IDENTIFYING ACOUSTIC CAUSES OF SPEAKER-DEPENDENT VARIATION IN SLOWED SPEECH INTELLIGIBILITY: A HYBRIDIZATION APPROACH

Frits van Brenk[1], Kris Tjaden[1], & Alexander Kain[2]

[1]Communicative Sciences and Disorders, University at Buffalo
[2]Computer Science & Electrical Engineering, Oregon Health & Science University
brenk@buffalo.edu; tjaden@buffalo.edu; kaina@ohsu.edu

ABSTRACT

This study sought to identify acoustic variables explaining rate-related variation in intelligibility for speakers with dysarthria for sentences produced at habitual and a slower than normal rate.

Four speakers with dysarthria due to Multiple Sclerosis (MS) produced the same 25 Harvard sentences at habitual and slow rates. Speakers were selected from a larger corpus based on intelligibility characteristics. Two speakers demonstrated improved intelligibility and two speakers demonstrated reduced intelligibility when rate was slowed. A speech-analysis resynthesis paradigm termed hybridization was used in which segmental (i. e. short-term spectral) and suprasegmental variables (i. e. sentence-level fundamental frequency, energy characteristics, duration) were manipulated individually or in combination. Six hybridized sentence types were studied.

On-line crowd-sourced orthographic transcription was used to quantify intelligibility for hybridized stimuli and the original habitual and slow productions. The results indicated that combined changes in short-term spectrum and duration affect intelligibility variation associated with a slowed rate.

Keywords: dysarthria; resynthesis; hybridization; slow rate; intelligibility

1. INTRODUCTION

Speech motor control issues associated with dysarthria may manifest in all speech subsystems namely respiration, phonation, resonance, articulation, and prosody, potentially contributing to reduced intelligibility [1]. Understanding specific speech production characteristics or combinations of characteristics underlying intelligibility variation is an important goal of dysarthria research. In addition to advancing conceptual understanding of intelligibility, this line of inquiry potentially helps shaping targeted and patient-tailored treatments that address specific and predefined speech production variables contributing to reduced intelligibility.

Kain and colleagues [3] developed an analysis-resynthesis approach termed *hybridization* in which one or more acoustic parameters of a set of speech stimuli, such as the fundamental frequency (F0) trajectory or the energy trajectory are manipulated while holding other parameters constant. This approach allows for statements concerning the causal role of specific speech production measures to intelligibility, as reflected in the acoustic signal. While their study used speech materials from one neurotypical speaker, hybridization may help with further untangling the acoustic basis of reduced intelligibility in dysarthria. Thus far, hybridization has been applied to only a single published study investigating dysarthria, where it was used to investigate segmental and suprasegmental variables explaining intelligibility variation in speech produced in conversational and clear speaking styles by two speakers with mild hypokinetic dysarthria secondary to Parkinson's disease [7]. Findings indicated that the increased intelligibility of clear speech resulted primarily from adjustments to the short-term spectrum (i. e. segmental articulation) and the energy trajectory, demonstrating that hybridization may be used to identify acoustic variables that cause (as opposed to correlate with) intelligibility variation in mild dysarthria.

The current study extended the use of hybridization to investigate the acoustic basis of intelligibility variation during slowed speech produced by speakers with MS. Rate reduction is a popular behavioral management technique in dysarthria, as it may be associated with improved intelligibility [12]. Slower than normal rate may facilitate achievement of more canonical or extreme vocal tract configurations that are distinctive from each other. However, not all speakers with dysarthria exhibit improved speech intelligibility when slowing rate [2, 8, 10]. It has been observed that a slower-than-normal articulation rate is associated with prosodic adjustments including reduced phrase-level fundamental frequency range that may negatively affect intelligibility [9]. Overall, little is known about rate-related changes in the acoustic signal that explain differences in intelligibility.

The goal of this study was to identify acoustic variables explaining rate-related variations in intelligibility in speakers with MS, potentially improving the scientific evidence base for dysarthria treatment. To investigate whether similar acoustic variables account for both reduced and increased intelligibility when slowing rate, this study focused on two speakers with MS for whom intelligibility was improved and two speakers with MS for whom intelligibility was reduced when rate was slowed.

