JSLHR

Research Article

Surface Electromyography–Based Recognition, Synthesis, and Perception of Prosodic Subvocal Speech

Jennifer M. Vojtech,[a,b] Michael D. Chan,[a] Bhawna Shiwani,[a] Serge H. Roy,[a] James T. Heaton,[c] Geoffrey S. Meltzner,[d] Paola Contessa,[a] Gianluca De Luca,[a] Rupal Patel,[d,e] and Joshua C. Kline[a]

Purpose: This study aimed to evaluate a novel communication system designed to translate surface electromyographic (sEMG) signals from articulatory muscles into speech using a personalized, digital voice. The system was evaluated for word recognition, prosodic classification, and listener perception of synthesized speech.
Method: sEMG signals were recorded from the face and neck as speakers with (n = 4) and without (n = 4) laryngectomy subvocally recited (silently mouthed) a speech corpus comprising 750 phrases (150 phrases with variable phrase-level stress). Corpus tokens were then translated into speech via personalized voice synthesis (n = 8 synthetic voices) and compared against phrases produced by each speaker when using their typical mode of communication (n = 4 natural voices, n = 4 electrolaryngeal [EL] voices). Naïve listeners (n = 12) evaluated synthetic, natural, and EL speech for acceptability and intelligibility in a visual sort-and-rate task, as well as phrasal stress discriminability via a classification mechanism.
Results: Recorded sEMG signals were processed to translate sEMG muscle activity into lexical content and categorize variations in phrase-level stress, achieving a mean accuracy of 96.3% (SD = 3.10%) and 91.2% (SD = 4.46%), respectively. Synthetic speech was significantly higher in acceptability and intelligibility than EL speech, also leading to greater phrasal stress classification accuracy, whereas natural speech was rated as the most acceptable and intelligible, with the greatest phrasal stress classification accuracy.
Conclusion: This proof-of-concept study establishes the feasibility of using subvocal sEMG-based alternative communication not only for lexical recognition but also for prosodic communication in healthy individuals, as well as those living with vocal impairments and residual articulatory function.
Supplemental Material: https://doi.org/10.23641/asha.14558481

Augmentative and alternative communication (AAC) systems enable those with limited speech production capabilities to input physical gestures or text, which may then be visually displayed and/or acoustically synthesized to support communication. AAC systems may be used to provide an alternative voice source for those who rely on alaryngeal speech; however, these technologies can be limited by poor intelligibility and difficulty training (e.g., with esophageal speech; Doyle & Eadie, 2005), tissue viability constraints and complicated air valve pressure maintenance (e.g., with tracheoesophageal speech; Hillman et al., 2005; Kramp & Dommerich, 2009), and/or the constant use of one's hand as in electrolaryngeal (EL) speech (Meltzner et al., 2005; Meltzner & Hillman, 2005). To overcome these limitations, research has explored subvocal speech recognition (SSR) as an alternative communication method. Subvocal speech is a form of communication in which a speaker makes articulatory movements in the absence of glottal excitation and thereby silently mouths their speech. SSR refers to the method of receiving the message and may occur through techniques such as lipreading or via sensors to

[a]Delsys/Altec, Inc., Natick, MA
[b]Boston University, MA
[c]Massachusetts General Hospital Department of Surgery, Boston
[d]VocaliD, Inc., Belmont, MA
[e]Northeastern University, Boston, MA
Correspondence to Joshua C. Kline: jkline@delsys.com

Editor-in-Chief: Cara E. Stepp
Editor: Jonathan S. Brumberg
Received May 17, 2020
Revision received October 15, 2020
Accepted January 29, 2021
https://doi.org/10.1044/2021_JSLHR-20-00257
Publisher Note: This article is part of the Special Issue: Selected Papers From the 2020 Conference on Motor Speech—Basic Science and Clinical Innovation.

