53	0,56	0,6	0,64	0,69	0,74	0,78	0,83	0,87	0,9
Italian	0,56	0,6	0,64	0,7	0,75	0,8	0,85	0,89	0,93	0,95
Portuguese	0,53	0,56	0,6	0,65	0,7	0,76	0,81	0,86	0,9	0,94
Spanish	0,61	0,66	0,72	0,78	0,83	0,88	0,91	0,94	0,96	0,98

Table A.3. Mean of AdA

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	0,8	0,75	0,68	0,58	0,44	0,31	0,19	0,1	0,05	0,02
French	0,99	1,02	1,02	0,96	0,83	0,65	0,45	0,28	0,16	0,08
German	1,51	1,52	1,5	1,42	1,3	1,14	0,95	0,75	0,55	0,36
Italian	1,57	1,65	1,67	1,59	1,39	1,09	0,74	0,45	0,24	0,11
Portuguese	0,88	0,75	0,67	1,02	0,63	0,79	0,45	0,36	0,19	0,08
Spanish	1,13	1,02	0,85	0,65	0,47	0,3	0,17	0,09	0,04	0,02

Table A.4. Variance of AdA (scaled by 10^{-3})

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	0,69	0,73	0,78	0,82	0,86	0,89	0,92	0,94	0,96	0,97
French	0,59	0,62	0,67	0,71	0,76	0,81	0,85	0,89	0,92	0,95
German	0,6	0,64	0,69	0,74	0,79	0,83	0,88	0,91	0,94	0,96
Italian	0,63	0,67	0,71	0,76	0,8	0,84	0,88	0,91	0,94	0,96
Portuguese	0,62	0,65	0,69	0,73	0,77	0,81	0,85	0,89	0,92	0,95
Spanish	0,67	0,72	0,76	0,8	0,84	0,88	0,91	0,94	0,96	0,97

Table A.5. Mean of AfA

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	1,21	1,11	0,95	0,77	0,58	0,42	0,28	0,17	0,09	0,05
French	1,07	1,1	1,11	1,08	1	0,85	0,66	0,45	0,27	0,14
German	1,3	1,4	1,42	1,35	1,18	0,92	0,63	0,38	0,2	0,1
Italian	0,55	0,5	0,44	0,38	0,31	0,24	0,17	0,11	0,07	0,04
Portuguese	0,76	0,92	0,74	0,87	0,85	0,82	0,65	0,43	0,23	0,14
Spanish	0,75	0,71	0,65	0,55	0,42	0,29	0,18	0,1	0,05	0,02

Table A.6. Variance of AbfA (scaled by 10^{-3})

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	0,63	0,68	0,74	0,79	0,84	0,88	0,92	0,94	0,96	0,98
French	0,52	0,56	0,61	0,66	0,72	0,77	0,83	0,87	0,91	0,94
German	0,55	0,59	0,64	0,69	0,73	0,78	0,82	0,86	0,9	0,93
Italian	0,51	0,56	0,61	0,67	0,73	0,78	0,84	0,88	0,92	0,94
Portuguese	0,5	0,55	0,6	0,67	0,73	0,79	0,85	0,9	0,93	0,96
Spanish	0,57	0,63	0,69	0,75	0,8	0,86	0,9	0,93	0,95	0,97

Table A.7. Mean of AgA

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	1,18	1,11	0,98	0,8	0,6	0,4	0,24	0,13	0,06	0,02
French	0,95	0,99	1,02	1,01	0,94	0,79	0,61	0,41	0,24	0,13
German	1,46	1,47	1,44	1,34	1,22	1,07	0,9	0,7	0,49	0,31
Italian	1,54	1,68	1,76	1,71	1,52	1,2	0,84	0,52	0,29	0,15
Portuguese	1,44	2,17	2,57	2,58	2,35	1,66	1,29	0,7	0,3	0,14
Spanish	1,78	1,66	1,48	1,23	0,9	0,57	0,32	0,16	0,07	0,03

Table A.8. Variance of AgA (scaled by 10^{-3})

