

Speech naturalness ratings and perceptual correlates of highly natural and unnatural speech in hypokinetic dysarthria secondary to Parkinson's disease

Marie Klopfenstein

Southern Illinois University at Edwardsville, IL, USA

Abstract

Despite the importance of speech naturalness to treatment outcomes, little research has been done on what constitutes speech naturalness and how to best maximize naturalness in relationship to other treatment goals like intelligibility. This study investigated the speech naturalness ratings of individuals with dysarthria and the associated perceptual correlates of highly natural and unnatural speech. Four speakers with hypokinetic dysarthria secondary to Parkinson's disease were recorded and rated for naturalness by 69 students in Communication Disorders. Students were presented with 436 speech samples and asked to provide speech naturalness ratings on a 1–9 Likert scale. After rating speech samples, subjects listed perceptual cues associated with samples rated most and least natural and weighted each cue on a visual analog scale. The data on naturalness ratings showed that spontaneous speech was rated the least natural on average, while sentences from a short story were rated slightly more natural and individually read sentences were rated the most natural of all of the utterance types. Thirteen themes emerged from the perceptual cues collected. Of the 13 themes, intelligibility was rated significantly more important than other cues in highly natural speech and intelligibility and articulation were rated significantly more important than other cues in highly unnatural speech.

KEYWORDS: HYPOKINETIC DYSARTHRIA; PROSODY; SPEECH NATURALNESS; SPEECH PERCEPTION

* Correspondence address: Email: maklopf@siue.edu

1. Introduction

In the field of speech-language pathology, speech naturalness is a useful measure of how an individual's participation at the society level¹ may be affected by a speech disorder. Depending on the social, and possibly professional, roles the individual wishes to maintain, the clinician will prioritize speech naturalness accordingly when planning intervention. A common emphasis in speech therapy is increasing the intelligibility of speech, keeping with the tradition of functional communicative goals. However, an increase in intelligibility may be accompanied by a decrease in how natural the individual's speech sounds and vice versa (Yorkston, Beukelman, Strand, and Bell, 1999). **Therefore, a common dilemma experienced by clinicians is how to balance intelligibility and naturalness.**

Due to the high level of motor control required for 'normal' or natural speech, clinicians' therapy goals for dysarthria clients are for 'best possible speech' and may not meet the clients' own expectations for their speech after therapy (Yorkston *et al.*, 1999). For the client, **sounding 'natural' may be speaking like she or he did before the onset of dysarthria, even if such a goal is out of reach.** Consequently, speech-language pathologists must evaluate functional goals in respect to clients' goals. For example, a speaker whose profession requires public speaking will be very concerned with maintaining speech naturalness, but will also want to remain intelligible. Furthermore, from studies of how people are perceived based on their 'tone' of voice, it is understandable why clients would want to speak as naturally as possible, as personal and professional interactions can be affected by listener perceptions of how a person sounds when he or she speaks. For example, a study of doctor-patient interactions found that doctors perceived to have a 'dominant' tone of voice were more likely to be sued for malpractice than those judged to express concern or anxiety when talking with patients (Ambady *et al.*, 2002). Also, as illustrated by the work of Miller and colleagues, the effect of a speech disorder such as dysarthria is felt in social interactions beyond the workplace (Miller *et al.*, 2006, 2008). The potential for a disconnect between clinician and client expectations, along with limited ability to provide a prognosis of speech naturalness based on speech characteristics, can lead to lack of client motivation and compliance with therapy.

For that reason, clinicians are paying increasing attention to naturalness in the clinic, but face several limitations with the current state of research on speech naturalness. While there are informal assessment procedures in use for the naturalness of dysarthric speech, no standardized assessment or evidence of what type of speech sample for assessment is most suitable exists. Finally, because there is limited research on the relationship between various speech characteristics and naturalness, current therapeutic approaches that attempt

to maximize both the intelligibility and naturalness of a client's speech must proceed on a trial-and-error basis (Yorkston *et al.*, 1999).

Unfortunately for the speech-language pathologist consulting research on naturalness and speech disorders, many questions one might ask when attempting to balance these goals remain largely unanswered. The concept of speech naturalness is commonly found in assessment and research, yet it is often left undefined or inconsistently defined. This inconsistency poses a problem for researchers and speech-language pathologists attempting to maintain consistency between procedures and assessments. One reason for this lack of consistency is the use of the terms 'bizarreness', 'normalcy', 'acceptability', 'severity', 'proficiency', and even 'intelligibility' interchangeably with the concept of naturalness (Whitehill, 2002). Another difficulty in defining the concept may lie in the phenomenon of easily recognizing and identifying something that is seen or heard, but performing significantly worse when asked to describe the same thing, a tendency known as verbal overshadowing (Dodson, Johnson, and Schooler, 1997; Schooler and Engstler-Schooler, 1990). Judges are very consistent in their assessments of naturalness when using rating scales (Ingham *et al.*, 1985; Onslow *et al.*, 1992), lending support to the notion that naturalness is a perceivable aspect of speech, but a difficult one to put in words.

Another issue to be faced is the need for more research on what aspects of speech relate to perceived naturalness (Linebaugh and Wolfe, 1984; Metz, Schiavetti, and Sacco, 1990; Southwood and Weismer, 1993; Yorkston *et al.*, 1999). Many dysarthrias present with prosodic involvement, regardless of the classification of the dysarthria. A recent definition of naturalness by Yorkston and colleagues (1999) places an emphasis on prosody and there is support for this focus in light of studies that have found decreased naturalness in the absence of segmental-level errors (Ingham and Onslow, 1985; Onslow and Ingham, 1987; Runyan, Hames, and Prosek, 1982; Sacco, Metz, and Schiavetti, 1992). Others have also argued that although naturalness and intelligibility are related dimensions of speech, prosody may serve as a distinguishing feature between the two (Dagenais and Wilson, 2002; Whitehill and Chun, 2002). Therefore, prosodic characteristics of speech represent a reasonable place to begin investigating what features of speech correlate with perceived naturalness. Additionally, a better understanding of how different characteristics of speech relate to naturalness can aid in therapy planning. By establishing treatment priorities, clinicians are directed to therapies that interfere less with naturalness of speech. It may very well be the case that clients would be motivated if naturalness were targeted earlier in therapy, but there is no information on the pros and cons of the timing of such goals.