Disclosure: *Delsys, Inc., is a private company that manufactures, markets, and sells electromyographic sensors and other physiological measurement systems. VocaliD, Inc., is a private company that develops, markets, and sells personalized digital voices. The authors have declared that no competing interests existed at the time of publication.*

capture different aspects of the speech production and output the recognized message (Denby et al., 2010). These systems can leverage articulatory motions via electromagnetic articulography (EMA; Fagan et al., 2008; Kim et al., 2017, 2018), ultrasound and optical imaging (Crevier-Buchman et al., 2011; Denby et al., 2010; Hueber et al., 2010), or the neural pathways contributing to speech via implants in the speech motor cortex (Brumberg et al., 2013; Guenther et al., 2009) or electroencephalographic sensors (Porbadnigk et al., 2009).

Speech recognition using surface electromyography (sEMG) is a particularly promising method, as it can be performed noninvasively while an individual produces subvocal speech. sEMG-based SSR operates by recording the electrical signals generated by articulatory muscle contractions during speech production using sensors placed over muscles on the face and neck skin surface. Numerous groups have worked to advance sEMG-based SSR using signal-based classification methods for recognizing recited numbers, isolated words, or small phrases among speakers with typical voices (Jorgensen et al., 2003; Jou et al., 2006; Maier-Hein et al., 2005). However, the efficacy of such existing technologies has yet to be established efficacy among those with speech or voice impairments and/or remain limited to relatively small (less than 60 isolated words) vocabulary sets that fall short of continuous speech generation (Betts & Jorgensen, 2005; Jorgensen et al., 2003; Lee, 2008; Manabe & Zhang, 2004; Meltzner et al., 2008).

Prior subvocal speech research has attempted to overcome these shortcomings by coupling advancements in sEMG sensor technology with automated speech recognition (ASR) using machine learning techniques. For instance, Meltzner et al. (2017, 2018) used small sEMG sensors to conform to the complex anatomy of small articulatory muscles of the face and neck, known to provide high-fidelity recordings of these relatively low amplitude sEMG signals during natural speech (De Luca et al., 2012; Roy et al., 2007). The authors further developed phoneme-based speech recognition models that utilized many of the methods evolved from ASR applications, including Mel-frequency cepstral coefficients (MFCCs) that were adapted to the sEMG signal spectrum. When tested on a 2,500-word vocabulary of continuous phrases among speakers with typical voices and speakers with laryngectomy, their recognition models were able to translate sEMG-based subvocal speech to text with a mean word recognition rate (WRR) of over 90% (Meltzner et al., 2017, 2018).

By design, many SSR systems (among other AAC technologies) have demonstrated the ability to recover lexical content yet struggle to convey the *suprasegmental (prosodic)* attributes of the content—vital components of natural speech (Pullin et al., 2017). Specifically, vocal characteristics such as pitch, voice quality, loudness, and temporal variability are essential mechanisms for conveying emotion, mood, and personality (Drager et al., 2010; Evitts & Searl, 2006; Fucci et al., 1995; Kangas & Allen, 1990; McCall et al., 1997). The absence of such suprasegmental content of speech has considerable impact on the psychosocial life of individuals living with vocal impairments, including a loss of identity, lack of confidence, social isolation, depression, and reduced intimacy (Garcia et al., 2002; Hegde & Freed, 2011; Lúcio et al., 2013; Meltzner & Hillman, 2005; Patel & Threats, 2016). Preliminary investigations of the feasibility to derive continuous estimates of prosody for SSR systems by Gonzalez et al. (2017) have shown that it is possible to approximate changes in vocal pitch via permanent magnetic articulography (PMA)–based estimates of fundamental frequency (f0) during subvocal speech when the system is trained using acoustic signals prior to use. The authors demonstrated that the modulated pitch contours from their synthesized speech achieved more favorable listener ratings of naturalness than did a monotonic pitch. While these findings highlight the feasibility of tracking continuous changes in prosody, further work is needed to overcome the invasive nature of PMA—which is known to have limited utility outside controlled laboratory settings—and the reliance on acoustic data for training the SSR for individuals who are unable to vocalize, such as people with laryngectomy.

sEMG-based AAC devices have the potential to overcome the shortcomings of other devices to noninvasively harness both segmental and suprasegmental attributes of speech. In particular, the extrinsic laryngeal muscles are indirectly responsible for changing the vibratory rate of the vocal folds (perceptually known as vocal pitch; Honda et al., 1999; Ueda et al., 1972) via altering laryngeal height and tilt (Broniatowski et al., 1999; Sataloff et al., 2007; Suárez-Quintanilla et al., 2019). Indeed, recent investigations have demonstrated a natural association between the sEMG activity of extrinsic laryngeal muscles and pitch modulation of human voice (Goldstein et al., 2004, 2007; Heaton et al., 2004; Kubert et al., 2009; Stepp et al., 2010). Additional studies examining the sEMG activity of orofacial muscles demonstrated an increase in activity following increases in vocal rate and loudness (McClean, 2000; McClean & Tasko, 2002, 2003). Thus, sEMG-based AAC shows promise for detecting suprasegmental (prosodic) attributes, such as pitch, loudness, and rate.

