AJSLP

Research Note

The Effects of Modulating Fundamental Frequency and Speech Rate on the Intelligibility, Communication Efficiency, and Perceived Naturalness of Synthetic Speech

Jennifer M. Vojtech,[a,b] Jacob P. Noordzij Jr.,[a,b] Gabriel J. Cler,[b,c] and Cara E. Stepp[a,b,c,d]

Purpose: This study investigated how modulating fundamental frequency (f0) and speech rate differentially impact the naturalness, intelligibility, and communication efficiency of synthetic speech.
Method: Sixteen sentences of varying prosodic content were developed via a speech synthesizer. The f0 contour and speech rate of these sentences were altered to produce 4 stimulus sets: (a) normal rate with a fixed f0 level, (b) slow rate with a fixed f0 level, (c) normal rate with prosodically natural f0 variation, and (d) normal rate with prosodically unnatural f0 variation. Sixteen listeners provided orthographic transcriptions and judgments of naturalness for these stimuli.
Results: Sentences with f0 variation were rated as more natural than those with a fixed f0 level. Conversely, sentences with a fixed f0 level demonstrated higher intelligibility than those with f0 variation. Speech rate did not affect the intelligibility of stimuli with a fixed f0 level. Communication efficiency was highest for sentences produced at a normal rate and a fixed f0 level.
Conclusions: Sentence-level f0 variation increased naturalness ratings of synthesized speech, whether the variation was prosodically natural or not. However, these f0 variations reduced intelligibility. There is evidence of a trade-off in naturalness and intelligibility of synthesized speech, which may impact future speech synthesis designs.
Supplemental Material: https://doi.org/10.23641/asha.8847833

Functional speech is described as using speech to convey needs, wants, feelings, or preferences in a way that others can successfully understand (American Speech-Language-Hearing Association, 2015). It is a critical factor in maintaining employment, pursuing education, establishing relationships, and participating in social activities (L. J. Garcia, Laroche, & Barrette, 2002; Hegde & Freed, 2011; Lúcio, Perilo, Vicente, & Friche, 2013). Many augmentative and alternative communication (AAC) devices incorporate speech synthesis methods to enable individuals with limited functional speech capabilities (e.g., as a result of stroke, traumatic brain injury, spinal cord injury, amyotrophic lateral sclerosis, or chronic Guillain–Barré syndrome) to communicate with others. Yet, despite enabling speech capabilities for these individuals, synthesized speech often fails to incorporate the prosodic cues we find in natural speech, such as using voice quality, pitch contour, or timing to convey emotion, mood, or personality (Drager, Reichle, & Pinkoski, 2010; Evitts & Searl, 2006; Fucci, Reynolds, Bettagere, & Gonzales, 1995; Kangas & Allen, 1990; McCall, Marková, Murphy, Moodie, & Collins, 1997). Thus, AAC users cannot manipulate the synthesized speech to convey an emotional state, irony, or emphasis, and listeners must often derive meaning through only the words

[a]Department of Biomedical Engineering, Boston University, MA
[b]Department of Speech, Language, and Hearing Sciences, Boston University, MA
[c]Graduate Program for Neuroscience–Computational Neuroscience, Boston University, MA
[d]Department of Otolaryngology–Head and Neck Surgery, Boston University School of Medicine, MA
Correspondence to Jennifer M. Vojtech: jmvo@bu.edu

Editor-in-Chief: Kristie Spencer
Editor: Julie Wambaugh
Received March 16, 2018
Revision received September 28, 2018
Accepted February 4, 2019
https://doi.org/10.1044/2019_AJSLP-MSC18-18-0052
Publisher Note: This article is part of the Special Issue: Selected Papers From the 2018 Conference on Motor Speech—Clinical Science and Innovations.

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

in the message. This leads to increased effort for both the speaker and the listener: For example, an AAC user might need to use more words to inform the listener of an urgent situation, rather than relying on tone and speech rate to convey the urgency. Similarly, without the prosodic cues found in natural speech, the listener may require a higher cognitive processing load to accurately decode the message (Evitts & Searl, 2006; McCall et al., 1997).

The development of an AAC interface that incorporates prosody may be an important step toward improving the quality of life of these individuals. Allowing users to directly control prosody may lead to increased system complexity and cognitive load for the user. As a result, most work into synthesized speech prosody involves automatically generating prosodic contours within text-to-speech (TTS) systems (e.g., Dutoit, Pagel, Pierret, Bataille, & Van der Vreken, 1996; Reddy & Rao, 2016; Rendel, Fernandez, Hoory, & Ramabhadran, 2016; Scordilis & Gowdy, 1989). A variety of algorithms are used to automatically generate prosody, generally relying on linguistic rules or inferring phrase structure directly from the text input. Another method under development is voice conversion, in which the user's linguistic content could be mapped onto target spectral and prosodic features of a given natural voice (Toda et al., 2016; Wester, Wu, & Yamagishi, 2016; Zhao, Kuruvilla-Dugdale, & Song, 2018). Despite these technical advances, the extent to which different, basic prosodic cues differentially affect the reception of synthesized speech has not yet been adequately studied. In the current study, we investigate the extent to which basic suprasegmental prosodic cues impact a listener's reception of synthesized speech. Results will inform the implementation of effective suprasegmental prosodic control in AAC devices.

Naturalness and Intelligibility

Two global descriptors of speech have been employed to capture the effects of various prosodic cues on speech: naturalness and intelligibility. *Speech naturalness* can be described as "how speech conforms to the listener's standards of rate, rhythm, intonation, and stress patterning and if it conforms to the syntactic structure of the utterance being produced" (Yorkston, Beukelman, Strand, & Bell, 2010, p. 288). By integrating perceptual cues from respiratory, phonatory, and articulatory systems, naturalness allows listeners to focus on the meaning of the message (Ratcliff, Coughlin, & Lehman, 2002). Distinct from naturalness, *speech intelligibility* refers to the degree to which the speaker's utterance is understood by a listener (Hustad, Beukelman, & Yorkston, 1998; Lindblom, 1990; Tjaden, Kain, & Lam, 2014; Yorkston, Strand, & Kennedy, 1996); intelligibility allows listeners to focus on recognizing the words in a message. Speech intelligibility in standard listening and speaking environments is used as a measure that is reflective of the speaker rather than the performance of the listener (Hustad et al., 1998).

Naturalness and intelligibility have previously been implemented as tools for (a) evaluating the acceptability of synthesized speech as an output from an AAC device (see Pampoulou, 2018), (b) examining the effects of various speaker and listener factors on the resulting speech output (e.g., Jones, Berry, & Stevens, 2007; Ratcliff et al., 2002), (c) comparing synthesized speech outputs between AAC devices (e.g., Crabtree, Mirenda, & Beukelman, 1990; Hustad et al., 1998), and (d) comparing the reception of synthesized speech to that of typical speech (e.g., Mirenda & Beukelman, 1987) or dysarthric speech (Drager, Hustad, & Gable, 2004). Work in these areas has shown that the naturalness and intelligibility of synthetic speech are each impacted by speech rate and fundamental frequency (f0); however, it is unclear how these factors differentially impact naturalness and intelligibility.