Despite the potential importance of speech naturalness to functional and social outcomes for individuals with dysarthria, there is little research avail-

able to guide clinicians. Currently, the literature only hints at a relationship between prosody and naturalness and does not address whether other linguistic characteristics are also important in listeners’ perceptions of speech naturalness. Consequently, this study investigates the speech naturalness ratings of individuals with dysarthria and the associated perceptual correlates of highly natural and unnatural speech.

2. Method

2.1. Participants

Four adults (3 males, 1 female) with **hypokinetic dysarthria** secondary to Parkinson’s disease were included in the study. Due to the heterogeneous nature of the dysarthrias (Wenke, Theodoros, and Cornwell, 2010) and the difficulty of finding subjects matched by age, etiology, or length or severity of illness in significant numbers, a case study approach was used. Efforts were made to find participants with a diagnosis of hypokinetic dysarthria because of the prosodic components typical of this dysarthria. Participants were recruited through local Parkinson’s disease support groups and speech-language pathologists’ referrals. All subjects were monolingual English speakers with no other history of nervous system injury or impairment. Subjects 2 and 3 were the only participants to report hearing impairment; both acquired their hearing impairment later in life and wore hearing aids. None of the subjects reported a history of other speech or language disorders, except for subject 4, who had a history of stuttering as a child and adolescent.

Table 1: Demographic data

Subject	Age	Sex	Formal Education	Time Since Diagnosis	Severity of Dysarthria
1	78	M	Bachelor’s degree	5 years	Mild
2	76	F	Master’s degree	4 years	Moderate
3	79	M	Bachelor’s degree	18 years	Mild
4	67	M	Associate’s degree	9 years	Severe

The severity of each subjects’ dysarthria was determined by two experienced speech-language pathologists and the author. Each listened to a short sample of connected speech from each subject and independently assigned an estimate of severity to each speaker. All judges were in agreement on the severity of each speaker’s dysarthria.

2.2. Recording of dysarthric speech samples

All speech samples were collected in a sound-isolated booth. Subjects were recorded using a cardioid condenser microphone (Audio Technica AT2020)

placed at a distance of 30 cm. Samples were directly digitized using a digital audio solid state recorder (Marantz PMD660) at a sampling rate of 44.1 kHz with 16-bit quantization. Subjects read four sets of sentences containing either words from a set of 25 monosyllabic or a set of 25 disyllabic words. The first two sets of sentences contained the words in the carrier sentence, 'No, he said _____. Subjects were asked to read the sentences with emphasis on the last word, as if correcting someone who had misheard a statement they had made. The final two sets of sentences contained the same words used previously, but in 50 short, i.e. containing five to ten words, sentences at a simple reading level. Subjects were given no instructions to emphasize any particular words in these sets. Subjects were also asked to read the children's short story, *The Pokey Little Puppy*. Speakers were instructed to read the story aloud as if they were reading to a child in order to maximize prosodic aspects of their speech. The story was divided up approximately into thirds and the middle third, containing 404 words, was selected for analysis in order to control for the effects of warming up and vocal fatigue. Finally, two samples of spontaneous speech were collected. In the first, subjects were asked to watch a short (6 minute) Pink Panther cartoon, 'Pickled Pink'. Subjects were then asked to recall what occurred in the cartoon to the best of their ability and retell the action, as if telling a friend about what they saw. In the second sample, subjects were recorded in conversation dyads or triads with either the author, or family members and caretakers. Topics of conversation were chosen based on background information gathered at the beginning of data collection. Conversations were recorded for at least five minutes. Sentences were selected for analysis from the middle of the conversation to control for the effects of vocal fatigue.

The speech samples selected for inclusion in this study were taken from the sentences and short paragraphs read by participants and spontaneous speech samples as described previously. Seventeen and nine sentences were included from each set of sentences with monosyllabic and multisyllabic words in carrier sentences, respectively, and 19 and 20 sentences were included from each set of sentences with monosyllabic and multisyllabic words in short sentences, respectively. In total, 65 sentences per speaker were included in the sample to be rated for naturalness, for a total of 260 sentences from all four participants. The same sentences were chosen for each speaker; sentences containing any hesitations longer than one second or reading errors that changed the essential meaning of a sentence for any one speaker were eliminated from the pool of possible sentences to be included for all speakers. This was done in order to reduce the influence of aspects of reading aloud, rather than the effect of PD on speech prosody, on naturalness judgments. Thirty-four sentences were eliminated in this manner from inclusion. Fourteen sentences per speaker, for a total of 56

sentences, were selected for inclusion from the short paragraph reading task in the same manner, i.e. the same sentences for all speakers with no hesitations or errors in reading. Fourteen sentences from the reading task were excluded from the rating task due to hesitations, disfluencies, or reading errors.

Sentences from the two recorded spontaneous speech tasks recorded in Experiment 1 were also included, but due to their nature, these sentences could not be as closely matched between speakers. The length of the spontaneous speech sample for the video recall task varied between speakers, so five sentences per speaker were included for naturalness rating, for 20 sentences total. Sentences that described the same action in the video were chosen as much as possible. The recorded conversation dyads and triads were divided up approximately into thirds for each speaker, in order to obtain a representative sample from the beginning, middle, and end of a conversation. Five sentences were selected for naturalness judgments from each third for each speaker, for a total of 15 sentences per speaker and 60 sentences overall. In total, 396 sentences were selected from the previous speaking tasks for naturalness ratings. Of those sentences, 40 (10%) were repeated in the rating task to test for intra-judge reliability. Selected speech samples were prepared for presentation to listeners and randomized using Praat's ExperimentMFC function as adapted by the author.

2.3. Naturalness rating procedure

Naturalness raters were 69 students enrolled in a BA, MA, or PhD program in the Department of Communicative Disorders at the University of Louisiana at Lafayette who reported having no history of hearing problems. A required sample size of 62 was calculated using a power analysis at an alpha level of 0.05. The listeners included one PhD student, 20 second year Master's level students who had completed coursework in motor speech disorders, 22 first year Master's level students enrolled in a craniofacial anomalies course, and 26 undergraduate students enrolled in a course on articulation and phonological disorders.