Prior work shows promise for sEMG-based devices to extract prosodic information from subvocal speech. Janke and Diener (2017) used sEMG sensors placed over orofacial and submental muscles of three speakers with typical voices to simultaneously record acoustic and sEMG signals as the speakers read sentences aloud. The authors processed the acoustic data to train their SSR system to identify the lexical content and f0 contour of each sentence, ultimately observing a substantial decline in the perceptual evaluation of speech naturalness when comparing reference audio signals to the synthesized speech. Diener et al. (2019) built upon this work to improve the generation of the f0 contour from sEMG data by quantizing f0 values; this is in contrast to the work from Janke and Diener, wherein baseline regression is used to predict f0 rather than choose an f0 from a finite set of values. Although the results of this work demonstrated marked improvements in f0 contour generation within an sEMG-based SSR device, the SSR system was trained using audible speech data, which may not be possible to acquire from speakers postlaryngectomy

since the vocal folds are removed in a total laryngectomy (i.e., no audible f0). Furthermore, the articulatory musculature may change following a total laryngectomy, such that a system trained for an individual able to phonate prior to laryngectomy may not function effectively postlaryngectomy. While these studies lend substantial support to the use of sEMG-based SSR technology that can identify both lexical and prosodic content, additional research work is crucial to alleviate the required dependence on acoustic information.

The aim of the current study was to adapt sEMG-based SSR technology that can recognize lexical content and basic prosodic manipulations in the "absence" of an acoustic signal to meet the needs of those with existing vocal impairments. This work was performed by (a) developing an sEMG-based SSR system that could not only recognize subvocal phrase content but also discriminate between simple manipulations in prosody and (b) evaluating the categorization of the recognized prosodic content by listeners when translated into synthetic speech output. To achieve these goals, algorithms were adapted from previous work (Meltzner et al., 2017, 2018) to recognize subvocal phrases produced with specific prosodic patterns (e.g., phrases with first-word or last-word stress) and separately determine where these patterns occurred. Lexical content detected from the SSR algorithms were then combined with text-to-speech synthesis engines to produce audible speech using a personalized, digital voice. Speech generated by the synthesized voices was then evaluated in a series of perceptual listening experiments to assess acceptability, intelligibility, and phrasal stress discriminability as compared to natural and EL alternatives. The following hypotheses were proposed:

1. Modifying existing speech recognition models to be robust to variations in phrasal stress will enable the sEMG-based SSR system to accurately detect lexical content of varying phrase-level stress.
2. The spectral and temporal characteristics of the sEMG recordings during subvocally stressed words will differ from unstressed words, such that the sEMG-based SSR system will be able to discriminate phrase-level changes in stress.
3. The acceptability and intelligibility of synthetic speech will be significantly greater than that of speech produced using an EL.
4. Listeners will discriminate phrasal stress to a higher degree in natural and synthetic speech than EL speech.

Method for Protocol I: sEMG-Based Voice Synthesis

Speakers

Four adults with typical voices (three women, one man; M = 25.8 years, SD = 2.7 years, range: 23.0–29.4 years) and four adults who had undergone a total laryngectomy (four men; M = 70.3 years, SD = 5.4 years, range: 65.4–77.4 years) participated in the study. The speakers with laryngectomy were fluent in English and had undergone a total laryngectomy at least 1 year prior to study enrollment (M = 4.4 years, SD = 2.5 years, range: 1.6–7.2 years), as described in Table 1. The speakers with typical voices were fluent in English; were nonsmokers; and reported no history of voice, speech, language, or hearing disorders. Prior to participation, speakers provided informed, written consent in compliance with the Western Institutional Review Board.