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	0,65	0,69	0,74	0,79	0,84	0,88	0,91	0,94	0,96	0,97
French	0,58	0,63	0,68	0,73	0,78	0,83	0,87	0,91	0,94	0,96
German	0,68	0,73	0,78	0,82	0,86	0,9	0,92	0,95	0,96	0,98
Italian	0,65	0,69	0,74	0,78	0,83	0,87	0,91	0,94	0,96	0,97
Portuguese	0,63	0,68	0,73	0,79	0,85	0,89	0,93	0,95	0,97	0,98
Spanish	0,74	0,79	0,84	0,88	0,91	0,94	0,96	0,97	0,98	0,99

Table A.9. Mean of AkA

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	0,77	0,68	0,56	0,45	0,33	0,22	0,14	0,08	0,04	0,02
French	1,66	1,65	1,54	1,34	1,09	0,82	0,56	0,35	0,19	0,1
German	1,84	1,63	1,37	1,07	0,78	0,53	0,34	0,2	0,1	0,05
Italian	0,79	0,81	0,78	0,67	0,51	0,35	0,21	0,12	0,06	0,03
Portuguese	1,73	2,22	2,08	1,76	1,21	0,75	0,39	0,17	0,07	0,03
Spanish	1,68	1,47	1,15	0,8	0,48	0,26	0,13	0,06	0,02	0,01

Table A.10. Variance of AkA (scaled by 10^{-3})

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	0,54	0,6	0,66	0,72	0,79	0,84	0,89	0,93	0,95	0,97
French	0,45	0,5	0,56	0,63	0,7	0,76	0,82	0,87	0,91	0,94
German	0,38	0,44	0,5	0,56	0,61	0,67	0,73	0,78	0,82	0,86
Italian	0,4	0,46	0,52	0,59	0,66	0,74	0,8	0,86	0,9	0,93
Portuguese	0,3	0,36	0,41	0,48	0,56	0,65	0,72	0,79	0,85	0,9
Spanish	0,64	0,7	0,76	0,81	0,86	0,9	0,93	0,95	0,97	0,98

Table A.11. Mean of AlA

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	1,19	1,14	1,04	0,88	0,66	0,43	0,24	0,12	0,05	0,02
French	1,65	1,65	1,57	1,39	1,12	0,81	0,53	0,31	0,17	0,08
German	1,71	1,67	1,57	1,43	1,27	1,12	0,97	0,8	0,62	0,45
Italian	2,17	2,07	1,89	1,63	1,33	1,02	0,72	0,44	0,23	0,11
Portuguese	2,12	1,69	1,64	2,05	1,93	1,55	1,09	0,78	0,37	0,25
Spanish	2,79	2,4	1,98	1,52	1,03	0,62	0,34	0,16	0,07	0,03

TABLE A.12. Variance of ALA (scaled by 10^{-3})

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	0,5	0,55	0,61	0,66	0,71	0,76	0,8	0,84	0,88	0,91
French	0,46	0,5	0,55	0,61	0,67	0,73	0,78	0,84	0,88	0,92
German	0,38	0,43	0,48	0,54	0,6	0,66	0,72	0,78	0,83	0,88
Italian	0,5	0,56	0,62	0,69	0,76	0,82	0,87	0,91	0,94	0,96
Portuguese	0,54	0,61	0,66	0,72	0,77	0,82	0,87	0,91	0,94	0,96
Spanish	0,6	0,66	0,72	0,78	0,83	0,88	0,92	0,94	0,96	0,98

TABLE A.13. Mean of AMA

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	1,47	1,35	1,18	0,98	0,78	0,61	0,47	0,36	0,27	0,18
French	1,41	1,42	1,41	1,33	1,18	0,97	0,74	0,5	0,3	0,16
German	2,04	1,96	1,83	1,64	1,43	1,2	0,94	0,68	0,46	0,28
Italian	2,36	2,34	2,23	1,97	1,57	1,09	0,66	0,36	0,18	0,08
Portuguese	1,66	1,6	1,69	1,49	1,31	1	0,78	0,51	0,33	0,15
Spanish	2,42	2,26	1,99	1,61	1,16	0,73	0,4	0,2	0,09	0,04

TABLE A.14. Variance of AMA (scaled by 10^{-3})