Listeners were asked to complete two listening tasks, divided into three activities. The method used for these tasks was adapted from Martin and colleagues (1984) and Coughlin (1998). Each listener completed all activities seated at a computer in a quiet room using headphones adjusted to a comfortable listening level. The first task and activity involved rating each provided speech sample on a 1-9 Likert scale (Martin *et al.*, 1984). The second task involved listening again to samples identified as either highly natural or unnatural by using the extreme ends of the scale and (1) listing perceptual cues identifying what made each sample sound either natural or unnatural, then (2) weighting each perceptual cue given on a provided visual analog scale (Coughlin, 1998).

For the first task, listeners were instructed to run a script that used Praat's Experiment MFC function to provide samples for rating and record the rating given for each item. Listeners were instructed in how to play each sample and provide a rating and afterward completed a practice set of five samples unrelated to the experimental samples in order to familiarize themselves with the procedure.² Each listener was instructed to wear headphones adjusted to his or her most comfortable listening level in order to reduce ambient noise in the room and maximize audio reception of the speech signal. The following instructions, adapted from (Martin *et al.*, 1984), were provided on the computer screen for listeners before beginning the rating task:

I am studying what makes speech sound natural or unnatural. In this experiment, you will hear a number of short speech samples and see the orthographic transcription of each sample on the screen. Your task is to rate the naturalness of each speech sample. If the speech sample sounds highly natural to you, click 1 on the scale. If the sample sounds highly unnatural, click 9 on the scale. If the sample sounds somewhere between highly natural and highly unnatural, click the appropriate number on the scale. Do not hesitate to use the ends of the scale (1 or 9) when appropriate.

'Naturalness' will not be defined for you. Make your rating based on how natural or unnatural the speech sounds to you. You may hear another voice overlap with the speaker's in some samples, but you may disregard this when making your rating. This is not a timed task, so you may proceed at whatever speed is most comfortable for you. You may also replay samples if you wish, up to 3 times per sample. If you make a mistake, you may click the 'oops' button to go back. You may also take short breaks if needed and will be prompted to do so after a certain number of samples. Please remember, however, that it is important to rate all of the speech samples provided. You will begin by rating 5 practice samples to familiarize yourself with the rating procedure.

Click your mouse ONCE when you are ready to begin the practice portion of the experiment.

In order to control for the potentially confounding variable of intelligibility, listeners were provided orthographic transcriptions on the computer screen of each sample that they rated. After listeners completed the practice rating task, they were asked to contact the author if they had any questions or did not understand the task. Once listeners had no questions and understood the task, they were allowed to begin the naturalness rating task. Listeners were prompted to take a short break after every 100 samples. Practice rating tasks involving raters not part of this study indicated that the rating task would take about an hour for most people to complete.

2.4. Procedure for listing and weighting perceptual cues

In this task, each listener was presented with all of the speech samples he or she rated most natural and most unnatural, beginning with those rated highly

natural, and was asked to provide and weight perceptual cues that were factors in his or her ratings on a 100 mm visual analog scale (VAS). This activity took place after listeners were instructed to take a short break upon completion of the first task. To begin the second task, listeners were instructed to run a script on Praat that selected samples to be replayed. The script first looked for all samples rated in the first experiment as 1, or highly natural, to be included in the sample. If no samples were rated as 1, 2 or the next most natural rating on the scale was used and the appropriate samples were included for replay. The same procedure was then followed for samples rated as highly unnatural, starting with a rating of 9 or the next most unnatural rating chosen by the listener.

Before beginning the second task, listeners were presented with the following instructions adapted from Coughlin (1998):

You have completed the first part of your job. For the second part of this job, you will listen to the samples you rated as highly natural followed by the samples that you rated as highly unnatural. You will be able to listen to each group a maximum of three times. After each playback group, think about your recent use of the 1–9 point scale and using the provided paper, list why you rated samples as either highly natural or highly unnatural. You may list as many reasons as you wish. Ten spots are provided for you to list those items, but if you do not need all ten spots, you do not have to use them all. If you need more spaces, you may use an additional paper. If you cannot think of a single word to describe a reason for your rating, please describe what you mean, using a phrase or sentence.

Once you have listed or described as many factors as possible, please indicate how important each factor was for rating along the provided scales. The leftmost portion of the scale is least important and the rightmost portion is most important. Do not hesitate to use the end of the scales. Please place the factor at a point along the scale based on the amount you feel that item influenced your decision. An example is given to you on the provided paper.

When you are ready to begin listening to the first group, click your mouse. When you are done listening to both groups, you may close this window and follow the directions for submitting your results.

A sample response sheet used by listeners for listing and weighting perceptual cues is included in Appendix A. After listing and weighting the perceptual cues associated with highly natural ratings, the process was repeated for the samples rated most unnatural.

2.5. Data analysis

The data collected on perceptual cues related to the speech samples rated most and least natural were also analyzed according to qualitative methods (Braun and Clarke, 2006; Denzin and Lincoln, 2000; Morse and Richards, 2002; Polit

and Beck, 2006). The data were entered into a spreadsheet, with each perceptual cue listed on the rating sheet associated with a rating value from the visual analog scale. In total, 774 listed cues (367 cues for the most natural sounding samples, 407 cues for the least natural sounding samples) were collected from 69 raters. The cues from five raters were excluded from this count because it appeared the raters did not understand the task (e.g. raters gave cues that would normally be associated with unnatural speech, like 'strain', on sheets for most natural cues), which cast doubt upon the validity of the data in those cases. As the cues were analyzed for commonalities between different raters, patterns emerged. Collapsing patterns and creating themes occurred next, producing more refined results. This occurred as common threads between patterns were identified, revealing similarities that supported the emerging themes. In certain cases, cues fell under two different themes, as in one cue provided by a student: 'articulations of words; difficulty of understanding'. Cues like these were assigned to two different themes (in this example, Articulation and Intelligibility) and both assigned the same value according to the visual analog scale. There were 31 cues total that fell into more than one theme in the least natural category and eight in the most natural category.