2. METHOD

2.1. Participants

2.1.1. Speakers

Speakers were females diagnosed with MS, native speakers of American English, had achieved at least a high school diploma, did not use a hearing aid, and reported no other history of neurologic disease. The participants were selected from a speaker database based on their transcription intelligibility scores, summarized and analyzed in previous work from our lab [5]. Two speakers (MSF03; 44 y/o and MSF17; 51 y/o) demonstrated *lower* intelligibility during slowed speech (MSF03: 79.6% and MSF17: 61.2%) compared to habitual speech (MSF03: 96.0% and MSF17: 78.0%), and formed the group ‘Low’. Two speakers (MSF04; 51 y/o and MSF20; 53 y/o) showed *higher* intelligibility during speech produced at a slow rate (MSF04: 79.6% and MSF20: 73.2%) compared to habitual (MSF04: 61.2% and MSF20: 60.8%), and formed the group ‘High’.

2.1.2. Listeners

A total of 507 adults (334 females, 170 males, 3 not specified), 18 to 80 years of age (M=35.9, SD=11.4) judged intelligibility. Participants were recruited using the crowdsourcing website Amazon Mechanical Turk (MTurk; mturk.com). Previous studies have demonstrated the feasibility of using MTurk sourced transcription scores to quantify intelligibility in dysarthria [4, 11]. Participants were allowed to participate after they fulfilled a number of prerequisites, including an approval rate of greater than or equal to 99%, and confirmed status of U. S. resident. All listeners were self-reported native speakers of American English, living in the United States, and without a history of speech, language, or hearing problems.

2.2. Production tasks

Speakers read 25 Harvard Psychoacoustic Sentences [5] in habitual and slow speaking conditions. Each sentence ranged in length from seven to nine words and contained five key words. For detailed information on data collection, see [7]. For each speaker, a random sample of the same 10 sentences produced in the habitual and slow conditions was of interest. For the slow condition, speakers were instructed to produce the sentences half as fast as their habitual rate, were encouraged to stretch out words rather than solely insert pauses, and were asked to produce each sentence on a single breath.

2.3. Stimuli preparation

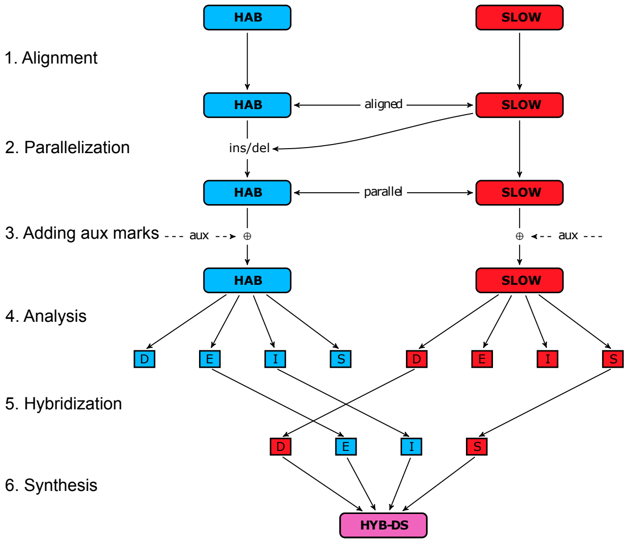
2.3.1. Hybridization Algorithm and Speech Resynthesis

The hybridization technique used in this study is a residual-excited linear predictive coding waveform resynthesis of each habitual sentence while selectively imposing the energy envelope, F0 envelope, segment durations, or short-term spectra from the same sentence produced at slow rate by the same speaker. The procedure is summarized below and displayed in Fig. 1; a comprehensive description of the hybridization process may be found in [7].