Speech Rate

Speech rate is commonly defined as words or phonemes produced per minute, including pauses (Tsao, Weismer, & Iqbal, 2006). Previous studies examining speech rate found that decreasing the rate of natural speech (typical: Aihong, Chundan, & Jingjing, 2014; dysarthric: Yorkston & Beukelman, 1981b) and synthetic speech (Syrdal et al., 2012; Venkatagiri, 1991) presentations each led to increases in intelligibility. Conversely, increases in rate with fewer and shorter pauses led to improvements in perceived naturalness of typical speech (Yorkston, Hammen, Beukelman, & Traynor, 1990) and synthetic speech (Ratcliff et al., 2002). Yet, it must be noted there are two methods for modulating speech rate that may affect the interpretability of these results. One method is to statically increase or decrease rate, such that syllables or phonemes are all stretched or compressed by the same amount, respectively (Aihong et al., 2014; Ratcliff et al., 2002; Syrdal et al., 2012; Venkatagiri, 1991; Yorkston et al., 1990). The second method is to mimic the speech rate changes produced in natural speech, in which pauses and vowels are lengthened whereas consonants stay relatively constant (Yorkston & Beukelman, 1981b; Yorkston et al., 1990). In a related study, Tjaden, Sussman, and Wilding (2014) found no significant improvements in intelligibility when speakers were asked to naturally reduce their articulatory rate, wherein the speed of articulatory gestures is computed without accounting for pause frequency (Tsao et al., 2006). Overall, the results of these studies suggest a complex relationship with regard to global speech timing between the naturalness and intelligibility of synthesized speech.

Fundamental Frequency Contours

Sentence-level measures of f0 have also been implicated in influencing the naturalness and intelligibility of speech. Although mean f0 level is not an influential factor on perceived naturalness of synthesized speech (Ratcliff et al., 2002), Meltzner and Hillman (2005) demonstrated that sentence-level modulations to f0 did affect the perceived naturalness of atypical (i.e., nonnatural) speech. Specifically, the authors examined naturalness ratings of electrolaryngeal (EL) speech that was modulated using a combination of enhancements, which included increasing low-frequency energy, reducing noise, and introducing f0 variation to mimic prosodically

natural pitch intonation. It was observed that EL speech stimuli with any type of modulation that included prosodically natural variations in f0 were rated as sounding the most natural in comparison to typical speech. However, further work must be done to determine how sentence-level modulations to f0 impact the perceived naturalness of synthesized speech.

Previous work also examined how f0 contours affect intelligibility. Tjaden and Wilding (2011) observed decreases in f0 variation relative to typical speakers when speakers with dysarthria (i.e., impoverished, natural speech) were instructed to reduce their articulatory rate. The authors speculated that there may be a relationship between reduced articulatory rate and f0 variation that could detrimentally affect measures of intelligibility. Additionally, studies have shown that flattening natural f0 variations to produce monopitch speech reduces the intelligibility of both typical and dysarthric speech (Bunton, Kent, Kent, & Duffy, 2001; Laures & Weismer, 1999; Watson & Schlauch, 2008). Conversely, Tjaden, Kain, and Lam (2014) found that enhancing sentence-level f0 variation of habitual speech did not produce any meaningful improvement in the intelligibility of resynthesized speech. The authors reported that this finding was unexpected but may be a consequence of participants' confounding intelligibility with the unnaturalness of the resynthesized speech. In terms of synthesized speech, mean f0 level has not been shown to be a large or consistently influential factor (Venkatagiri, 1991) on intelligibility, and the effects of f0 variation on intelligibility are inconsistent. As a result, it is unclear how sentence-level variation in f0 affects the intelligibility of synthesized speech.

Intelligibility and Naturalness Trade-Offs

Klopfenstein (2016) describes an ongoing struggle with balancing naturalness and intelligibility: Naturalness is linked to social communication, whereas intelligibility is often linked to efficacious communication (Anand & Stepp, 2015; Klopfenstein, 2016; Yorkston et al., 2010). An increase in one measure may be accompanied by a decrease in the other, and vice versa (Nusbaum, Francis, & Henly, 1997). Patel, Connaghan, and Campellone (2013) observed this relationship in dysarthric speakers, wherein slowed speech rates improved intelligibility while inadvertently minimizing prosodic contrasts such as f0, duration, and intensity. These findings may be a result of listeners weighting different elements of the speech signal when evaluating intelligibility and perceived naturalness. Specifically, intelligibility is highly dependent on the quality of the acoustic signal (Lindblom, 1990; Miller, 2013). On the other hand, perceived naturalness is based on suprasegmental features (e.g., intonation pattern, syllable stress, timing, and word juncture; Lindblom, 1990; Yorkston et al., 1990) extracted from the acoustic signal, in addition to verbal and nonverbal cues (e.g., semantics and gestures; Miller, 2013). Thus, relying solely on one of these measures may not provide a complete indication of overall speech function. Additionally, it is unclear how introducing prosody to synthesized speech differentially impacts intelligibility and perceived naturalness.

Communication Efficiency Ratio

In addition to naturalness and intelligibility, another relevant measure to assess speech function may be the communication efficiency ratio (Miller, 2013; Yorkston & Beukelman, 1981a). Communication efficiency evaluates the effectiveness of a conveyed message as the rate of intelligible speech per minute. As a result, communication efficiency depends not only on the words of the message but also on an additional cue of speaking rate. For instance, slowing down the overall rate of synthetic speech may be a good strategy to improve intelligibility, but the decrease in speed may counteract these improvements in terms of efficiency. Examining speech intelligibility, naturalness, and listener communication efficiency may allow for a more comprehensive assessment of speech than using one or two of these measures alone.

Current Investigation

This study evaluated the extent to which adding variations in f0 and changes in speech rate impacts the naturalness, intelligibility, and communication efficiency of synthesized speech. This study is a necessary step in determining which simple characteristics of prosody are most effective in improving the reception of synthesized speech. Here, f0 was varied at the sentence level using prosodically "natural" and "unnatural" modulations to evaluate how type of modulation affects the reception of synthetic speech. Results will inform future development of simple, automated prosodic control in AAC systems. For instance, an f0 contour could be automatically generated using general variations in f0 without the need for linguistic rules or inferring phrase structure. Normal and reduced speech rates were also compared to examine how synthetic speech reception was impacted; these results will further inform speech synthesis development for both social and functional communication. As such, the following hypotheses are proposed:

1. Reducing the rate of synthesized speech will improve intelligibility but will reduce communication efficiency and perceived naturalness.
2. Introducing sentence-level variations in f0 will enhance perceived naturalness of synthesized speech but will not significantly impact intelligibility or communication efficiency.

Method

Listeners

Sixteen listeners aged 18–29 years (seven women, nine men; M = 21.5 years, SD = 2.9 years) provided orthographic transcriptions of speech and ratings of speech naturalness. The orthographic transcriptions were used to estimate intelligibility and listener communication efficiency, and ratings of naturalness were directly extracted to estimate the perceived naturalness of speech. All participants were healthy adults who reported no history of voice, speech,

language, or hearing disorders. All participants passed a bilateral, pure-tone hearing screening at 25 dB HL at 125, 250, 500, 1000, 2000, 4000, and 8000 Hz. All participants were native speakers of American English; were naive to speech, language, and hearing sciences; and had no prior experience with TTS AAC devices.