Data Acquisition

Sensor Preparation and Configuration

sEMG signals were recorded using wireless double parallel bar Trigno Quattro sensors (Delsys, Inc.). Prior to sensor placement, the surface of each speaker's skin was prepared by removing excessive hair and cleaning using alcohol wipes and tape peel exfoliation (Hermens et al., 2000; Roy et al., 2007; Stepp, 2012). Eight single differential sEMG sensors (25 × 12 × 7 mm) were then placed in specific locations of the face and neck shown in Figure 1 using double-sided hypoallergenic adhesive tape (refer to Meltzner et al., 2017, 2018, for details on sensor location). Briefly, four sensors were placed on the right ventral neck, with two submental sensors (1 and 2 in Figure 1) and two ventromedial sensors (3 and 4 in Figure 1). These sensors were placed over the (1) anterior belly of the digastric, mylohyoid, and geniohyoid (Palmer et al., 1999); (2) platysma, mylohyoid, and stylohoid (Palmer et al., 1999); and (3, 4) platysma, thyrohyoid, omohyoid, and sternohyoid (Ding et al., 2002). Four sensors were placed on the right side of the face, with two supralabial sensors (5 and 6 in Figure 1) and two infralabial sensors (7 and 8 in Figure 1). These sensors were placed over (5) zygomaticus major and/or minor, levator labii superioris, and levator anguli oris; (6) orbicularis oris (upper lip); (7) orbicularis oris (lower lip); and (8) mentalis (Eskes et al., 2017). All sensors were adhered using double-sided hypoallergenic adhesive tape. sEMG signals were recorded at 2222 Hz, bandpass filtered between 20 and 450 Hz, and amplified by a gain of 300 using custom Delsys software.

Speech Corpus

The speech corpus implemented in the current study was derived from a subset of phrases used in previous work examining sEMG-based speech recognition. Specifically, Meltzner et al. (2017) constructed a 2,500-word corpus that included 980 phrases from standard data corpora, including (a) the Boston Children's Hospital corpus for message banking (Costello, 2014), (b) the TIMIT-SI corpus (Garofolo et al., 1993), (c) the TIMIT-SX corpus (Garofolo et al., 1993), and (d) a set of the most common English phrases.[1] Given that prior work has demonstrated sEMG-based alaryngeal speech recognition, the corpus size was reduced to minimize vocal fatigue that may occur during the recording process while ensuring a minimal set of training vocabulary with comprehensive, balanced suprasegmental phoneme combinations (Searl & Knollhoff, 2018; Welham & Maclagan, 2003).

[1] https://www.englishspeak.com/en/english-phrases

Table 1. Demographic information of laryngectomee speakers.

ID	Sex	Age (years)	Years since total laryngectomy	Communication modality	Notes
L1	M	71.3	7.2	Servox EL	Monopitch EL
L2	M	66.9	5.1	TruTone EL	Partial glossectomy due to tongue cancer EL set to moderate pitch range
L3	M	77.4	3.3	Servox EL	Monopitch EL
L4	M	65.4	1.9	TruTone EL	EL set to narrow pitch range

Note. M = male; EL = electrolaryngeal.

To achieve this goal, the 980-token corpus was reduced to 650 unique phrases (1,739 unique words) by balancing the frequency of phoneme and triphone combinations. Of the 650-token corpus, 50 phrases were repeated twice more to include variable prosody by including stress on the first word (e.g., "JANE loves Bob," with phrasal stress denoted by capitalized words) and then stress on the last word (e.g., "Jane loves BOB"). The final corpus thus comprised 750 tokens, including 600 unique phrases that did not contain cues for phrasal stress ("lexical corpus") and 150 tokens that were repeats of the same 50 base phrases and contained variations in phrasal stress (no stress, first word, last word; henceforth "stress corpus").

Experimental Overview

Data collection software was designed to allow speakers to self-control the start/stop of recording during subvocal recitation: Speakers were provided with a push-button switch (7700DW Crest Healthcare Supply) that was programmed to control mouse clicks (StealthSwitch3, H-Mod, Inc.), such that the speaker would press the push-button to start and stop sEMG data collection. Upon initiating data collection, a prompt would appear on a computer monitor (U2414H oriented at a resolution of 720 W × 1280 H, Dell), instructing the speaker to subvocally recite a phrase displayed on the screen.