1. Alignment: segment boundaries and individual glottal pulses were identified using Praat, and aligned across sentence types to compensate for possible differences in phonemic content.
2. Parallelization: the waveform of the habitual rate sentences was modified by implementing phoneme deletions or insertions relative to the slow-rate waveforms using a time-domain crossfade technique to avoid discontinuities. As a result, the phoneme sequence of the resulting hybridized waveform was the same as that of the slow-rate waveform. Hybridization did not take place during inserted phoneme segments.
3. Auxiliary Marks: a speech analysis was carried out on the parallelized waveforms, consisting of first determining the location of auxiliary marks used for subsequent prosodic modification.
4. Analysis: extracting short-term spectra, energy trajectories, F0 trajectories, and segment durations.
5. Hybridization: acoustic characteristics of slow rate speech were combined with complementary acoustic characteristics of habitual speech to form hybrid acoustic parameters.
6. Synthesis: creation of six hybrid speech waveforms: intonation (I); energy (E); duration (D); prosody, defined as the combination of intonation, energy, and duration (IED); short-term spectrum (S); and the combination of duration and short-term spectrum (DS).

A total of 320 stimuli were created: 4 speakers × 10 sentences × 8 sentence variants (Habitual, Slow, and six hybrid versions listed above).

Figure 1: Block diagram summarizing hybridization process. HAB = habitual rate; SLOW = slow rate; ins = insert; del = delete; aux = auxiliary; D = duration; E = energy; I = intonation; S = short-term spectrum; HYB-DS = hybrid of duration and short-term spectra.



2.3.2. *Mixing with multitalker babble*

Speakers with MS had mostly preserved intelligibility, based on Sentence Intelligibility Task scores, reported as group means in [6]. Thus to prevent ceiling effects, speech materials were mixed with multitalker babble. Sentences were first intensity-normalized and then mixed with 20-talker babble, at a signal-to-noise ratio (SNR) of 0 dB. This SNR minimized floor and ceiling effects, as determined by pilot testing.

2.4. Perceptual task

After consenting to participate, listeners were instructed to transcribe a series of stimuli while using headphones and working in a quiet environment. Sentences were presented one at a time. Following each presentation, listeners were asked to transcribe the sentence as accurately as possible, and to guess if unsure. Listeners were allowed to listen multiple times to each sentence. Stimuli were presented in a blocked and randomized fashion, ensuring that no identical sentence text was presented twice, irrespective of speakers or sentence variant. After completing the sentence transcription, participants were asked to complete a demographic questionnaire. The experiment took between five to nine minutes with associated remunerations between $0.80 and $1.20. Listeners were allowed to participate only once. For each of the 320 stimuli, a minimum of 20 valid transcription scores were obtained.

2.5. Transcription analysis

The mean percentage of correctly transcribed key words was calculated for each stimulus, with each sentence containing five key words. For quality control purposes, listeners failing to respond to more than 20% of the presented stimuli, or who correctly transcribed less than four out of five key words in a control sentence without noise, were excluded.

The percentage of correctly transcribed key words was the primary dependent variable for comparing groups, namely speakers with lower intelligibility in slow vs. habitual ('Low') and speakers with the reverse effect ('High'). Data were analyzed using an analysis of variance with group and sentence variant as fixed factors. Further effects were explored with Bonferroni-corrected post-hoc tests. A significance level of .05 was used for all hypothesis testing.