Speech Stimuli

Sixteen sentences (see Appendix) were synthesized using the MBROLA 2.00 program (Dutoit et al., 1996). MBROLA is an algorithm that uses diphone concatenation to assemble a string of phonemes into a fluid TTS output. The resulting TTS output is of constant intensity with each phoneme produced according to user-specified durations and pitch patterns. This speech synthesizer was chosen due to its accessibility and versatility: MBROLA is an open-source system that is compatible with a variety of operating systems and contains 74 unique voices across 36 languages that are available for TTS output. Due to its open-source nature and widespread use in the area of speech synthesis (e.g., Chabchoub & Cherif, 2011; Chandra & Akila, 2012; Gibbon & Bachan, 2008; Moreton, 2008; Pierre-Yves, 2003; Schröder & Trouvain, 2003), results can be immediately extrapolated to AAC development. MBROLA is available as a speech synthesizer with associated voice databases and takes sequences of phonemes with target durations and f0 values as input. Thus, it can be combined with other software, such as the open-source software eSpeak, to act as a complete TTS system (Panoiu, Rat, & Panoiu, 2016). In this example, eSpeak would provide spelling-to-phoneme translations and prosodic information, which MBROLA would use to generate the desired speech sounds. In the current study, we investigated how basic modulations to f0 and speech rate affect the reception of synthesized speech, as could be generated automatically by an AAC system using any type of target selection, including direct phoneme selection, customized (nondictionary) words, or words in any variety of languages. This is in contrast to combining MBROLA with a prosodic filter that makes use of linguistic rules or an inferred phrase structure to generate prosodic contours. In other words, we sought to examine how direct, simple modulations (i.e., without the use of linguistic rule-sets) to suprasegmental prosodic cues impact the intelligibility, naturalness, and communication efficiency of the TTS output.

We employed the female "Us1" MBROLA voice/language database to produce sentences in American English at a sampling rate of 16 kHz and a fixed f0 of 180 Hz; the sampling rate and f0 parameters used in this study are defaults of the voice/language database. Each of the 16 sentences produced using MBROLA was prosodically unique in terms of sentence length, utterance form (i.e., statement, question, or command), and semantic predictability in order to model a range of utterances that may be heard in daily life. As an example, the sentence "How are you?" is short in sentence length, is of question form, and is highly semantically predictable; in contrast, the sentence "This is my pomegranate smoothie, not his" is longer in sentence length, is of statement form, and is less semantically predictable. Four groups of sentences were created by altering different aspects of the 16 MBROLA-synthesized sentences, described in detail below. In summary, modulations to speech rate and variations in f0 were introduced as follows:

- Group NF: synthesized speech produced at a normal rate and a fixed f0 level;
- Group SF: synthesized speech produced at a slow rate and a fixed f0 level;
- Group NN: synthesized speech produced at a normal rate with variations in f0 to mimic prosodically natural speech; and
- Group NU: synthesized speech produced at a normal rate with variations in f0 to mimic prosodically unnatural speech.

In order to investigate how altering speech rate and sentence-level f0 variation affect the reception of synthesized speech, we included sentences produced at a slow speech rate to examine whether a simple reduction in rate could improve intelligibility. Additionally, we included sentences with prosodically natural and unnatural f0 contours to determine if introducing sentence-level f0 variation could enhance speech naturalness.

Sentence Groups With a Fixed f0 Level: NF and SF

Two sentence groups were produced as direct outputs from the MBROLA program. The sentences of this first set were generated at constant rate of 95 ms/phoneme with no specified pitch pattern; the resulting sentences had a flat f0 contour at 180 Hz and were termed *NF* for "normal" rate and "fixed" f0. Within the second sentence group, speech rate was purposefully slowed: Each phoneme was empirically chosen to have a constant duration of 142.5 ms (1.5 × the duration of NF phonemes) so that the resulting stimuli were not exceptionally slow but were still perceptibly different from those at the normal speech rate of 95 ms/phoneme (Nejime & Moore, 1998; Yorkston et al., 1990). This set was termed *SF* for "slow" rate and "fixed" f0. The NF and SF stimuli were then processed in Praat (Boersma, 2001) to verify a flat f0 contour at 180 Hz.

Sentence Groups With Variations in f0: NN and NU

Two additional groups were synthesized using the MBROLA program and subsequently modified using Praat (Boersma, 2001). The first of these sets, termed *NN* for "normal" rate and prosodically "natural" variations in f0, was developed by overlaying the f0 contours of the productions of a native speaker of American English onto the flat f0 contours of the NF stimuli using the open-source Vocal Toolkit (Corretge, 2012). The native speaker was a 24-year-old woman with no prior history of speech, language, or hearing disorders. The speaker was instructed to produce each sentence three consecutive times using her typical rhythm, loudness, and pitch. A single repetition of each

sentence that contained no pauses or misarticulations was selected for further processing.

The speaker productions were then modified in Praat to have a mean f0 of 180 Hz so as to match that of the synthesized MBROLA sentences. Thereafter, the f0 contours of the speaker productions were copied onto the NF stimuli using the Vocal Toolkit to produce sentences with a modulated f0 contour mimicking prosodically natural speech. Each sentence was manually examined to ensure that the f0 contour of the speaker was properly copied onto the synthesized sentences. Specifically, the Vocal Toolkit "pitch contour overlay" function failed to operate on areas with no visible glottal pulses; these areas were manually manipulated to match the f0 contour of the speaker. Manual manipulation of the f0 contour in Praat was necessary in approximately 25% of stimuli. Resulting samples were then processed in Praat by using the Vocal Toolkit "pitch contour smoothing" function at a level of 25% in order to optimize correspondence between the suprasegmental features of the speaker and that of the synthesized sentences. The median f0 of each sample was then shifted to 180 Hz.

A final sentence group was produced by adapting the NF set using the Vocal Toolkit in Praat. Here, the inverted f0 contours of the native speaker were overlaid onto the flat f0 contours of each respective NF sentence to simulate "random" intonation. This method of simulating random intonation was chosen instead of truly randomizing the f0 contours in order to maintain sample statistics. The median f0 of each sentence was shifted to 180 Hz to match that of the other stimulus sets. Prosodically unnatural intonation was used in this experiment to investigate how random intonations, such as those that could be implemented automatically in speech synthesis, would affect the reception of the resulting synthesized speech. This sentence group was termed *NU* or "normal" rate and prosodically "unnatural" variations in f0.

Stimulus Processing

Four sentence groups were produced, each containing a modification to the same 16 sentences (see Appendix) with a mean f0 of 180 Hz, for a total of 64 stimuli. The sentence "Surprisingly, no other coffee place in town has free Internet" is available in Supplemental Materials S1–S4 in .wav format for each group (i.e., NF [S1], SF [S2], NN [S3], and NU [S4]). Speech-shaped noise was added for judgments of intelligibility to simulate everyday speech in competing noise environments (Anand & Stepp, 2015; Laures & Weismer, 1999; Miller, 2013). A signal-to-noise ratio of +3.5 dB was chosen in order to minimize the potential for ceiling effects to occur in listener orthographic transcriptions, as determined by pilot testing. No noise was added to speech samples used for judgments of naturalness.

Experimental Overview

Listeners completed two separate tasks within the same session, presented in the following order: (a) orthographic transcriptions and (b) naturalness ratings. The order in which these tasks were presented was kept consistent to minimize ceiling effects that may have otherwise occurred if the listeners heard the speech samples without noise—as in the naturalness task—prior to orthographically transcribing them. Because the level of experience with and exposure to synthesized speech have been shown to influence word recognition ability (Hustad et al., 1998; Venkatagiri, 1994), no training regime was provided to the naive listeners before completing the two tasks. All stimuli were presented to listeners using headphones (Sennheiser HD280 PRO) in a quiet room. The computer volume level was set to play audio samples at an average level of 80 dB SPL. The sound pressure level was calibrated using a sound-level meter with headphone coupler (Type 2250 Hand-Held Analyzer with Type 4947 ½-in. Pressure Field Microphone, Bruel & Kjaer, Inc.). Listeners were allowed to listen to each audio sample a maximum of twice per task.