Figure 1. Depiction of sensor configurations targeting (1) anterior belly of the digastric, mylohyoid, and geniohyoid; (2) platysma, mylohyoid, and stylohoid; (3, 4) platysma, thyrohyoid, omohyoid, and sternohyoid; (5) zygomaticus major and/or minor, levator labii superioris, and levator anguli oris; (6, 7) orbicularis oris; and (8) mentalis.

Prior to starting data collection, speakers were familiarized to the experiment. This included outfitting the speakers with the sEMG sensors (described in the Sensor Preparation and Configuration section) and instructing the speakers to practice subvocal recitations of an example phrase shown on the screen. To subvocally recite speech tokens, speakers were instructed to deliberately reduce their speaking rate and hyperarticulate (involving increased mouth opening) when silently mouthing phrases, similar to the use of clear speech (Cox & Doyle, 2018; Krause & Braida, 2002; Picheny et al., 1985; Smiljanić & Bradlow, 2009) yet in the absence of an acoustic signal. This method of instructing speakers to subvocalize was used in the current study to increase articulatory effort (e.g., via movement distances and durations as in clear speech; Lam & Tjaden, 2013), as is typically instructed for persons learning to use an EL following laryngectomy (Adler & Zeides, 1986; Diedrich & Youngstrom, 1966; Hočevar-Boltežar & Žargi, 2001). During the speaker recitations of the example phrase, signal quality was visually inspected by the experimenter, and sensors were adjusted as necessary to reduce poor skin contact, movement artifacts, or poor sensor positioning relative to the muscles of interest (Meltzner et al., 2018).

From here, speakers began recording the 750-token corpus. Speakers were requested to notify the experimenter if they misarticulated a phrase at any time during the recording session so that the phrase could be repeated. The sEMG data from each of the eight sensors were displayed in real time on a secondary computer monitor (U2414H oriented at a resolution of 1280 W × 720 H, Dell) for the experimenter to monitor signal quality throughout the recording session. Corpus tokens were presented in the same order for each speaker to ensure that a consistent number of phrases could be obtained across all speakers. For the tokens that incorporated phrasal stress (first word, last word), speakers were instructed to stress the capitalized word according to how they would typically stress a word. For instance,

one speaker with laryngectomy reported using a longer word duration when trying to stress a word when using an EL; as such, this speaker used the same method during subvocal recordings in the current study.

After subvocally reciting the 600 tokens of the lexical corpus, speakers were asked to recite the 150-token stress corpus in their typical mode of communication. As in the subvocal recordings, all speakers were instructed to introduce phrasal stress as they typically would, without experimenter guidance. For the experiments involving the speakers with typical voices, each speaker also produced the stress corpus in their typical pitch and loudness while their voice was recorded by a microphone. For the experiments involving speakers with laryngectomy, each speaker produced the stress corpus using their preferred EL speech aid (see Table 1). Speech aids were a Servox Digital XL Electrolarynx (Servox Digital) and a TruTone EMOTE Electrolarynx (Griffin Laboratories). Although the TruTone EMOTE Electrolarynx can provide push-button pitch modulation, only one of the two TruTone users was comfortable using the dynamic pitch functionality. This speaker used the TruTone EMOTE with the dynamic pitch control feature set at the medium range (Mode 3), which he felt best represented the dynamic pitch control range he used daily. The speaker used the dynamic pitch control functionality via increasing the pressure of his finger on the power button to stress the intended word. Natural and EL speech were captured via a condenser headset microphone (PLM31, PylePro, Pyle Audio) and sampled at 44.1 kHz with 16-bit resolution. These recordings were used to guide the development and facilitate the perceptual evaluation of synthetic voices described in the sEMG-Based Voice Transcription and Protocol II: Listener Perceptual Assessment of Synthetic Speech sections, respectively. In total, the sEMG acquisition session required approximately 2–3 hr to capture all 750 tokens, including time for consent, task introduction, and multiple 5-min stretch breaks between subvocal recitation trials to minimize fatigue and boredom.