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	0,65	0,69	0,73	0,77	0,8	0,84	0,87	0,9	0,93	0,95
French	0,47	0,52	0,57	0,63	0,69	0,75	0,8	0,85	0,89	0,92
German	0,37	0,42	0,47	0,53	0,6	0,67	0,73	0,79	0,84	0,89
Italian	0,54	0,6	0,67	0,73	0,79	0,85	0,89	0,93	0,95	0,97
Portuguese	0,46	0,53	0,6	0,66	0,72	0,79	0,84	0,89	0,92	0,95
Spanish	0,63	0,68	0,74	0,79	0,84	0,88	0,91	0,94	0,96	0,97

Table A.15. Mean of AnA

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	1,29	1,12	0,95	0,81	0,69	0,59	0,48	0,37	0,25	0,16
French	1,48	1,41	1,3	1,15	0,95	0,74	0,53	0,35	0,22	0,13
German	1,27	1,17	1,08	1,01	0,95	0,86	0,71	0,53	0,36	0,22
Italian	3,04	3,08	2,91	2,49	1,89	1,25	0,72	0,37	0,17	0,08
Portuguese	2,49	2,09	2,06	2,8	1,75	1,6	1,14	0,83	0,54	0,32
Spanish	2,47	2,22	1,89	1,52	1,14	0,79	0,48	0,26	0,13	0,06

Table A.16. Variance of AnA (scaled by 10^{-3})

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	0,65	0,7	0,74	0,79	0,83	0,87	0,91	0,94	0,96	0,97
French	0,56	0,6	0,64	0,69	0,74	0,8	0,84	0,88	0,92	0,94
German	0,56	0,6	0,64	0,69	0,74	0,79	0,83	0,87	0,91	0,94
Italian	0,65	0,68	0,71	0,75	0,79	0,84	0,87	0,91	0,93	0,95
Portuguese	0,55	0,59	0,63	0,69	0,74	0,79	0,84	0,89	0,92	0,95
Spanish	0,67	0,71	0,76	0,81	0,85	0,89	0,92	0,95	0,97	0,98

Table A.17. Mean of ApA

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	0,7	0,62	0,52	0,42	0,32	0,23	0,15	0,09	0,05	0,02
French	1,17	1,14	1,06	0,94	0,78	0,61	0,45	0,29	0,17	0,08
German	1,09	1,13	1,14	1,11	1,03	0,89	0,72	0,52	0,33	0,18
Italian	0,64	0,65	0,64	0,61	0,55	0,46	0,35	0,24	0,15	0,08
Portuguese	1,6	1,4	2,05	1,97	1,83	1,45	0,97	0,74	0,4	0,23
Spanish	1,11	1,05	0,94	0,77	0,56	0,36	0,2	0,1	0,04	0,02

TABLE A.18. Variance of ApA (scaled by 10^{-3})

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	0,53	0,59	0,65	0,71	0,78	0,84	0,89	0,93	0,95	0,97
French	0,4	0,44	0,49	0,54	0,6	0,66	0,72	0,78	0,83	0,88
German	0,35	0,4	0,46	0,52	0,58	0,65	0,71	0,77	0,83	0,87
Italian	0,29	0,35	0,41	0,48	0,55	0,62	0,69	0,76	0,82	0,87
Portuguese	0,37	0,41	0,46	0,52	0,59	0,66	0,74	0,8	0,86	0,91
Spanish	0,6	0,67	0,73	0,79	0,84	0,88	0,92	0,94	0,96	0,98

TABLE A.19. Mean of ArA

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	1,61	1,7	1,69	1,51	1,16	0,75	0,42	0,2	0,09	0,04
French	1,68	1,57	1,43	1,26	1,1	0,95	0,79	0,62	0,44	0,28
German	1,98	1,96	1,86	1,7	1,49	1,25	1	0,77	0,56	0,37
Italian	1,99	1,95	1,87	1,73	1,55	1,31	1,02	0,74	0,49	0,29
Portuguese	1,58	1,96	1,45	1,82	2,13	1,71	1,61	1,03	0,66	0,31
Spanish	2,93	2,67	2,21	1,65	1,1	0,66	0,37	0,19	0,1	0,04