Orthographic transcriptions were collected through a custom-built MATLAB GUI. Listeners transcribed five stimuli from each of the four sets for a total of 20 sentences, with one production per set repeated at the end for intralistener reliability measures. Stimuli were counterbalanced so that no listener heard the same sentence more than once during the transcription task. Stimuli were also pseudorandomized within and between sets for each listener so that each of the 64 stimuli was rated by four different listeners, as is approximately consistent with prior intelligibility studies (Cannito et al., 2012; Fontan, Tardieu, Gaillard, Woisard, & Ruiz, 2015; Lagerberg, Johnels, Hartelius, & Persson, 2015; Stipancic, Tjaden, & Wilding, 2016; Tjaden, Richards, Kuo, Wilding, & Sussman, 2013).

Next, listeners were familiarized with a description of speech naturalness, which remained visible on a secondary desktop screen throughout the naturalness session. *Speech naturalness* was defined as "how speech conforms to the listener's standards of rate, rhythm, intonation, and stress patterning and if it conforms to the syntactic structure of the utterance being produced" (Yorkston et al., 2010). Additionally, naturalness was described as "NOT the degree to which speech can be understood or comprehended" to attempt to control for the potential confound of intelligibility. All participants were allowed to inquire about the meaning of any of the words contained in the two definitions on the secondary desktop screen; in such cases, a standard dictionary definition was presented. Naturalness ratings were then elicited through a custom-built MATLAB GUI that presented a total of 76 sentences, including all stimuli (16 sentences × 4 sentence groups; three sentences from each condition repeated at the end for reliability). Because all sentences were presented to each listener, no counterbalancing was required as with the intelligibility task. Listeners were asked to rate each stimulus on a 100-mm visual analog scale (VAS), in which 0 represented *completely unnatural* and 100 represented *completely natural*. The VAS was chosen to evaluate naturalness because speech naturalness is considered to be metathetic, with a substitutive and qualitative perceptual continuum, such that equal-appearing interval

scale, VAS, and direct magnitude estimation methods are each valid methods to elicit listener judgments of naturalness (Eadie & Doyle, 2002; Metz, Schiavetti, & Sacco, 1990). Ratings were elicited for all sentences from all 16 listeners, which is approximately consistent with or greater than prior studies (Anand & Stepp, 2015; Eadie & Doyle, 2004; Meltzner & Hillman, 2005; Ratcliff et al., 2002; Yorkston et al., 1990).

Data Analysis

Performance Metrics

Naturalness was directly extracted from the mean listener VAS scores, whereas intelligibility and communication efficiency were calculated from the orthographic transcriptions. Intelligibility was evaluated using a custom MATLAB script as the number of correctly identified words in the transcription divided by the total number of words in the original sentence. All transcriptions were manually inspected for misspellings and homophones, which were counted as correct. The communication efficiency ratio, which was originally developed by Yorkston and Beukelman (1981a) to evaluate speech performance in dysarthric speakers, was adapted to reflect the rate of intelligible synthesized speech (intelligible words per minute) from the orthographic transcriptions when normalized by the mean rate of intelligible natural speech produced by a group of normal speakers (190 intelligible words/min). When overlaying the f0 contours of the native speaker onto the synthesized samples, the phonemic duration of the synthesized speech samples was preserved. Because of this, the mean rate of intelligible speech was set to the standard rate for a group of healthy speakers rather than to that of the native speaker; this allows for comparison with other literature. The final data set consisted of 256 ratings for intelligibility and communication efficiency (64 stimuli × 4 ratings for each stimulus) and 1,024 ratings for naturalness (64 stimuli × 16 ratings for each stimulus).

Statistical Analysis

Intralistener reliability was evaluated for speech naturalness by computing the Pearson product–moment correlation coefficients for each of 16 listeners using the naturalness ratings of the repeated 20% of stimuli. Interlistener reliability was assessed for speech naturalness using a two-way random intraclass correlation (ICC) to evaluate the consistency of agreement. A repeated-measures analysis of variance (ANOVA) test was performed on the speech naturalness ratings to evaluate the overall differences in mean naturalness judgments between sentence groups. Participant was included as a random factor with sentence group, sentence length (in number of words), and the interaction between sentence group and sentence length as fixed factors. An α level of .05 was used for significance testing. Effect sizes were calculated using a squared partial curvilinear correlation (η_p^2). A post hoc analysis was conducted using a Tukey simultaneous test to examine differences in mean naturalness ratings between the four sentence sets. This post hoc analysis was selected in order to evaluate the effects of f0 and speech rate on perceived speech naturalness.

Intralistener reliability was evaluated for the orthographic transcriptions by computing Pearson product–moment correlation coefficients for the number of words correctly transcribed in the original and repeated 20% of stimuli. Following Neel (2009), interlistener reliability was assessed separately for the orthographic transcriptions of each of the four sentence sets because listeners judged different sentences in each group. A two-way mixed-effects model was employed to evaluate the consistency of ratings. Two one-way ANOVAs were then performed to evaluate the effect of sentence group, sentence length, and the interaction between these two variables on outcome measures of intelligibility and communication efficiency, with participant as a random factor. An α level of .05 was used for significance testing. Effect sizes were calculated using a squared partial curvilinear correlation. Post hoc Tukey simultaneous comparison tests were conducted between sentence sets in order to investigate the impact of f0 and speech rate on intelligibility and communication efficiency.

Results

Speech Naturalness

Intralistener reliability of speech naturalness ratings was completed on 16 participants; listeners with reliability lower than .50 were removed from further analysis. The average intralistener reliability across the remaining 14 listeners was calculated as $r = .74$ ($SD = .13$, range: .67–.80). Interlistener reliability for the 14 listeners was calculated as ICC = .89 (95% CI [.85, .93]).

Model 1 of Table 1 displays the summary for the repeated-measures ANOVA test performed on speech naturalness ratings for 14 listeners. The model for speech naturalness explained approximately 51% (48% adjusted) of the variance of the data.

A statistical power analysis revealed that we could sufficiently detect effects as small as $\eta_p^2 = .005$ at the .05 significance level (two-tailed) with 80% power while using a repeated-measures ANOVA design. Sentence group showed a large effect size on speech naturalness ($\eta_p^2 = .27$), whereas sentence length showed a medium–large effect ($\eta_p^2 = .14$; Witte & Witte, 2010). The interaction between sentence group and sentence length was not significant. Figure 1a shows that, on average, prosodically natural sentences were rated as the most natural ($M = 55.8\%$), with slow-rate sentences rated as the least natural ($M = 18.3\%$). Post hoc pairwise analyses revealed that the mean naturalness ratings of all sentence groups were statistically different from each other ($p_{adj} < .05$).

Speech Intelligibility and Communication Efficiency

Average intralistener reliability for orthographic transcriptions was calculated for 16 participants, yielding a mean of $r = .95$ ($SD = .07$, range: .8–1.0). All original transcriptions were retained for further processing. The

Table 1. Results of analysis of variance tests performed for speech naturalness (Model 1), intelligibility (Model 2), and communication efficiency (Model 3).