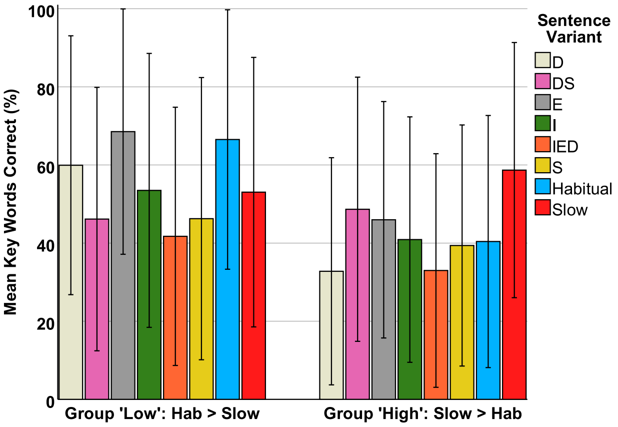
3. RESULTS

3.1. Intelligibility measures

Fig. 2 reports mean percent correct scores pooled across sentences and listeners. The statistical analysis indicated a significant main effect of Group: $F(1, 8409) = 283.9$, $p<.001$. There was a main effect of Sentence Variant: $F(7, 8409) = 46.3$, $p<.001$, as well as a significant Group by Sentence Variant interaction effect: $F(7, 8409) = 38.2$, $p<.001$. Post-hoc comparisons of transcription scores were performed within the Low and High groups. The results for the habitual and slow rate sentences confirmed the expected pattern in the Low group, with significantly higher transcription scores in the habitual compared to the slow version (mean difference 6.9%, p=.019). Similarly, the High group had higher transcription scores in the slow compared to the habitual condition (mean difference 18.3%, $p<.001$). These results confirmed the data reported in [8], and indicated that, as a baseline, the two groups showed meaningful differences between the slow and habitual rates.

Of particular interest were the hybrid versions that may explain acoustically driven intelligibility differences between the slow and habitual conditions. For the High group, these were hybrids that scored higher than the habitual variant (thus becoming more slow-like with respect to intelligibility). Post-hoc results comparing the habitual variant to each hybrid variant associated with improved intelligibility were as follows. There was a non-significant increase in intelligibility in the E hybrid (mean difference 5.6%, p=.153, and a significant increase in intelligibility in the DS hybrid (mean difference 8.3%, p=.002), indicating that a combination of spectral and durational characteristics contributed to increased intelligibility for the slow rate in the

Figure 2: Percent correct scores across sentences and listeners, reported by Group and compared by Sentence Variant. Error bars indicate ± 1SD



High group. For the Low group, the hybrids of interest were those that yielded a significant increase in intelligibility relative to the slow condition (thus becoming more habitual-like). Post-hoc testing identified two hybrids: the D hybrid (mean difference 6.9%, *p*=.019) and the E hybrid (mean difference 15.5%, *p*<.001). Thus, for the Low group, acoustic characteristics other than strictly energy, and to a lesser extent duration, contributed to decreased intelligibility when rate was slowed.

3.2. Acoustic measures

A variety of acoustic measures were obtained to verify that the four speakers were slowing rate when instructed to do so, and to characterize additional acoustic changes in segmental articulation by means of vowel space area (VSA), vocal intensity in terms of sound pressure level (SPL) and fundamental frequency (F0). Descriptive statistics for the acoustic measures are reported in Table 1 in the form of averages for the 10 Harvard sentences. When qualitatively comparing habitual versus slow speaking conditions, speakers from all groups showed an increase in tense and lax vowel space area, in mean F0 and F0 range, and a decrease in articulation rate. Furthermore, all speakers except MSF03 showed an increase in mean SPL and a decrease in SPL variation. Overall, the acoustic measures showed consistent changes from habitual to slow rate conditions, and these changes were found to be fairly comparable across speakers of both groups.

4. DISCUSSION

This study investigated the contribution of segmental and suprasegmental acoustic variables to intelligibility variation in speakers with MS. Specifically, we examined acoustic features that cause speaker-dependent variation in intelligibility when slowing speaking rate.

Table 1: Acoustic measures for both habitual and slow rate sentence production. Top panel: Low group, bottom panel: High group.

Speaker	Rate	VSA	VSA	SPL	SPL	F0	F0	Art
		(tense)	(lax)	M	SD	M	Range	rate
MSF03	Hab	428310	111513	68.2	7.78	147	119	3.83
	Slow	463160	130720	63.8	8.78	149	135	2.55
MSF17	Hab	495966	105924	70.9	9.45	174	173	3.41
	Slow	620985	144807	75.1	9.30	211	222	1.16
MSF04	Hab	488949	104956	71.6	9.31	160	179	2.99
	Slow	537483	125677	74.7	9.98	162	184	2.01
MSF20	Hab	167872	54696	73.3	7.41	147	143	4.24
	Slow	206070	80392	73.9	8.77	173	148	2.89