TABLE A.20. Variance of ArA (scaled by 10^{-3})

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	0,69	0,73	0,77	0,82	0,86	0,9	0,93	0,95	0,97	0,98
French	0,7	0,74	0,78	0,82	0,86	0,89	0,92	0,94	0,96	0,98
German	0,56	0,6	0,65	0,7	0,75	0,8	0,85	0,89	0,92	0,94
Italian	0,67	0,7	0,74	0,78	0,82	0,86	0,89	0,92	0,94	0,96
Portuguese	0,66	0,69	0,73	0,76	0,8	0,84	0,88	0,91	0,94	0,96
Spanish	0,63	0,68	0,72	0,77	0,82	0,86	0,9	0,93	0,95	0,97

TABLE A.21. Mean of AsA

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	0,92	0,8	0,68	0,54	0,39	0,25	0,14	0,07	0,03	0,01
French	0,79	0,78	0,74	0,66	0,55	0,41	0,26	0,15	0,08	0,04
German	1,48	1,58	1,61	1,56	1,4	1,15	0,85	0,56	0,33	0,18
Italian	0,67	0,61	0,54	0,45	0,35	0,25	0,17	0,11	0,06	0,03
Portuguese	0,58	0,66	0,69	0,58	0,9	0,51	0,38	0,19	0,11	0,06
Spanish	1,55	1,42	1,28	1,08	0,84	0,59	0,36	0,19	0,09	0,04

TABLE A.22. Variance of AsA (scaled by 10^{-3})

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	0,67	0,72	0,76	0,81	0,85	0,89	0,92	0,95	0,96	0,98
French	0,59	0,62	0,66	0,7	0,74	0,79	0,83	0,87	0,91	0,94
German	0,65	0,69	0,73	0,77	0,81	0,85	0,88	0,91	0,94	0,96
Italian	0,67	0,71	0,75	0,79	0,84	0,87	0,91	0,94	0,96	0,97
Portuguese	0,56	0,59	0,64	0,69	0,75	0,8	0,85	0,89	0,93	0,95
Spanish	0,64	0,69	0,74	0,79	0,84	0,88	0,91	0,94	0,96	0,97

TABLE A.23. Mean of AtA

SNR [dB]	-12	-10	-8	-6	-4	-2	0	2	4	6
English	0,73	0,66	0,56	0,46	0,33	0,22	0,12	0,06	0,03	0,01
French	0,71	0,67	0,62	0,57	0,53	0,47	0,38	0,27	0,17	0,09
German	0,78	0,81	0,79	0,73	0,64	0,52	0,39	0,26	0,15	0,08
Italian	1,02	1,01	0,95	0,84	0,68	0,49	0,32	0,18	0,09	0,04
Portuguese	1,64	1,88	1,66	1,66	1,78	1,39	0,89	0,58	0,28	0,15
Spanish	1,16	1,03	0,87	0,68	0,48	0,3	0,17	0,09	0,04	0,02

TABLE A.24. Variance of AtA (scaled by 10^{-3})

Bibliography

[Albrecht et al., 2005] Albrecht, J., van Elewout, L., Verhage, L., and Verweij, C. (2005). Oorcheck: de validering van een interactief screeningsinstrument. *Internal report, LUMC Leiden*.

[American Speech Language Hearing Association and others, 1997] American Speech Language Hearing Association and others (1997). Guidelines for audiologic screening.

[Arlinger, 2003] Arlinger, S. (2003). Negative consequences of uncorrected hearing loss-a review. *International journal of audiology*, 42:2S17–2S20.

[Association, 1999] Association, I. P. (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

[Baer et al., 2002] Baer, T., Moore, B. C., and Kluk, K. (2002). Effects of low pass filtering on the intelligibility of speech in noise for people with and without dead regions at high frequencies. *The Journal of the Acoustical Society of America*, 112(3):1133–1144.

[Beddor, 1983] Beddor, P. S. (1983). Phonological and phonetic effects of nasalization on vowel height.

[Boldt and Ellis, 2009] Boldt, J. B. and Ellis, D. P. (2009). A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation. In *Signal Processing Conference, 2009 17th European*, pages 1849–1853. IEEE.