Model							
Number	Response	Type	Factor	*df*	η_p^2	*F*	*p*
1	Naturalness	Repeated measures	Participant	15	.29	26.3	< .001
			Sentence group	3	.27	105	< .001
			Sentence length	8	.14	16.7	< .001
			Sentence Group × Sentence Length	24		0.85	.669
2	Intelligibility	Mixed effects	Participant	15		1.54	.094
			Sentence group	3	.32	31.6	< .001
			Sentence length	8		1.97	.051
			Sentence Group × Sentence Length	24		1.45	.087
3	Communication efficiency	Mixed effects	Participant	15		1.57	.085
			Sentence group	3	.31	30.9	< .001
			Sentence length	8	.39	16.3	< .001
			Sentence Group × Sentence Length	24	.20	2.10	.003

interlistener reliability was measured for each sentence group, yielding an average of ICC = .48 (*SD* = .12, range: .31–.59).

Model 2 of Table 1 shows the mixed-effects ANOVA results for speech intelligibility for 16 listeners. The model for intelligibility explained approximately 49% (37% adjusted) of the variance of the data, with sentence group producing a large significant effect (η_p^2 = .31). Sentence length and the interaction between sentence group and sentence length were not significant. Post hoc pairwise analyses revealed that the mean intelligibility ratings of sentences produced at a fixed f0 level (i.e., NF, SF) were statistically different from those of sentences produced with variation in f0 (i.e., NN, NU; $p_{adj} < .05$). Unlike findings for speech naturalness, fixed-f0 sentences were found to be, on average, more intelligible than variable-f0 sentences. Specifically, NF sentences were the most intelligible (*M* = 73.0%), whereas NU sentences were the least intelligible (*M* = 24.9%; see Figure 1b).

Model 3 of Table 1 shows the mixed-effects ANOVA results for communication efficiency for 16 listeners. The model for communication efficiency explained approximately 61% (51% adjusted) of the variance of the data, for which sentence group and sentence length each produced a large effect (η_p^2 = .31 and .39, respectively). The interaction between sentence group and sentence length was significant in the model for communication efficiency, producing a medium–large effect size (η_p^2 = .20). Post hoc pairwise analyses revealed that (a) the mean communication efficiency ratios of sentences produced at a normal rate and a fixed f0 level were statistically different from those of all other sentence groups ($p_{adj} < .05$) and (b) mean communication efficiency ratios were not statistically different between the sentence group produced with variation in f0. On average, sentences produced at a normal rate and a fixed f0 level resulted in the highest communication efficiency ratios (*M* = 55.0%; see Figure 1c). Of note, a statistical power analysis revealed that, with power set at 0.8 and $\alpha = .05$ (two-tailed), our sample size of 16 was sufficient to detect only large effects (η_p^2 = .33) when using this mixed-effects ANOVA design.

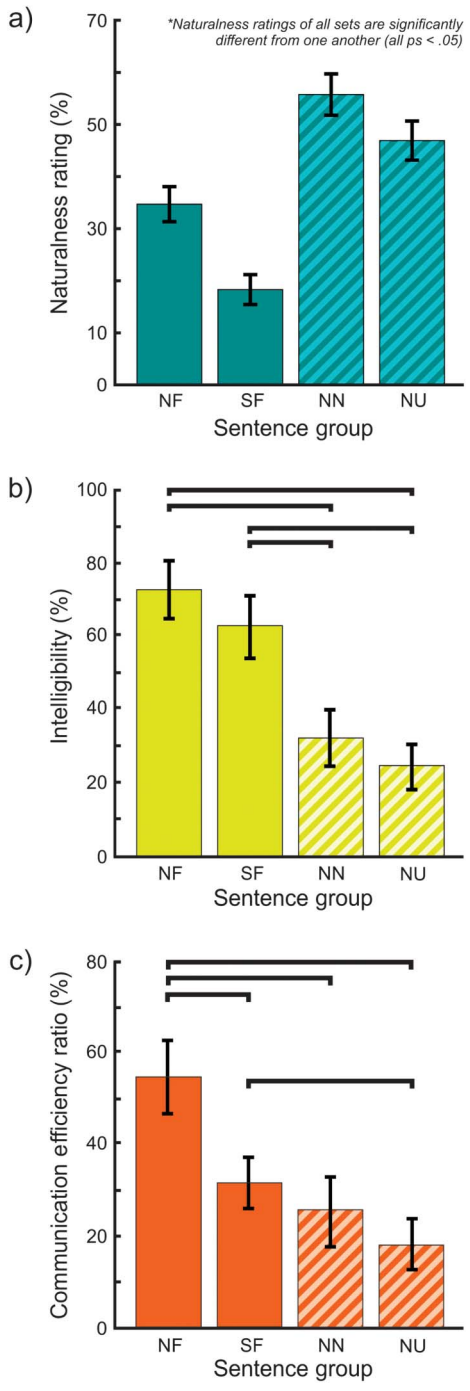
Discussion

The purpose of this study was to assess the extent to which modulating the prosodic cues of speech rate and f0 differentially affects the social and functional reception of synthesized speech. Four sets of synthesized sentences were examined: (a) normal speech rate and fixed f0 level, (b) slow speech rate and fixed f0, (c) normal speech rate with variations in f0 to mimic prosodically natural speech, and (d) normal speech rate with variations in f0 to mimic prosodically unnatural speech. We elicited judgments of naturalness and orthographic transcriptions of speech from 16 listeners. Three outcome measures were examined: perceived naturalness, intelligibility, and communication efficiency.

Perceived Naturalness

We hypothesized that sentences produced at a slower rate would be perceived as significantly less natural than those produced at a normal rate. Indeed, our results show that sentences produced at a slower rate were rated as the least natural of the four sentence groups. We also hypothesized that introducing sentence-level variations in f0 would enhance the perceived naturalness of synthesized speech, regardless of whether the variations were prosodically natural or not. Our results showed that sentences with prosodically natural variations in f0 were found to be, on average, perceived as the most natural. These findings are consistent with those from Meltzner and Hillman (2005), wherein prosodically natural variations in the f0 contour led to improvements in the perceived naturalness of EL speech. Most notable, however, is that both sentence groups with variations in f0 were rated as more natural than those without any f0 modulations. Overall, our findings suggest that sentence-level f0 variation is a perceptually important factor on the naturalness of synthesized speech, regardless of whether the variation is prosodically natural or not. These results can be used to inform the development of AAC devices to enhance social communication: An f0 contour could be automatically generated using general variations in f0 instead

Figure 1. Mean (a) naturalness ratings, (b) intelligibility scores, and (c) communication efficiency ratios for sentence groups produced at a fixed f0 (solid bars) and sentence groups produced with variations in f0 (striped bars). NF = normal rate, fixed f0; SF = slow rate, fixed f0; NN = normal rate, prosodically natural variations in f0; NU = normal rate, prosodically unnatural variations in f0. Error bars represent 95% confidence intervals. For (a), post hoc pairwise comparisons revealed that the mean naturalness ratings between all sets are statistically significant (all p values < .05). For (b) and (c), brackets indicate mean differences between sets that are statistically significant (p < .05).



of requiring complex linguistic rulesets—or the user directly—to provide a natural f0 contour.