Twenty pairs of identical speech samples were included in the rating task to test for intra-judge reliability. Reliability was determined using Cronbach's alpha for this investigation. Following George and Mallery (2003), the following guidelines for interpreting the coefficient were used: $\alpha \geq 0.9$ – Excellent, $\alpha \geq 0.8$ – Good, $\alpha \geq 0.7$ – Acceptable, $\alpha \geq 0.6$ – Questionable, $\alpha \geq 0.5$ – Poor, and $\alpha < 0.5$ – Unacceptable. Alpha values were calculated for each rater using the duplicated speech samples; 63 out of 69 raters had alpha values above 0.7, the minimum reliability level selected for this investigation. Because of their unreliable judgments, six raters³ were eliminated from further data analysis.

Krippendorff's *alpha* was chosen as an index for inter-rater agreement and reliability because it operates even with different levels of measurement, including nominal and ordinal data, and it considers the observed frequency of agreement as well as the amount of agreement expected by chance alone. It also can accommodate multiple raters and can calculate a single reliability coefficient, unlike Cohen's *kappa* (Hayes and Krippendorff, 2007). Krippendorff's interval α was 0.4079 for naturalness ratings given to speech samples by the remaining 63 raters, which indicates a low level of agreement between listeners. Inter-rater agreement may be on the low side compared to intra-rater reliability because raters each used the term 'naturalness' differently. Following the methods used in previous studies, no explicit definition of naturalness was given and no examples of highly natural and unnatural speech, as a kind of calibration, were provided to raters.

3. Results

3.1. Naturalness Ratings

In the end, data from 436 speech samples rated by 63 raters, for a total of 27,468 ratings ($M = 4.92$, $SD = 2.623$), were analyzed. Overall, the more frequent rating given to speech samples was 7 and the least frequent rating was 9.

Table 2: Frequency of Ratings Used from the Naturalness Scale

Rating	1	2	3	4	5	6	7	8	9
Frequency	3,581	3,251	2,972	2,592	2,609	3,053	3,595	3,200	2,615
Percent of Total	13.0	11.8	10.8	9.4	9.5	11.1	13.1	11.6	9.5

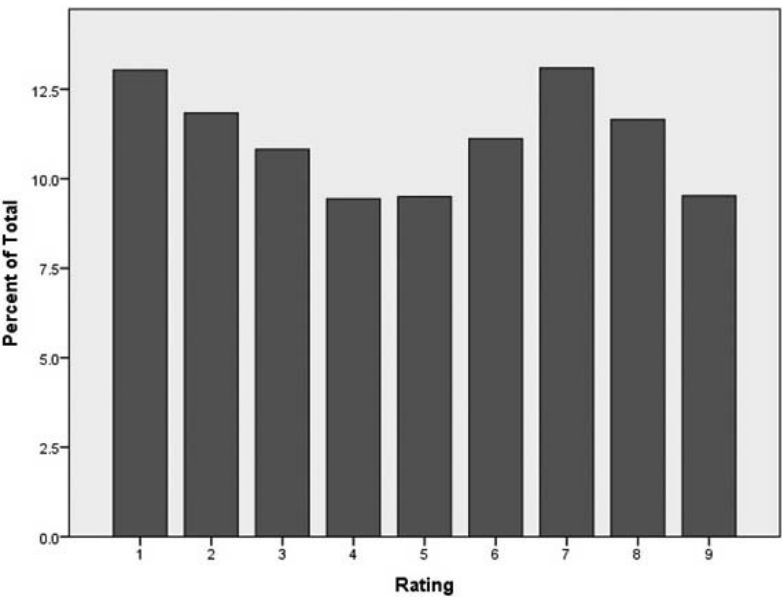


Figure 1: Percentage that each rating contributed to all naturalness ratings.

The mean rating given by all raters to each subject is shown below in Table 3 and Figure 2. Each rater made full use of the naturalness scale for each subject, so all subjects had a minimum rating of 1 and a maximum rating of 9.

Table 3: Descriptive Statistics for Naturalness Ratings Given by Subject

Subject	S1	S2	S3	S4
Mean	3.92	6.23	3.13	6.42
SD	2.245	2.178	2.164	2.200

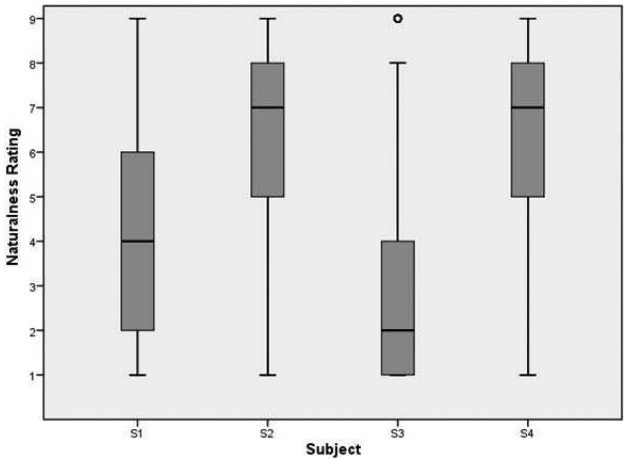


Figure 2: Boxplot of naturalness ratings received by each subject.

The data were also analyzed in order to determine whether the type of speech sample being rated (e.g. read sentences, stories or spontaneous speech) had any interaction with naturalness ratings. The means and standard deviations for each type of speech sample across all subjects are given in Table 4.