sEMG-Based Voice Synthesis

sEMG-Based Word and Prosody Recognition

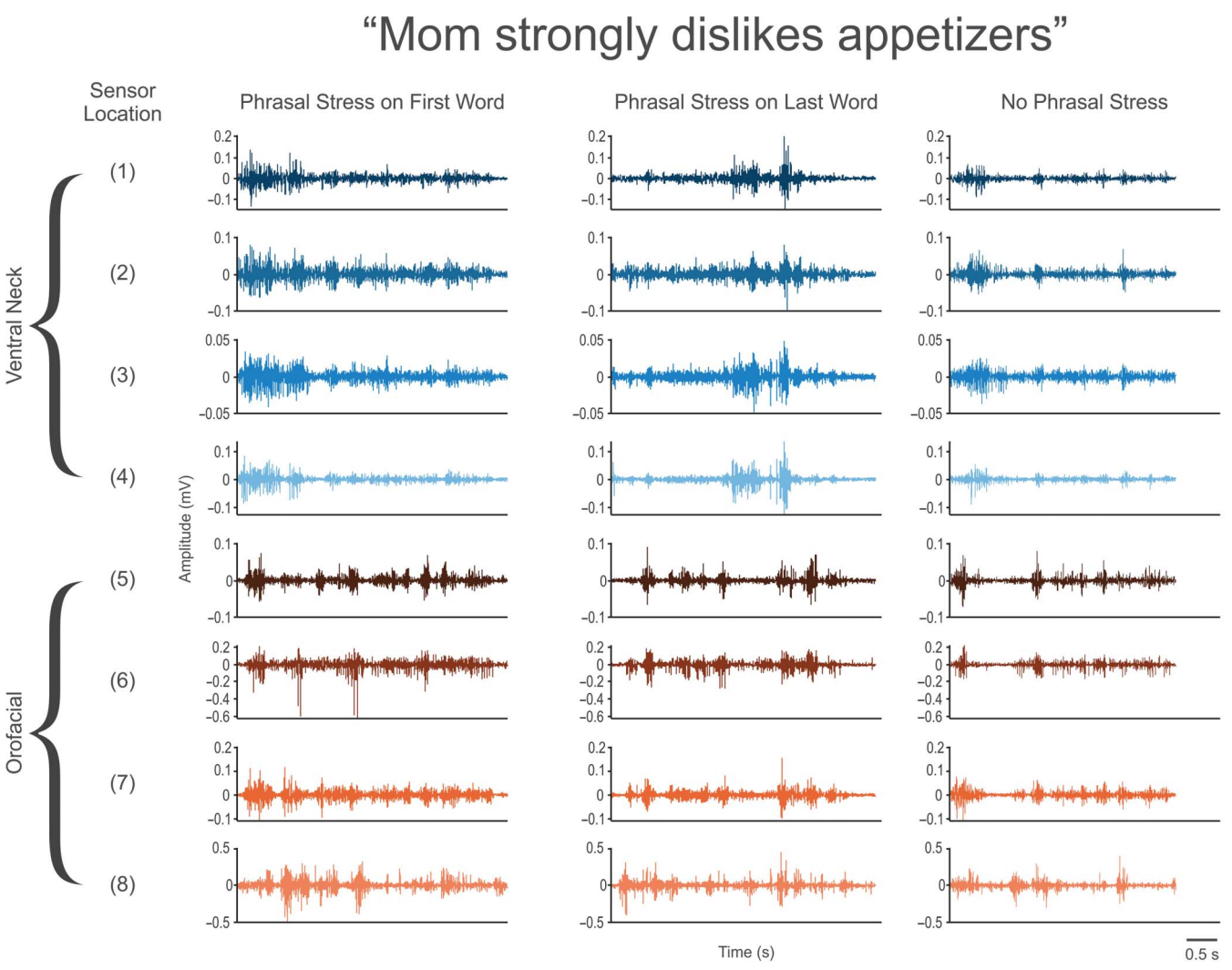
1. Algorithm development. Following data collection, SSR algorithms were developed to detect the lexical and prosodic content from the sEMG signals of the subvocally recited phrases of the data corpus. It is important to note that sEMG activity can vary substantially not only across speaker, sensor location, and phonemic content, as shown in previous work (Meltzner et al., 2017, 2018), but also according to phrasal stress type. For example, Figure 2 shows the sEMG signals from one speaker with laryngectomy (L2 in Table 1) who produced the token "Mom strongly dislikes appetizers" with phrasal stress on the first word of the token ("Mom") and on the last word of the token ("appetizers"), as well as without any phrasal stress. It can be observed that signal amplitude increased when this speaker introduces phrasal stress, particularly when detected from the ventral neck (i.e., Sensors 1–4). In addition, the duration of the token was longer when phrasal stress was introduced on either the first or last word. In a set of messages with the same lexical content but variable phrasal stress content, the goal of recognizing words was to identify similarities among these tokens, whereas the goal of categorizing stress was to detect differences among these tokens. Sensor subsets that characterize these prosodic and lexical features for achieving desired goals of this work varied across speakers. For this reason, the full eight-sensor data set was utilized for all subsequent analyses rather than isolating a small subset of sensors that show observable differences in activity for each speaker. Likewise, the features used for lexical recognition and phrasal stress categorization were combined across sensor locations when possible to provide insight into the cumulative contribution of the feature to the SSR algorithms.