[Byrne et al., 1994] Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., Hagerman, B., Hetu, R., Kei, J., Lui, C., et al. (1994). An international comparison of long-term average speech spectra. *The Journal of the Acoustical Society of America*, 96(4):2108–2120.

[Chauhan et al., 1993] Chauhan, B. C., Tompkins, J. D., LeBlanc, R. P., and McCormick, T. A. (1993). Characteristics of frequency-of-seeing curves in normal subjects, patients with suspected glaucoma, and patients with glaucoma. *Investigative ophthalmology & visual science*, 34(13):3534–3540.

[Cooke et al., 2010] Cooke, M., Lecumberri, M. L. G., Scharenborg, O., and van Dommelen, W. A. (2010). Language-independent processing in speech perception: Identification of english intervocalic consonants by speakers of eight european languages. *Speech Communication*, 52(11):954–967.

[Culling et al., 2005] Culling, J. F., Zhao, F., and Stephens, D. (2005). The viability of speech-in-noise audiometric screening using domestic audio equipment: La viabilidad del tamizaje audiométrico con lenguaje en ruido utilizando equipo doméstico de audio. *International journal of audiology*, 44(12):691–700.

[Dau et al., 1996] Dau, T., Püschel, D., and Kohlrausch, A. (1996). A quantitative model of the effectivesignal processing in the auditory system. i. model structure. *The Journal of the Acoustical Society of America*, 99(6):3615–3622.

[Davis et al., 2007] Davis, A., Smith, P., Ferguson, M., Stephens, D., and Gianopoulos, I. (2007). Acceptability, benefit and costs of early screening for hearing disability: a study of potential screening tests and models. *Health Technology Assessment Southampton*, 11(42).

[Deprez et al., 2013] Deprez, H., Yilmaz, E., Lievens, S., and Van Hamme, H. (2013). Automating speech reception threshold measurements using automatic speech recognition. In *Proceedings of the Fourth Workshop on Speech and Language Processing for Assistive Technologies*, pages 35–40.

[Donaldson and Kreft, 2006] Donaldson, G. S. and Kreft, H. A. (2006). Effects of vowel context on the recognition of initial and medial consonants by cochlear implant users. *Ear and hearing*, 27(6):658–677.

[Dorman et al., 1990] Dorman, M. F., Soli, S., Dankowski, K., Smith, L. M., McCandless, G., and Parkin, J. (1990). Acoustic cues for consonant identification by patients who use the ineraid cochlear implant. *The Journal of the Acoustical Society of America*, 88(5):2074–2079.

[Dubno and Dirks, 1982] Dubno, J. R. and Dirks, D. D. (1982). Evaluation of hearing-impaired listeners using a nonsense-syllable test i. test reliability. *Journal of Speech, Language, and Hearing Research*, 25(1):135–141.

[Ellis et al., 2006] Ellis, N., Kuijpers, M., van der Pijl, S., and Verbiest, E. (2006). Ontwikkeling van een gehoorscreening voor het bedrijfsleven. *Internal report, LUMC Leiden.*

[Etymotic Research, 2018] Etymotic Research (2018). `https://www.etymotic.com/`. Last accessed on March 2018.

[Falk et al., 2015] Falk, T. H., Parsa, V., Santos, J. F., Arehart, K., Hazrati, O., Huber, R., Kates, J. M., and Scollie, S. (2015). Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools. *IEEE signal processing magazine*, 32(2):114–124.

[Finney and Tattersfield, 1952] Finney, D. J. and Tattersfield, F. (1952). *Probit analysis.* Cambridge University Press; Cambridge.

[Francart et al., 2009] Francart, T., Moonen, M., and Wouters, J. (2009). Automatic testing of speech recognition. *International Journal of Audiology*, 48(2):80–90.

[Francart et al., 2008] Francart, T., Van Wieringen, A., and Wouters, J. (2008). Apex 3: a multi-purpose test platform for auditory psychophysical experiments. *Journal of Neuroscience Methods*, 172(2):283–293.

[Garc´a-Pérez, 1998] Garc´a-Pérez, M. A. (1998). Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties. *Vision research*, 38(12):1861–1881.