Table 4: Descriptive Statistics for Naturalness Ratings by Utterance Type

Utterance Type	Mean	SD	SE of Mean	Min.	Max.
Sentence	4.63	2.558	0.019	1	9
Spontaneous	5.66	2.577	0.033	1	9
Story	5.02	2.737	0.043	1	9

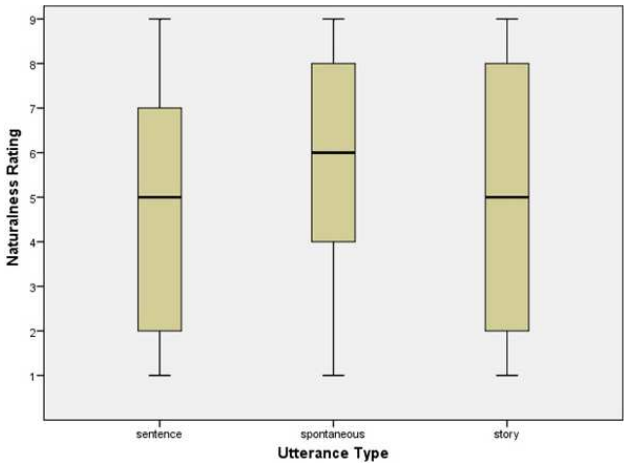


Figure 3: Naturalness ratings for each utterance type across all subjects.

3.2. Perceptual Cues

As described before, 13 themes emerged that covered all of the perceptual cues given by participants in regards to unnatural and natural speech alike. Each of these themes are explained below.

Articulation

Raters discussed perceptions of how subjects produced their speech, without commenting on the understandability of the speech. Cues in this theme included keywords like ‘articulation’, ‘pronunciation’, or ‘enunciation’ in relation to natural speech and ‘slurring’, ‘mumbling’, or ‘muffled’ for unnatural speech. As imprecise articulation is a common characteristic of dysarthric speech, it was not unexpected for raters to mention speech articulation as a perceptual feature of naturalness. These cues were assigned to a theme separately from the intelligibility theme because we cannot know based on the given data whether raters found the speech unintelligible. It is possible for mumbled speech to still be understandable. Therefore only cues that described speech articulation without commenting on intelligibility were included in this theme.

Content Delivery

Some raters mentioned their impressions of whether what they were listening to was read from a script or not, or whether the speaker had confidence when speaking. Other raters made judgments about the speakers, such as whether the speaker was relaxed, believable, or struggling with knowing what to say. While cues like these were infrequent, they diverged enough from other themes for them to be assigned to a theme related to how the speaker delivered meaning or content in their speech. We can speculate on some of the raters’ perceptions, such as the perception of reading versus conversational speech being related to speech rate, but cannot know this for sure without follow up data.

Dialect

Although PD subjects and raters come from the same general geographic region of the US, many cues commented on the perceived dialect of the speakers. Some of these comments included statements that the subject did not have a disorder, but a strong accent, or simply stated the presence or ‘lack’ of accent. Any mention of the terms ‘dialect’ and ‘accent’ in cues were included in this theme. Most raters used these terms with no additional explanation of how this contributed to their perception of naturalness, regardless of whether these cues were in the most or least natural category.

Fluency

Many cues were comments on features that are associated with fluency of speech. These qualities include hesitations, breaks, repetitions, stuttering, pauses, interjections, pauses, and flow. These cues were categorized separately from the content delivery theme because raters commented exclusively on speech qualities associated with fluency in these cues, while other cues related more to perceived speaker personality or prompted speech.

Gender Qualities

Interestingly, some raters explicitly mentioned whether they thought the speaker's voice appropriately matched their gender or not, or they just wrote down 'gender'. These cues were grouped under this theme and not the voice quality theme, as they otherwise made no comment on voice quality or pitch.

Intelligibility

Cues included in this theme explicitly mentioned intelligibility or comprehensibility. Some raters stated that they did not need the provided captions to understand what the speaker was saying instead. While intelligibility is multidimensional and may include aspects of other themes that emerged from other cues, many cues only referred to intelligibility and not to speech features that contributed to the subject's intelligibility.

Loudness

Keywords such as intensity, 'audible', and loudness that appeared in cues were included in this theme. Many raters quantified the loudness level and judged that as contributing to either more or less naturalness. Other raters also mentioned variable loudness as a factor. As raters were instructed to adjust their listening volume to a comfortable level at the beginning using the practice rating task, differences between subjects' loudness should have been apparent.

Nasal Resonance

In these cues listeners either noticed nasality, by which they may have meant hypernasality, or they noted resonance as a cue. Generally, cues that were grouped under this theme were not specific about the kind of resonance perceived.

Other

Cues that did not fit with any other theme and also held no other kind of cohesive relationship with each other were grouped here. As described in the

following sections, these cues made up a small proportion of total cues, being less than 5% of both the natural and unnatural speech cues. Essentially these cues were eliminated from further analysis.

Prosody

Cues with keywords like intonation, monotone voice, pitch variability or instability were robust in the data. Raters also contributed stress and rhythm as perceptual cues for naturalness. When cues strictly conformed to the aforementioned prosodic speech features, they were grouped in this theme.

Speech Rate

Cues that described subjects' rate of speech were included in this theme. Common key words included speed, pace, rate, fast, slow, and hurried. Some raters expanded on their cues by stating that their perception of speech rate as too fast or too slow depended on the subject.

Syntactic Boundaries

Although such cues were uncommon, some raters provided perceptual cues describing subjects' speech in relation to punctuation and commas. These cues were not limited to just one rater. The raters could have been referencing whether subjects paused at expected places, like clause boundaries or not, but without additional clarification or follow up possible this is not certain. Therefore, these cues were grouped into this theme, rather into another theme like prosody.

Voice Qualities

Cues included in this theme included keywords characterizing speakers' voices, like rough, hoarse, breathy, raspy, strained, pitch breaks, whisper, scratchy, gurgly, gravely, creaky, quiver, and diplophonia. Some raters also wrote 'voice quality' as a perceptual cue without going into more detail as well. These descriptions occurred in cues for natural and unnatural speech, depending on whether raters were noting these qualities were present or not.

3.3. Perceptual Correlates of Highly Natural Speech

Some types of perceptual cues were mentioned more frequently by listeners than others. The percentage total for each theme, as well as the descriptive statistics relating to the rated importance of each theme, is given in Table 6.