To develop an SSR system that could both recognize lexical content (i.e., similarities in sEMG features in the presence of prosodic modulations) and categorize phrasal stress (i.e., differences in sEMG features in repetitions of the same token), algorithms were adapted from those developed by Meltzner et al. (2017, 2018), which comprised a multistage architecture that utilized modified MFCCs derived from the raw sEMG signals recorded during monotonic subvocal recitation of continuous phrases. These MFCCs are first used to train a three-state, left-to-right hidden Markov model on a per-phoneme scale, followed by context-dependent triphone models, which, taken together, facilitate the recognition of each phoneme whose sEMG-based manifestation may be influenced by the presence of adjacent phonemes. Subspace Gaussian mixture models are then applied to train the model for recognizing patterns of phonemes from previously unseen words and phrases. A data-driven decision tree (using the KALDI toolkit decision tree algorithm; Povey et al., 2011) is used to cluster the triphone models for the observed training data. Additional details on these algorithms can be found in Meltzner et al. (2017, 2018).

In this study, these SSR algorithms were modified to accurately recognize tokens produced with variable prosody. Specifically, subject adaptive training was introduced alongside two primary modifications: a feature space maximum likelihood linear regression (fMLLR) processing stage to improve the recognition of lexical content and a multilayer perceptron neural network to categorize differences in phrasal stress.

a. Lexical content recognition. A new fMLLR processing stage was used to adjust the influence of different modified MFCC input features to the word recognition models. In particular, MFCCs that are most affected by variations in stress would have less influence on the recognition models than more stress stationary MFCC features. In this work, fMLLR was implemented as an unsupervised process where the Gaussian mixture models are trained using MFCC features normalized for each subject based on the subject adaptive training algorithm (Anastasakos et al., 1996); this algorithm allows the model to operate on a more homogenized feature set with lower variability from suprasegmental emphasis on the phonemic content. By adapting

Figure 2. Example of raw surface electromyographic signals obtained from one speaker with laryngectomy from the token "Mom strongly dislikes appetizers." Each column shows where phrasal stress occurred within the token, including on the first word ("Mom"; left column) and on the last word ("appetizers"; middle column), as well as without any phrasal stress (right column). Electromyographic sensors located on the ventral neck (1–4) are colored in blue, and those located on the face (5–8) are colored in orange.



to changes in phrasal stress, the fMLLR algorithm increases the probability of recognizing variations of the same triphone across multiple tokens.