[Gierut, 1989] Gierut, J. A. (1989). Maximal opposition approach to phonological treatment. *Journal of Speech and Hearing Disorders*, 54(1):9–19.

[Goldsworthy and Greenberg, 2004] Goldsworthy, R. L. and Greenberg, J. E. (2004). Analysis of speech-based speech transmission index methods with implications for nonlinear operations. *The Journal of the Acoustical Society of America*, 116(6):3679–3689.

[Gordon-Salant, 2005] Gordon-Salant, S. (2005). Hearing loss and aging: new research findings and clinical implications. *Journal of rehabilitation research and development*, 42(4):9.

[Grandori et al., 2009] Grandori, F., Parazzini, M., Tognola, G., and Paglialonga, A. (2009). Hearing screening in older adults is gaining momentum-the european project ahead iii on

adult hearing. In *Proceedings of the 2nd Phonak International Adult Conference: Hearing care for adults*, pages 191–202.

[Green et al., 1989] Green, D. M., Richards, V. M., and Forrest, T. (1989). Stimulus step size and heterogeneous stimulus conditions in adaptive psychophysics. *The Journal of the Acoustical Society of America*, 86(2):629–636.

[Hu and Loizou, 2008] Hu, Y. and Loizou, P. C. (2008). Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on audio, speech, and language processing*, 16(1):229–238.

[Humes, 2013] Humes, L. E. (2013). Understanding the speech-understanding problems of older adults. *American Journal of Audiology*, 22(2):303–305.

[ISO 7029:2017, 2017] ISO 7029:2017 (2017). Acoustics Statistical distribution of hearing thresholds related to age and gender. Standard.

[ISO 8253-3:1996(E), 1996] ISO 8253-3:1996(E) (1996). Acoustics audiometric test methods part 3: Speech audiometry. Standard.

[Jensen and Taal, 2014] Jensen, J. and Taal, C. H. (2014). Speech intelligibility prediction based on mutual information. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 22(2):430–440.

[Kaernbach, 2001] Kaernbach, C. (2001). Adaptive threshold estimation with unforced-choice tasks. *Perception & Psychophysics*, 63(8):1377–1388.

[Kandagatla and Subbaiah, 2018] Kandagatla, R. K. and Subbaiah, P. (2018). Speech enhancement using mmse estimation of amplitude and complex speech spectral coefficients under phase-uncertainty. *Speech Communication*, 96:10–27.

[Kates and Arehart, 2005] Kates, J. M. and Arehart, K. H. (2005). Coherence and the speech intelligibility index. *The journal of the acoustical society of America*, 117(4):2224–2237.

[Killion and Niquette, 2000] Killion, M. C. and Niquette, P. A. (2000). What can the pure-tone audiogram tell us about a patients snr loss. *Hear J*, 53(3):46–53.

[Killion et al., 2004] Killion, M. C., Niquette, P. A., Gudmundsen, G. I., Revit, L. J., and Banerjee, S. (2004). Development of a quick speech-in-noise test for measuring signal-

to-noise ratio loss in normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 116(4):2395–2405.

[Kollmeier et al., 1988] Kollmeier, B., Gilkey, R. H., and Sieben, U. K. (1988). Adaptive staircase techniques in psychoacoustics: A comparison of human data and a mathematical model. *The Journal of the Acoustical Society of America*, 83(5):1852–1862.

[Kryter, 1962] Kryter, K. D. (1962). Methods for the calculation and use of the articulation index. *The Journal of the Acoustical Society of America*, 34(11):1689–1697.

[Laplante-Lévesque et al., 2011] Laplante-Lévesque, A., Hickson, L., and Worrall, L. (2011). Predictors of rehabilitation intervention decisions in adults with acquired hearing impairment. *Journal of Speech, Language, and Hearing Research*, 54(5):1385–1399.

[Leek, 2001] Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception & psychophysics*, 63(8):1279–1292.

[Leek et al., 1992] Leek, M. R., Hanna, T. E., and Marshall, L. (1992). Estimation of psychometric functions from adaptive tracking procedures. *Perception & Psychophysics*, 51(3):247–256.