Table 6: Percentage and Descriptive Statistics for Themes Derived from Cues for Highly Natural Speech

Theme	Mean	SD	Median	N	Min	Max	Range	%
Articulation	73.29	24.232	81.00	38	10	100	90	10.2%
Content Delivery	70.91	20.602	79.00	23	26	97	71	6.2%
Dialect	56.67	7.638	55.00	3	50	65	15	0.8%
Fluency	65.25	23.497	67.00	63	3	100	97	17.0%
Gender Qualities	46.00	30.633	50.00	6	7	79	72	1.6%
Intelligibility	84.42	18.623	93.00	50	10	100	90	13.5%
Loudness	66.28	22.305	72.50	18	26	98	72	4.9%
Nasal Resonance	67.00	32.527	67.00	2	44	90	46	0.5%
Other	66.31	28.389	75.00	16	5	99	94	4.3%
Prosody	71.26	20.821	71.50	62	15	100	85	16.7%
Speech Rate	70.02	21.653	75.00	45	5	100	95	12.1%
Syntactic Boundaries	74.67	20.781	69.50	6	54	100	46	1.6%
Voice Quality	64.03	28.422	71.00	39	5	100	95	10.5%
Total	70.34	23.649	75.00	371	3	100	97	100.0%

While most of each theme's cues made up less than 10% of the sample, the three most frequent themes were all above: fluency (17%), prosody (16.7%), and intelligibility (13.5%).

3.4. Perceptual Correlates of Highly Unnatural Speech

Some types of perceptual cues were mentioned more frequently by listeners than others. The percentage of total for each theme, as well as the descriptive statistics relating to the rated importance of each theme, is given in Table 7.

Table 7: Percentage and Descriptive Statistics for Themes Derived from Cues for Highly Unnatural Speech

Theme	Mean	SD	Median	N	Min	Max	Range	%
Articulation	81.09	18.209	86.00	64	13	100	87	14.7%
Content Delivery	67.00	33.707	80.50	12	12	97	85	2.8%
Dialect	23.33	22.546	25.00	3	0	45	45	0.7%
Fluency	65.99	24.818	69.50	72	9	100	91	16.6%
Gender Qualities	61.62	22.393	65.50	16	7	93	86	3.7%
Intelligibility	85.88	19.276	95.00	49	25	100	75	11.3%
Loudness	58.75	22.530	61.50	24	17	90	73	5.5%
Nasal Resonance	47.00	29.924	53.50	8	7	85	78	1.8%

Other	59.55	24.047	57.50	20	7	99	92	4.6%
Prosody	65.86	22.542	66.50	50	20	100	80	11.5%
Speech Rate	72.11	15.585	71.00	37	33	100	67	8.5%
Syntactic Boundaries	72.80	15.385	71.00	72.80	50	92	42	1.1%
Voice Quality	64.04	22.807	69.00	75	7	100	93	17.2%
Total	69.23	23.884	73.00	435	0	100	100	100.0%

Likewise, most of each unnatural speech theme’s cues made up less than 10% of the entire sample. The three most frequent themes were: voice quality (17.2%), fluency (16.6%), and articulation (14.7%).

4. Discussion

The overall naturalness ratings agree with the severity estimates given by independent judges. The subject with the severe dysarthria, S4, was rated the least natural overall. S2 was judged to have a moderate dysarthria and was rated second least natural after S4. The two subjects with mild forms of dysarthria, were rated most natural, with S3 rated the most natural overall among the four subjects. These results suggest that the severity of disordered speech and naturalness are related dimensions of speech and explains why these two different concepts are often confused in the literature (Whitehill, 2002).

Interestingly, speech that resulted from speakers reading pre-determined material, whether from individual sentences or from a story, was rated more natural than spontaneous speech. Naturalness judges often recognized spontaneous speech as such during the listening task and mentioned that speech that sounded like the speaker was reading aloud sounded less natural during the perceptual cue task. This indicates listeners may not have complete conscious insight into why they have rated a particular speech sample natural or not. This can occur in the results of this study because the collection of perceptual cues for naturalness was disassociated in a way from the naturalness rating task, since perceptual cues were only collected after all speech samples were rated.

4.1. Perceptual Cues for Highly Natural Speech

It is not surprising that intelligibility cues were frequently given by raters for highly natural speech. It is understandable that if one cannot comprehend another person’s speech, that one would consider it unnatural. This factor was recognized and the design of this study attempted to control for it by providing a transcription of all utterances during the rating task. However, it could

be that rather than increasing the intelligibility of utterances, the transcription may have enhanced the perception of raters that the linguistic content of some of the speech samples was difficult to understand. Therefore, intelligibility may have overshadowed the importance of other cues for speech naturalness, leading listeners to mostly think of intelligibility when re-listening to the speech samples they rated highly natural. The rating of intelligibility as more important than all other cues is also possibly a consequence of following previous methods of investigating speech naturalness, in which naturalness is not defined for raters.

One can also look at the frequency that each cue was mentioned by naturalness raters to get another perspective on how important these cues are for hypokinetic dysarthria. In order from most to least frequently mentioned types of cues, the cues are: (1) prosody and fluency (tied); (2) intelligibility; (3) speech rate; (4) voice quality; (5) articulation; (6) content delivery; (7) loudness; (8) other; (9) gender qualities and syntactic boundaries (tied); (10) dialect; and (11) nasal resonance. Further research is required to determine whether these cues would continue to be mentioned most frequently in a naturalness study and if any of the cues would be rated significantly more important than others if raters were explicitly instructed to regard intelligibility as a separate dimension of speech than naturalness.

4.2. Perceptual Cues for Highly Unnatural Speech

In order from most to least frequently mentioned types of cues, the cues are: (1) voice quality; (2) fluency; (3) articulation; (4) prosody; (5) intelligibility; (6) speech rate; (7) loudness; (8) other; (9) gender qualities; (10) content delivery; (11) nasal resonance; (12) syntactic boundaries; and (13) dialect. The same possibility of intelligibility overshadowing other cues, as mentioned previously, may apply here, although intelligibility was not as frequent a cue as for natural speech. Articulation was also rated more significant than several other cues, but was listed only slightly more frequently than prosody, a cue rated significantly less important. This suggests that students often noticed the prosodic aspects of the speech of individuals with hypokinetic dysarthria, but did not feel these factors were as important in rating speech samples highly unnatural. It could be that prosody plays a larger role in rating speech highly natural than less natural. However, it is also possible that articulation overshadowed the importance of prosody or other cues. Since articulation errors may have significant impact of the intelligibility of any given utterance, the same caution in teasing apart naturalness and intelligibility applies here as well. Further investigation of the frequency and rated importance of perceptual cues when intelligibility and naturalness are separately defined would shed more light on this issue.