The results showed that, even though all speakers demonstrated significant rate changes, duration in itself seems to play a minor role in intelligibility decline, as the associated D hybrid did not (or very minimally) contribute to intelligibility changes, irrespective of group. Furthermore, for the High group, a combination of spectral and durational hybridization adjustments (DS) contributed to the increase in intelligibility for slowed speech. These findings were not completely mirrored in the Low group (but neither contradicted), where a number of parameters seemed to have contributed to the intelligibility decline for slowed speech, including those associated with the DS hybrid. Moreover, whilst the DS hybrid was the most prominent contributor to increased intelligibility in slow speech, these findings were not captured in the acoustic measures, where vowel space areas consistently increased when rate was slowed, again irrespective of group. This indicates that the spectral cues mediating changes in intelligibility go beyond vowel characteristics, and may include spectral properties across a longer domain, including consonantal information.

Hybridization is a powerful technique to systematically manipulate and subsequently identify acoustic variables explaining intelligibility variation in dysarthria. While the current study identified both segmental and suprasegmental properties as sources of increased intelligibility during slowed speech, the current results indicate that additional studies are needed to identify factors explaining intelligibility variation associated with rate manipulation.

5. ACKNOWLEDGMENTS

This study was funded by an NIH grant (NIH-NIDCD R01DC004689) awarded to the second author. The authors gratefully thank all speaker and listener participants, and Sara Silverman for her help with data analysis.

6. REFERENCES

[1] Duffy, J. R. 2013. *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*. Mosby, St. Louis 3 edition.

[2] Fletcher, A. R., McAuliffe, M. J., Lansford, K. L., Sinex, D. G., Liss, J. M. 2017. Predicting intelligibility gains in individuals with dysarthria from baseline speech features. *J. Speech Lang. Hear. Res.* 60(11), 3043–3057.

[3] Kain, A., Amano-Kusumoto, A., Hosom, J.-P. 2008. Hybridizing conversational and clear speech to determine the degree of contribution of acoustic features to intelligibility. *J. Acoust. Soc. Am.* 124(4), 2308–2319.

[4] Lansford, K. L., Borrie, S. A., Bystricky, L. 2016. Use of crowdsourcing to assess the ecological validity of perceptual-training paradigms in dysarthria. *Am. J. Speech-Lang. Path.* 25(2), 233–239.

[5] Stipancic, K. L., Tjaden, K., Wilding, G. 2016. Comparison of intelligibility measures for adults with parkinson's disease, adults with multiple sclerosis, and healthy controls. *J. Speech Lang. Hear. Res.* 59(2), 230–238.

[6] Sussman, J. E., Tjaden, K. 2012. Perceptual Measures of Speech From Individuals With Parkinson's Disease and Multiple Sclerosis: Intelligibility and Beyond. *J. Speech Lang. Hear. Res.* 55(4), 1208–1219.

[7] Tjaden, K., Kain, A., Lam, J. 2014. Hybridizing conversational and clear speech to investigate the source of increased intelligibility in speakers with parkinson's disease. *J. Speech Lang. Hear. Res.* 57(4), 1191–1205.

[8] Tjaden, K., Sussman, J. E., Wilding, G. E. 2014. Impact of clear, loud, and slow speech on scaled intelligibility and speech severity in parkinson's disease and multiple sclerosis. *J. Speech Lang. Hear. Res.* 57(3), 779–792.

[9] Tjaden, K., Wilding, G. 2011. The impact of rate reduction and increased loudness on fundamental frequency characteristics in dysarthria. *Folia Phoniatrica et Logopaedica* 63(4), 178–186.

[10] Tjaden, K., Wilding, G. E. 2004. Rate and Loudness Manipulations in Dysarthria: Acoustic and Perceptual Findings. *J. Speech Lang. Hear. Res.* 47(4), 766–783.

[11] Yoho, S. E., Borrie, S. A. 2018. Combining degradations: The effect of background noise on intelligibility of disordered speech. *J. Acoust. Soc. Am.* 143(1), 281–286.

[12] Yorkston, K. M., Hakel, M., Beukelman, D. R., Fager, S. 2007. Evidence for effectiveness of treatment of loudness, rate, or prosody in dysarthria: A systematic review. *J. Med. Speech-Lang. Path.* 15(2), xi–xi.