[Leensen et al., 2011] Leensen, M. C., de Laat, J. A., Snik, A. F., and Dreschler, W. A. (2011). Speech-in-noise screening tests by internet, part 2: improving test sensitivity for noise-induced hearing loss. *International journal of audiology*, 50(11):835–848.

[Leensen and Dreschler, 2013] Leensen, M. C. J. and Dreschler, W. A. (2013). Speech-in-noise screening tests by internet, part 3: Test sensitivity for uncontrolled parameters in domestic usage. *International Journal of Audiology*, 52(10):658–669.

[Levitt, 1971] Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical society of America*, 49(2B):467–477.

[Liu et al., 2011] Liu, C.-F., Collins, M. P., Souza, P. E., and Yueh, B. (2011). Long-term cost-effectiveness of screening strategies for hearing loss. *Journal of rehabilitation research and development*, 48(3):235.

[Loizou et al., 2000] Loizou, P. C., Poroy, O., and Dorman, M. (2000). The effect of parametric variations of cochlear implant processors on speech understanding. *The Journal of the Acoustical Society of America*, 108(2):790–802.

[Lyregaard, 1997] Lyregaard, P. (1997). Towards a theory of speech audiometry tests. *Speech audiometry*, 2:34–62.

[Meyer-Bisch, 1996] Meyer-Bisch, C. (1996). Epidemiological evaluation of hearing damage related to strongly amplified music (personal cassette players, discotheques, rock concerts)-high-definition audiometric survey on 1364 subjects. *Audiology*, 35(3):121–142.

[Moore et al., 2010] Moore, D. R., Fuchs, P. A., Plack, C., Rees, A., and Palmer, A. R. (2010). *Oxford Handbook of Auditory Science: Hearing*, volume 3. Oxford University Press.

[Nachtegaal et al., 2009] Nachtegaal, J., Smit, J. H., Smits, C., Bezemer, P. D., Van Beek, J. H., Festen, J. M., and Kramer, S. E. (2009). The association between hearing status and psychosocial health before the age of 70 years: results from an internet-based national survey on hearing. *Ear and Hearing*, 30(3):302–312.

[Nielsen and Dau, 2009] Nielsen, J. B. and Dau, T. (2009). Development of a danish speech intelligibility test. *International journal of audiology*, 48(10):729–741.

[NIH - National Library of Medicine, 1993] NIH - National Library of Medicine (1993). https://www.nlm.nih.gov/.

[Nilsson et al., 1992] Nilsson, M., Gelnett, D., Sullivan, J., Soli, S. D., and Goldberg, R. L. (1992). Norms for the hearing in noise test: The influence of spatial separation, hearing loss, and english language experience on speech reception thresholds. *The Journal of the Acoustical Society of America*, 92(4):2385–2385.

[Nilsson et al., 1994] Nilsson, M., Soli, S. D., and Sullivan, J. A. (1994). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America*, 95(2):1085–1099.

[Nittrouer and Boothroyd, 1990] Nittrouer, S. and Boothroyd, A. (1990). Context effects in phoneme and word recognition by young children and older adults. *The Journal of the Acoustical Society of America*, 87(6):2705–2715.

[Ozimek et al., 2009] Ozimek, E., Kutzner, D., Sek, A., and Wicher, A. (2009). Polish sentence tests for measuring the intelligibility of speech in interfering noise. *International Journal of Audiology*, 48(7):433–443.

[Paglialonga, 2009] Paglialonga, A. (2009). *Advanced Methods and Models for The Investigation of the Human Hearing Function in Tinnitus*. PhD thesis, Politecnico di Milano.

[Paglialonga et al., 2011] Paglialonga, A., Fiocchi, S., Parazzini, M., Ravazzani, P., and Tognola, G. (2011). Influence of tinnitus sound therapy signals on the intelligibility of speech. *The Journal of laryngology and otology*, 125(8):795.

[Paglialonga et al., 2014] Paglialonga, A., Tognola, G., and Grandori, F. (2014). A user-operated test of suprathreshold acuity in noise for adult hearing screening: the sun (speech understanding in noise) test. *Computers in biology and medicine*, 52:66–72.