5. Conclusions

One implication of the overall rated importance of perceptual cues for naturalness is the importance of investigating naturalness as a distinct dimension of speech from intelligibility. While there is definitely a relationship between the two speech dimensions, unless one can investigate other aspects of speech that contribute to naturalness, one might as well use intelligibility as a measure of the social impact of disordered speech. The drawback to that approach is that there are clearly aspects of speech beyond intelligibility that contribute to how disordered speech is perceived. For example, individuals who stutter are still perceived as sounding different from individuals who do not stutter, even when sound or syllable repetitions are not present in the former's speech (Ingham, Gow, and Costello, 1985).

Another implication is that prosodic aspects of speech were not necessarily listed most frequently as a perceptual cue in both highly natural and highly unnatural speech. This raises the question of whether prosody should be considered to be as important for speech naturalness in individuals with dysarthria as the literature suggests. Before one can conclude that prosody is not as vital to naturalness as previously thought, however, two considerations should be made. First, would prosody still not be rated significantly important for the perception (or lack thereof) of naturalness if intelligibility was not a consideration? And second, would the same perceptual cues be listed in a similar study and given similar importance if more experienced, or even more naïve, judges were used?

The results of this study are suggestive as to the role of articulation, intelligibility, and other speech features for speech naturalness, but given that naturalness is a very complex speech dimension, further research is needed to clarify the role each suggested perceptual cue from this study plays towards its perception.

One limitation of this study is the sample size of individuals with hypokinetic dysarthria and characteristics of the population included in the sample. In order to make predictions about the characteristics of speech naturalness and prosody for the speech of all individuals with hypokinetic dysarthria secondary to PD in general, this study would ideally have a much larger and more random sample of the population. The participants of this study were not randomly selected because they were all members of a support group for individuals with PD in East Texas. Additionally, speakers with the same level of severity of hypokinetic dysarthria would be included in order to reduce potential intersubject variability. Because of these limitations, one cannot definitively say that the results of this study hold true for the speech of the general population. However, the sample size of this study allowed for a larger and deeper data set to be collected. One of the aims of this study was to collect

substantial data that would point to promising avenues for future, expanded research projects, in which the focus of the research needs to be narrowed in relation to the number of participants.

Limitations aside, this study has some important strengths, such as the perceptual cues that were reported by listeners that highlight all of the distinguishing characteristics of hypokinetic dysarthria and therefore likely represent a complete picture of what components make up judgments of naturalness for this particular speech disorder. However, the examination of how important listeners felt each cue was to their assessment of naturalness revealed some inconsistencies and suggested that naturalness may be a concept that is more intuitive and difficult to describe or define, even for students of speech-language pathology. Also, the influence of intelligibility as a perceptual cue for naturalness, even to the point of possibly overshadowing other perceptual cues, was revealed during the analysis of these cues.

One of the implications of these results is that while the data presented here confirm a relationship between prosody and naturalness – at the very least within speakers included in the study – there are also indications that listeners are not necessarily consciously aware of the characteristics of speech that lead to their naturalness judgments. Although the raters in this study were speech-language pathology students at different levels of training, one must consider the possibility that students could benefit from more experience listening to and describing disordered prosody, especially given its importance to speech disorders like dysarthria. In order to diagnose and treat such disorders, one must understand and be able to express what one hears in disordered prosody. Speech-language pathologists naturally gain this ability as they gain professional experience, but it could only benefit students to expose them to these skills as soon and as frequently as possible.

The current study has only scratched the surface in terms of investigating perceptual cues of naturalness in disordered speech. There is a myriad of possibilities for future research, starting with investigations in other kinds of disordered speech such as that in hearing impaired speakers, those with fluency disorders, other types of dysarthric speech, and so on. Given the different characteristics of these populations, different relationships between prosodic characteristics and naturalness would be expected to emerge, furthering our understanding of how humans judge speech naturalness.

Numerous possibilities for future research also exist in which variations of the methods employed in this study and previous studies can be tweaked to see how study outcomes are affected. These include comparisons of perceptual naturalness cues between different speech disorders or even between judges with varying levels of sophistication, i.e. advanced graduate level students and professionals with more than ten years of experience with a particular popu-

lation. Future studies could also attempt to expand on the results presented here by experimentally manipulating samples in terms of length, frequency, or other acoustic characteristics in order to determine whether certain thresholds for the naturalness of these variables exist. Additionally, prospective work in this area may need to use techniques like calibration in order to ensure that raters are consistent in their use of the naturalness scale and to try to walk the line between giving raters absolutely no definition of naturalness and unduly influencing raters' impression of what naturalness should mean.

Notes

1. According to the 2001 version of the International Classification of Functioning, Disability, and Health (ICF).
2. Practice samples were sentences taken from the Rainbow Passage as recorded by the author.
3. Upon closer investigation, these raters did not appear to be completely unreliable in their ratings, as many duplicate samples were given either the same rating or a rating within one or two points of the original. It appears that these raters changed their use of the scale partway through the experiment or for some other unknown reason were not consistent in their use of the scale, since some samples had widely divergent ratings, which explains the low alpha values.

About the author

Marie Klopfenstein is Assistant Professor in the Department of Applied Health at Southern Illinois University Edwardsville. Her doctoral research was undertaken at the University of Louisiana at Lafayette, and her main research interests are in motor speech disorders and phonetics.