Intelligibility

We hypothesized that sentence-level variations in f0 would reduce intelligibility while simultaneously enhancing naturalness (Klopfenstein, 2016). In line with this hypothesis, we found that synthetic speech produced at a fixed f0 level was more intelligible than when produced with sentence-level variation in f0. These results are in agreement with those from Tjaden, Kain, and Lam (2014), in which resynthesized speech with exaggerated prosodic contours did not produce significant improvements in intelligibility. In that study, the authors speculated that the exaggerated f0 range may have been perceived as unnatural and, as a result, counteracted improvements in intelligibility. Similarly in this study, it is possible that the unnaturalness of the sentences with variations in f0 counteracted any improvements in intelligibility in the orthographic transcription task.

We also hypothesized that reducing the rate of synthesized speech would improve intelligibility. In contrast with our hypothesis and findings from Venkatagiri (1991), speech rate did not have a significant impact on the intelligibility of synthesized speech when produced at a fixed f0. This finding may be a result of the diphone concatenation methods of the MBROLA 2.00 algorithm. Specifically, decreasing speech rate using MBROLA 2.00 only stretches the spectral content of the sentence in time. Thus, acoustically relevant properties of speech, such as voice onset time, are linearly stretched in time to produce speech that sounds perceptually slower but otherwise identical in spectral content. This is in contrast to reducing the rate of typical speech, wherein the temporal characteristics of phonetic segments are nonlinearly related to the duration of articulatory gestures as a result of coarticulation effects (Hertrich & Ackermann, 1995). As such, our results demonstrate that this method of decreasing speech rate was not associated with significant improvements in the intelligibility of the synthetic speech output.

Communication Efficiency

Within this study, we employed communication efficiency as an additional indicator of functional speech communication. We hypothesized that reducing the rate of synthesized speech produced in a noise-degraded environment would improve intelligibility but reduce communication efficiency. As hypothesized, the synthesized sentences produced at a normal speech rate and a fixed f0 level were significantly more efficiently communicated than those produced at a slow speech rate and a fixed f0 level. Moreover, the mean communication efficiency ratios between sentences produced with sentence-level variations in f0 were not statistically different from each other. These findings suggest that our method of reducing speech rate led to degradations in communication efficiency and that our methods

of varying f0 did not significantly differ in their effects on communication efficiency.

Effects of Sentence Length

Previous work has implicated sentence length as a factor affecting the naturalness (e.g., Metz et al., 1990), intelligibility (e.g., Allison & Hustad, 2014; Frearson, 1985; Hustad, 2007; Speaks & Jerger, 1965; Venkatagiri, 1994; Yorkston & Beukelman, 1981b), and communication efficiency (e.g., Frearson, 1985; Yorkston & Beukelman, 1981b) of speech. Indeed, our results suggest that sentence length had a significant impact on speech naturalness and communication efficiency, with the interaction between sentence group and sentence length producing a significant effect on communication efficiency. Yet even though communication efficiency incorporates intelligibility in its calculation, sentence length did not produce a significant effect in the model for intelligibility. This result was not wholly unexpected because Venkatagiri (1994) demonstrated that sentences with a mean length of 11 words were as intelligible as those with a mean length of five words. It should also be considered that the communication efficiency ratio depends not only on intelligibility but also on sentence duration and the rate of intelligible speech.

It is possible that sentence length is confounded with semantic predictability, particularly when considering the significant impact of sentence length on the model for speech naturalness. As an example, the sentence "How are you?" is highly predictable, whereas "It's a great place to meet with friends and it's often quiet enough to read a newspaper or magazine" is sufficiently less predictable (Kalikow, Stevens, & Elliott, 1977). In addition to this discrepancy in predictability, the difference in length between these two utterances is 16 words. Thus, it might be speculated that listeners perceived some sentences as more unnatural than others as a result of an interaction between sentence length and predictability.

Limitations and Future Work

Although our results provide insight into how modulating speech rate and variation in f0 differentially impact the intelligibility, naturalness, and communication efficiency of synthetic speech, prosody is not limited to these elements. Further investigation should be undertaken to encompass all aspects of prosody, such as loudness and rhythm. In the same vein, this study examined linguistic prosody; however, the impacts of affective prosody should be also considered for investigation.

It is also worth noting that the diphone concatenation methods used to decrease speech rate may not be representative of natural decreases in speech rate. As such, different methods of altering synthesized speech rate should be compared using similar outcome measures. In addition, the order of the orthographic transcription and naturalness rating tasks was not randomized. Although this was designed to minimize ceiling effects that may occur if listeners heard noise-free sentences prior to transcribing them when presented in noise, it is possible that an order effect occurred wherein listeners perceived the sentences as more natural after hearing them in noise during the orthographic transcription task. Furthermore, it is unclear how experience level affects auditory–perceptual ratings of synthetic speech; as a result, the lack of a training regime prior to the naturalness task may have affected the clarity of the task.

Although our repeated-measures ANOVA design to assess perceived naturalness was sufficiently powered, it is possible that our nonsignificant intelligibility and communication efficiency results were due to limited statistical power. In order to limit learning in the intelligibility task while using identical stimuli as in the naturalness task, we utilized a mixed-effects ANOVA design. However, only large effects could be detected with our sample size of 16 when using this mixed-effects design. Smaller effects could exist and would not be detected via this experiment. In this study, we found that sentence length significantly impacted naturalness and communication efficiency. Follow-up investigations should balance sentence and semantic predictability to minimize the impact of such characteristics on the outcome variables. Furthermore, we chose to modulate f0 using the f0 contour of a single speaker; however, prosodic cues differ among individuals, particularly across genders (Fitzsimons, Sheahan, & Staunton, 2001). A follow-up study could therefore investigate synthesized speech reception using samples constructed from the f0 contours of different individuals within a diverse pool of participants (i.e., varying in gender, age, occupation, location, etc.). Using these samples, the interaction between speech rate and sentence-level f0 manipulations could also be evaluated. The results of these studies would further inform the implementation of automatic prosodic contour generation in speech synthesis incorporated into AAC systems.

Conclusions

Providing AAC users with prosodic synthetic speech output is a crucial step toward natural and intelligible speech synthesis. Yet, the extent to which prosodic cues differentially impact the social and functional reception of synthesized speech must first be evaluated to inform the implementation of prosodic control. Our results support a trade-off between social and functional speech: Speech produced with sentence-level f0 variation was less intelligible and less efficiently communicated but perceived as more natural compared to speech produced at a fixed f0 level. Additionally, decreasing speech rate reduced perceived naturalness but did not significantly impact intelligibility. With these results in mind, next steps include expanding the investigation to AAC users to determine how acceptable these individuals find the prosodically manipulated speech as an output from an AAC device. Overall, the results from the current investigation highlight the importance of considering multiple measures to evaluate the effects of prosody on synthesized speech, in addition to demonstrating preliminary evidence for the differential effects of basic prosodic manipulation on the social and functional reception of synthesized speech. Advances in knowledge surrounding the effects

of prosodic cues on synthesized speech will help to inform the development of AAC devices to enhance the quality of life of AAC users with limited speech capabilities.

Acknowledgments

This work was supported by the National Science Foundation under Grants 1510563 (awarded to Cara E. Stepp) and 1247312 (awarded to Jennifer M. Vojtech) and the National Institutes of Health under Grant F31 DC014872 (Cara E. Stepp). We would like to thank Zachary Morgan for assistance with data recording.