[Potter, 2000] Potter, R. F. (2000). The effects of voice changes on orienting and immediate cognitive overload in radio listeners. *Media Psychology*, 2(2):147–177.

[Reis and Kluge, 2008] Reis, M. S. and Kluge, D. C. (2008). Intelligibility of brazilian portuguese-accented english realization of nasals in word-final position by brazilian and dutch efl learners. *Revista Crop*, 13:215–229.

[Schacter and Church, 1992] Schacter, D. L. and Church, B. A. (1992). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(5):915.

[Simons and Fennig, 2017] Simons, G. F. and Fennig, C. D. (2017). *Ethnologue: Languages of the World*. SIL International, Dalla, Texas, 20^{th} edition.

[Smits et al., 2004] Smits, C., Kapteyn, T. S., and Houtgast, T. (2004). Development and validation of an automatic speech-in-noise screening test by telephone. *International journal of audiology*, 43(1):15–28.

[Smits et al., 2006] Smits, C., Merkus, P., and Houtgast, T. (2006). How we do it: The dutch functional hearing–screening tests by telephone and internet. *Clinical Otolaryngology*, 31(5):436–440.

[Strasburger, 2001] Strasburger, H. (2001). Converting between measures of slope of the psychometric function. *Perception & psychophysics*, 63(8):1348–1355.

[Strom, 2003] Strom, K. E. (2003). The hr 2003 dispenser survey. *Hearing Review*, 10(6):22–41.

[Taal et al., 2011] Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2011). An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136.

[Taseska and Habets, 2017] Taseska, M. and Habets, E. A. (2017). Doa-informed source extraction in the presence of competing talkers and background noise. *EURASIP Journal on Advances in Signal Processing*, 2017(1):60.

[Tomczak and Tomczak, 2014] Tomczak, M. and Tomczak, E. (2014). The need to report effect size estimates revisited. an overview of some recommended measures of effect size. *Trends in Sport Sciences*, 21(1).

[Treutwein, 1995] Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision research*, 35(17):2503–2522.

[Vaez et al., 2014] Vaez, N., Desgualdo-Pereira, L., and Paglialonga, A. (2014). Development of a test of suprathreshold acuity in noise in brazilian portuguese: a new method for hearing screening and surveillance. *BioMed research international*, 2014.

[Versfeld et al., 2000] Versfeld, N. J., Daalder, L., Festen, J. M., and Houtgast, T. (2000). Method for the selection of sentence materials for efficient measurement of the speech reception threshold. *The Journal of the Acoustical Society of America*, 107(3):1671–1684.

[Vogel et al., 2007] Vogel, I., Brug, J., Van der Ploeg, C. P., and Raat, H. (2007). Young peoples exposure to loud music. *American Journal of Preventive Medicine*, 33(2):124–133.

[Websdale et al., 2015] Websdale, D., Le Cornu, T., and Milner, B. (2015). Objective measures for predicting the intelligibility of spectrally smoothed speech with artificial excitation. *Proceedings of Interspeech 2015*.

[Weingarten et al., 2016] Weingarten, E., Chen, Q., McAdams, M., Yi, J., Hepler, J., and Albarrac´n, D. (2016). From primed concepts to action: A meta-analysis of the behavioral effects of incidentally presented words. *Psychological Bulletin*, 142(5):472.

[WHO, 2018] WHO (2018). Deafness and hearing loss, Fact Sheet n. 300. `http://www.who.int/mediacentre/factsheets/fs300/en/`. Last accessed on March 2018.

[Wichmann and Hill, 2001] Wichmann, F. A. and Hill, N. J. (2001). The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & psychophysics*, 63(8):1293–1313.

[Won et al., 2007] Won, J. H., Drennan, W. R., and Rubinstein, J. T. (2007). Spectral-ripple resolution correlates with speech reception in noise in cochlear implant users. *Journal of the Association for Research in Otolaryngology*, 8(3):384–392.

[Xia et al., 2012] Xia, R., Li, J., Akagi, M., and Yan, Y. (2012). Evaluation of objective intelligibility prediction measures for noise-reduced signals in mandarin. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4465–4468. IEEE.