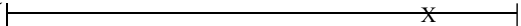
References


- Ambady, N., Laplante, D., Nguyen, T., Rosenthal, R., Chaumeton, N., and Levinson, W. (2002). Surgeons' tone of voice: a clue to malpractice history. *Surgey*, 132 (1): 5–9. <http://dx.doi.org/10.1067/msy.2002.124733>
- Braun, V., and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3: 77–101. <http://dx.doi.org/10.1191/1478088706qp063oa>
- Coughlin, S. S. (1998). *Speech naturalness of normal speaking children and adolescents*. (Unpublished doctoral dissertation). Michigan State University, East Lansing, MI.
- Dagenais, P. A., and Wilson, A. F. (2002). Acceptability and intelligibility of moderately dysarthric speech by four types of listener. In F. Windsor, M. L. Kelly, and N. Hewlett (Eds), *Investigations in Clinical Phonetics and Linguistics*, 363–372. Mahwah, NJ: Lawrence Erlbaum Associates.
- Denzin, N. K., and Lincoln, Y. S. (Eds) (2000). *Handbook of Qualitative Research* (2nd ed.). Thousand Oaks, CA: Sage.


- Dodson, C. S., Johnson, M. K., and Schooler, J. W. (1997). The verbal overshadowing effect: Why descriptions impair face recognition. *Memory and Cognition*, 25 (2): 129–139. <http://dx.doi.org/10.3758/BF03201107>
- George, D., and Mallery, P. (2003). *SPSS for Windows Step by Step: A Simple Guide and Reference. 11.0 Update* (4th ed.). Boston, MA: Allyn and Bacon.
- Hayes, A. F., and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1 (1): 77–89. <http://dx.doi.org/10.1080/19312450709336664>
- Ingham, R. J., and Onslow, M. (1985). Measurement and modification of speech naturalness during stuttering therapy. *Journal of Speech and Hearing Disorders*, 50 (3): 261–281. <http://dx.doi.org/10.1044/jshd.5003.261>
- Ingham, R. J., Gow, M., and Costello, J. M. (1985). Stuttering and speech naturalness: Some additional data. *Journal of Speech and Hearing Disorders*, 50 (2): 217–219. <http://dx.doi.org/10.1044/jshd.5002.217>
- Jaccard, J., and Wan, C. K. (1996). *LISREL Approaches to Interaction Effects in Multiple Regression*. Thousand Oaks, CA: Sage.
- Keppel, G. (1992). *Design and Analysis – A Researcher's Handbook* (3rd ed.). Engelwood Cliffs, NJ: Prentice Hall.
- Martin, R. R., Haroldson, S. K., and Triden, K. A. (1984). Stuttering and speech naturalness. *Journal of Speech and Hearing Disorders*, 49 (1): 53–58. <http://dx.doi.org/10.1044/jshd.4901.53>
- Miller, N., Noble, E., Jones, D., and Burn, D. (2006). Life with communication changes in Parkinson's disease. *Age and Ageing*, 35 (3): 235–239. <http://dx.doi.org/10.1093/ageing/afj053>
- Miller, N., Noble, E., Jones, D., Allcock, L., and Burn, D. J. (2008). How do I sound to me? Perceived changes in communication in Parkinson's disease. *Clinical Rehabilitation*, 22 (1): 14–22. <http://dx.doi.org/10.1177/0269215507079096>
- Morse, J. M., and Richards, L. (2002). *Readme First for a User's Guide to Qualitative Methods*. Thousand Oaks, CA: Sage.
- Onslow, M., and Ingham, R. J. (1987). Speech quality measurement and the management of stuttering. *Journal of Speech and Hearing Disorders*, 52 (1): 2–17. <http://dx.doi.org/10.1044/jshd.5201.02>
- Polit, D.F., and Beck, C. T. (2006). *Essentials of Nursing Research: Methods, Appraisal and Utilization* (6th ed.). Philadelphia, PA: Lippincott.
- Runyan, C. M., Hames, P. E., and Prosek, R. A. (1982). A perceptual comparison between paired stimulus and single stimulus methods of presentation of the fluent utterances of stutterers. *Journal of Fluency Disorders*, 7 (1): 71–77. [http://dx.doi.org/10.1016/0094-730X\(82\)90040-7](http://dx.doi.org/10.1016/0094-730X(82)90040-7)
- Sacco, P. R., Metz, D. E., and Schiavetti, N. (1992). Speech naturalness of nonstutterers and treated stutterers: Acoustical correlates. Presentation to the Annual Meeting of the American Speech–Language–Hearing Association, San Antonio, TX.


- Schooler, J. W., and Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, 22 (1): 36–71. [http://dx.doi.org/10.1016/0010-0285\(90\)90003-M](http://dx.doi.org/10.1016/0010-0285(90)90003-M)
- Wenke, R. J., Theodoros, D., and Cornwell, P. (2010). Effectiveness of Lee Silverman Voice Treatment (LSVT) on hypernasality in non-progressive dysarthria: The need for further research. *International Journal of Language and Communication Disorders*, 45 (1): 31–46. <http://dx.doi.org/10.3109/13682820802638618>
- Whitehill, T. L. (2002). Assessing intelligibility in speakers with cleft palate: A critical review of the literature. *Cleft Palate-Craniofacial Journal*, 39 (1): 50–58. [http://dx.doi.org/10.1597/1545-1569\(2002\)039<0050:AIISWC>2.0.CO;2](http://dx.doi.org/10.1597/1545-1569(2002)039<0050:AIISWC>2.0.CO;2)
- Whitehill, T. L., and Chun, J. C. (2002). Intelligibility and acceptability in speakers with cleft palate. In F. Windsor, M. L. Kelly, and N. Hewlett (Eds), *Investigations in Clinical Phonetics and Linguistics*, 405–415. Mahwah, NJ: Lawrence Erlbaum Associates.
- Yorkston, K. M., Beukelman, D. R., Strand, E. A., and Bell, K. R. (1999). *Management of Motor Speech Disorders in Children and Adults* (2nd ed.), 618. Austin, TX: Pro-Ed.


Appendix: Cues for Most Natural Sounding Samples


1. Knit 
 Least enjoy Most enjoy


1. 
 Least important Most important


2. 
 Least important Most important


3. 
 Least important Most important


4. 
 Least important Most important


5. 
 Least important Most important

6. 
 Least important Most important

7. 
 Least important Most important

8. 
 Least important Most important

9. 
 Least important Most important

10. 
 Least important Most important