References

Aihong, D., Chundan, L., & Jingjing, W. (2014). Effect of speech rate for sentences on speech intelligibility. In *2014 IEEE International Conference on Communication Problem-solving* (pp. 233–236). Bejing, China: IEEE.

Allison, K. M., & Hustad, K. C. (2014). Impact of sentence length and phonetic complexity on intelligibility of 5-year-old children with cerebral palsy. *International Journal of Speech-Language Pathology, 16*(4), 396–407.

American Speech-Language-Hearing Association. (2015). *Definition of communication and appropriate targets*. Retrieved from http://asha.org/NJC/Definition-of-Communication-and-Appropriate-Targets/

Anand, S., & Stepp, C. E. (2015). Listener perception of monopitch, naturalness, and intelligibility for speakers with Parkinson's disease. *Journal of Speech, Language, and Hearing Research, 58*(4), 1134–1144.

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

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

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

Chabchoub, A., & Cherif, A. (2011). An automatic MBROLA tool for high quality arabic speech synthesis. *International Journal of Computer Applications, 36*(1), 1–5.

Chandra, E., & Akila, A. (2012). An overview of speech recognition and speech synthesis algorithms. *International Journal of Computer Technology and Applications, 3*(4), 1426–1430.

Corretge, R. (2012). *Praat vocal toolkit*. Barcelona, Spain: Praat. Retrieved from http://praatvocaltoolkit.com

Crabtree, M., Mirenda, P., & Beukelman, D. R. (1990). Age and gender preferences for synthetic and natural speech. *Augmentative and Alternative Communication, 6*(4), 256–261.

Drager, K. D. R., Hustad, K. C., & Gable, K. L. (2004). Telephone communication: Synthetic and dysarthric speech intelligibility and listener preferences. *Augmentative and Alternative Communication, 20*(2), 103–112.

Drager, K. D. R., Reichle, J., & Pinkoski, C. (2010). Synthesized speech output and children: A scoping review. *American Journal of Speech-Language Pathology, 19*(3), 259–273.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & Van der Vreken, O. (1996). The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. In *International Conference on Spoken Language Processing* (pp. 1393–1396). Philadelphia, PA: IEEE.

Eadie, T. L., & Doyle, P. C. (2002). Direct magnitude estimation and interval scaling of naturalness and severity in tracheoesophageal (TE) speakers. *Journal of Speech, Language, and Hearing Research, 45,* 1088–1096.

Eadie, T. L., & Doyle, P. C. (2004). Auditory–perceptual scaling and quality of life in tracheoesophageal speakers. *Laryngoscope, 114*(4), 753–759.

Evitts, P. M., & Searl, J. (2006). Reaction times of normal listeners to laryngeal, alaryngeal, and synthetic speech. *Journal of Speech, Language, and Hearing Research, 49*(6), 1380–1390.

Fitzsimons, M., Sheahan, N., & Staunton, H. (2001). Gender and the integration of acoustic dimensions of prosody: Implications for clinical studies. *Brain and Language, 78*(1), 94–108.

Fontan, L., Tardieu, J., Gaillard, P., Woisard, V., & Ruiz, R. (2015). Relationship between speech intelligibility and speech comprehension in babble noise. *Journal of Speech, Language, and Hearing Research, 58*(3), 977–986.

Frearson, B. (1985). A comparison of the AIDS sentence list and spontaneous speech intelligibility scores for dysarthric speech. *Australian Journal of Human Communication Disorders, 13*(1), 5–21.

Fucci, D., Reynolds, M., Bettagere, R., & Gonzales, M. D. (1995). Synthetic speech intelligibility under several experimental conditions. *Augmentative and Alternative Communication, 11*(2), 113–117.

Garcia, L. J., Laroche, C., & Barrette, J. (2002). Work integration issues go beyond the nature of the communication disorder. *Journal of Communication Disorders, 35*(2), 187–211.

Gibbon, D., & Bachan, J. (2008). An automatic close copy speech synthesis tool for large-scale speech corpus evaluation. In *Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC'08)* (pp. 902–907). Marrakech, Morocco: European Language Resources Association.

Hegde, M. N., & Freed, D. B. (2011). *Assessment of communication disorders in adults*. San Diego, CA: Plural.

Hustad, K. C. (2007). Effects of speech stimuli and dysarthria severity on intelligibility scores and listener confidence ratings for speakers with cerebral palsy. *Folia Phoniatrica et Logopaedica, 59*(6), 306–317.

Hustad, K. C., Beukelman, D. R., & Yorkston, K. M. (1998). Functional outcome assessment in dysarthria. *Seminars in Speech and Language, 19*(3), 291–302.

Hertrich, I., & Ackermann, H. (1995). Coarticulation in slow speech: Durational and spectral analysis. *Language and Speech, 38*(2), 159–187.

Jones, C., Berry, L., & Stevens, C. (2007). Synthesized speech intelligibility and persuasion: Speech rate and non-native listeners. *Computer Speech & Language, 21*(4), 641–651.

Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America, 61*(5), 1337–1351.

Kangas, K. A., & Allen, G. D. (1990). Intelligibility of synthetic speech for normal-hearing and hearing-impaired listeners. *Journal of Speech and Hearing Disorders, 55*(4), 751–755.

Klopfenstein, M. (2016). Speech naturalness ratings and perceptual correlates of highly natural and unnatural speech in hypokinetic dysarthria secondary to Parkinson's disease. *Journal of Interactional Research in Communication Disorders, 7*(1), 123–146.

Lagerberg, T. B., Johnels, J. Å., Hartelius, L., & Persson, C. (2015). Effect of the number of presentations on listener transcriptions and reliability in the assessment of speech intelligibility in

children. *International Journal of Language & Communication Disorders, 50*(4), 476–487.

Laures, J. S., & Weismer, G. (1999). The effects of a flattened fundamental frequency on intelligibility at the sentence level. *Journal of Speech, Language, and Hearing Research, 42,* 1148–1156.

Lindblom, B. (1990). On the communication process: Speaker–listener interaction and the development of speech. *Augmentative and Alternative Communication, 6*(4), 220–230.

Lúcio, G. S., Perilo, T. V., Vicente, L. C., & Friche, A. A. (2013). The impact of speech disorders quality of life: A questionnaire proposal. *CoDAS, 25,* 610–613.

McCall, F., Marková, I., Murphy, J., Moodie, E., & Collins, S. (1997). Perspectives on AAC systems by the users and by their communication partners. *European Journal of Disorders of Communication, 32*(3), 235–256.

Meltzner, G. S., & Hillman, R. E. (2005). Impact of aberrant acoustic properties on the perception of sound quality in electrolarynx speech. *Journal of Speech, Language, and Hearing Research, 48,* 766–779.

Metz, D. E., Schiavetti, N., & Sacco, P. R. (1990). Acoustic and psychophysical dimensions of the perceived speech naturalness of nonstutterers and posttreatment stutterers. *Journal of Speech and Hearing Disorders, 55*(3), 516–525.

Miller, N. (2013). Measuring up to speech intelligibility. *International Journal of Language & Communication Disorders, 48*(6), 601–612.

Mirenda, P., & Beukelman, D. R. (1987). A comparison of speech synthesis intelligibility with listeners from three age groups. *Augmentative and Alternative Communication, 3*(3), 120–128.

Moreton, E. (2008). Analytic bias and phonological typology. *Phonology, 25*(1), 83–127.

Neel, A. T. (2009). Effects of loud and amplified speech on sentence and word intelligibility in Parkinson disease. *Journal of Speech, Language, and Hearing Research, 52*(4), 1021–1033.

Nejime, Y., & Moore, B. C. J. (1998). Evaluation of the effect of speech-rate slowing on speech intelligibility in noise using a simulation of cochlear hearing loss. *The Journal of the Acoustical Society of America, 103*(1), 572–576.

Nusbaum, H. C., Francis, A. L., & Henly, A. S. (1997). Measuring the naturalness of synthetic speech. *International Journal of Speech Technology, 2*(1), 7–19.

Pampoulou, E. (2018). Speech and language therapists' views about AAC system acceptance by people with acquired communication disorders. *Disability and Rehabilitation: Assistive Technology, 18,* 1–8.

Panoiu, M., Rat, C.-L., & Panoiu, C. (2016). A comparative study of text-to-speech systems in LabVIEW. In V. E. Balas, L. C. Jain, & B. Kovačevič (Eds.), *Soft computing applications: Proceedings of the 6th International Workshop Soft Computing Applications (SOFA 2014)* (Vol. 1, pp. 3–11). Cham, Switzerland: Springer.

Patel, R., Connaghan, K. P., & Campellone, P. J. (2013). The effect of rate reduction on signaling prosodic contrasts in dysarthria. *Folia Phoniatrica et Logopaedica, 65*(3), 109–116.

Pierre-Yves, O. (2003). The production and recognition of emotions in speech: Features and algorithms. *International Journal of Human–Computer Studies, 59*(1), 157–183.

Ratcliff, A., Coughlin, S., & Lehman, M. (2002). Factors influencing ratings of speech naturalness in augmentative and alternative communication. *Augmentative and Alternative Communication, 18*(1), 11–19.

Reddy, V. R., & Rao, K. S. (2016). Prosody modeling for syllable based text-to-speech synthesis using feedforward neural networks. *Neurocomputing, 171,* 1323–1334.

Rendel, A., Fernandez, R., Hoory, R., & Ramabhadran, B. (2016). Using continuous lexical embeddings to improve symbolic-prosody prediction in a text-to-speech front-end. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5655–5659). Shanghai, China: IEEE.

Schröder, M., & Trouvain, J. (2003). The german text-to-speech synthesis system mary: A tool for research, development and teaching. *International Journal of Speech Technology, 6*(4), 365–377.

Scordilis, M. S., & Gowdy, J. N. (1989). Neural network based generation of fundamental frequency contours. In *1989 IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 219–222). Glasgow, United Kingdom: IEEE.

Speaks, C., & Jerger, J. (1965). Method for measurement of speech identification. *Journal of Speech and Hearing Research, 8*(2), 185–194.

Stipancic, K. L., Tjaden, K., & Wilding, G. (2016). Comparison of intelligibility measures for adults with Parkinson's disease, adults with multiple sclerosis, and healthy controls. *Journal of Speech, Language, and Hearing Research, 59*(2), 230–238.

Syrdal, A. K., Bunnell, H. T., Hertz, S. R., Mishra, T., Spiegel, M. F., Bickley, C., . . . Makashay, M. J. (2012). Text-to-speech intelligibility across speech rates. *Proceedings of the Interspeech* (pp. 623–626). Portland, OR.

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. *Journal of Speech, Language, and Hearing Research, 57*(4), 1191–1205.

Tjaden, K., Richards, E., Kuo, C., Wilding, G., & Sussman, J. (2013). Acoustic and perceptual consequences of clear and loud speech. *Folia Phoniatrica et Logopaedica, 65*(4), 214–220.

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. *Journal of Speech, Language, and Hearing Research, 57*(3), 779–792.

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.

Toda, T., Chen, L. H., Saito, D., Villavicencio, F., Wester, M., Wu, Z. Z., & Yamagishi, J. (2016). The voice conversion challenge 2016. In *17th Annual Conference of the International Speech Communication Association (Interspeech 2016)* (Vols. 1–5, pp. 1632–1636). Baixas, France: International Speech Communications Association.

Tsao, Y.-C., Weismer, G., & Iqbal, K. (2006). Interspeaker variation in habitual speaking rate: Additional evidence. *Journal of Speech, Language, and Hearing Research, 49*(5), 1156–1164.

Venkatagiri, H. S. (1991). Effects of rate and pitch variations on the intelligibility of synthesized speech. *Augmentative and Alternative Communication, 7*(4), 284–289.

Venkatagiri, H. S. (1994). Effect of sentence length and exposure on the intelligibility of synthesized speech. *Augmentative and Alternative Communication, 10*(2), 96–104.

Watson, P. J., & Schlauch, R. S. (2008). The effect of fundamental frequency on the intelligibility of speech with flattened intonation contours. *American Journal of Speech-Language Pathology, 17*(4), 348–355.

Wester, M., Wu, Z. Z., & Yamagishi, J. (2016). Analysis of the voice conversion challenge 2016 evaluation results. In *17th Annual Conference of the International Speech Communication*

Association (Interspeech 2016) (Vols. 1–5, pp. 1637–1641). Baixas, France: International Speech Communications Association.
Witte, R. S., & Witte, J. S. (2010). *Statistics*. Hoboken, NJ: Wiley.
Yorkston, K. M., & Beukelman, D. R. (1981a). *Assessment of intelligibility of dysarthric speech*. Tigard, OR: CC Publishing.
Yorkston, K. M., & Beukelman, D. R. (1981b). Communication efficiency of dysarthric speakers as measured by sentence intelligibility and speaking rate. *Journal of Speech and Hearing Disorders, 46*(3), 296–301.
Yorkston, K. M., Beukelman, D. R., Strand, E. A., & Bell, K. (2010). *Management of motor speech disorders in children and adults* (3rd ed.). Austin, TX: Pro-Ed.
Yorkston, K. M., Hammen, V. L., Beukelman, D. R., & Traynor, C. D. (1990). The effect of rate control on the intelligibility and naturalness of dysarthric speech. *Journal of Speech and Hearing Disorders, 55*(3), 550–560.
Yorkston, K. M., Strand, E. A., & Kennedy, M. R. T. (1996). Comprehensibility of dysarthric speech: Implications for assessment and treatment planning. *American Journal of Speech-Language Pathology, 5*(1), 55–66.
Zhao, Y. X., Kuruvilla-Dugdale, M., & Song, M. G. (2018). Structured sparse spectral transforms and structural measures for voice conversion. *IEEE/ACM Transactions on Audio Speech and Language Processing, 26*(12), 2267–2276.

Appendix

Synthesized Sentences

1. I think the owner just bought a brand new location, but it's in a really bad section of town.
2. You're a magnificently beautiful person.
3. It's a great place to meet with friends, and it's often quiet enough to read a newspaper or magazine.
4. Surprisingly, no other coffee place in town has free Internet.
5. This is my pomegranate smoothie, not his.
6. How are you?
7. Give me the food, not the container.
8. Please call today if possible.
9. Departmental politics are beneficial to no one.
10. I should consult the encyclopedia for my experiment.
11. It's really upsetting that his winter coat is not reversible.
12. The composer collided with the percussionist yesterday afternoon.
13. Sophisticated people have their own beliefs.
14. I demand additional information to be willing to invest in privatization.
15. What a corrosive personality.
16. The passengers grew restless during the everlasting voyage.