b. Phrasal stress categorization. A set of per-channel, sEMG-based features was identified to categorize differences in phrasal stress (first word, last word, no stress) via capturing suprasegmental attributes of speech based on the patterns of articulatory muscle activity. In the current study, phrasal stress was characterized as the overall emphasis or prominence given to a certain word—in this case, the first word or last word of a phrase—rather than specific suprasegmental attributes that contribute to phrasal stress (e.g., pitch, loudness, and timing; Bolinger, 1958; Fry, 1955, 1958). As sEMG signals are detected during "subvocalized" speech, a range of spectral and temporal features were selected to capture changes in articulatory effort that would otherwise occur when introducing phrasal stress in audible speech (Fry, 1955; Harris, 1978; Tuller et al., 1981) and may differ across individuals. The details of these features are described in detail below:

i. Temporal Waveform Length: The temporal waveform length (TWL) was computed using methodology from Vojtech et al. (2018) to capture the combined time, frequency, and amplitude complexity of muscle activation (Ahsan et al., 2011; Englehart & Hudgins, 2003; Hudgins et al., 1993; Phinyomark et al., 2012). TWL was included in this feature set to capture changes in signal amplitude (Harris, 1978) and phasic modulations in the timing and frequency of motor unit firings (Fry, 1955; Tuller et al., 1981) that may result from the introduction of phrasal stress. The sEMG signal was first divided into two equally timed segments, then TWL was estimated for

each segment as the cumulative length of the sEMG waveform. This process was repeated for each of the eight sEMG sensors, resulting in a pair of TWL values at each sensor location. The TWL values were then summed across respective segments across all sensor locations. A single pair of TWL values (i.e., one TWL value for the first half of the signal and one TWL value for the second half of the signal) were therefore obtained per speech token that could be compared against each other to categorize phrasal stress (i.e., first word, last word, no stress).

ii. Spectral Waveform Length: The spectral waveform length (SWL) was included as an estimate of spectral compression, which has demonstrated effectiveness characterizing fatigue and changes in motor unit recruitment induced in the intrinsic laryngeal muscles by effortful vocalizations (Boucher et al., 2006). SWL was computed from the raw sEMG signal at each sensor location after segmenting the signal into two equally timed frames. The waveform length of the power spectrum of each frame was estimated to produce SWL features describing the cumulative output of power spectral density per sensor location per speech token. These SWL features were then summed over respective frames of all sensor locations to obtain an overall estimate of SWL at the beginning and end of each speech token.

iii. Pairwise Cross-Correlation: The pairwise cross-correlation (PCC) of muscle activity across sensors was included to capture the degree of coordinated activity patterns across muscles that are necessary to alter the configuration of the vocal tract to, in turn, produce certain speech sounds. The PCC was computed from the root-mean-square of signal amplitude across two sensor locations (De Luca & Mambrito, 1987). A single PCC estimate was then obtained by averaging the PCC values from all sensors to provide a cumulative estimate of the level of muscle coactivation.

iv. Dominant MFCCs: A subset of MFCC features from the sEMG-based recognition algorithms were also included within the phrasal stress discriminability algorithms. MFCCs are among the most widely used feature extraction methods for speech analysis, as they are computationally simple and highly robust to noise (Liu et al., 2018; Meltzner et al., 2017, 2018). The MFCC-based features were derived from the MFCC superset, which comprised 56 MFCC values per frame from each of eight sensor locations per speech token. Distinctive MFCC content was extracted from each sensor location to capture the contribution of that muscle to global changes in phrasal stress in the subvocal speech. The dominant MFCC features corresponded to different sensor locations across different phonemic composition of sentences in the training set. To attain a 10:1 sample-to-feature ratio needed for robust model training without overfitting (Duda et al., 2012), the 10 most significant singular vectors were extracted from each MFCC feature set via singular value decomposition (Zhang et al., 2017), a matrix factorization technique that uses an orthogonal transformation approach to convert a set of observations of possibly correlated variables into a set of uncorrelated singular vectors. The result of this process produced 10 MFCC features for subsequent analyses.

The final set of 15 features (two TWL, two SWL, one PCC, 10 MFCC) cover a broad range of spectral and temporal changes in the sEMG signal to account for potential changes in articulatory effort when introducing phrasal stress. The features served as inputs to a multilayer perceptron neural network, which comprised two hidden layers utilizing adaptive moment estimation (Kingma & Ba, 2014) optimization to maintain a per-parameter adaptive learning rate that would accommodate speaker-specific stress variations.

2. Algorithm evaluation. The algorithms were trained and tested on subsets of the 750-token corpus to assess the ability of the sEMG-based SSR system to (a) recognize lexical content and (b) categorize phrasal stress. These subsets are shown in Table 2, as well as described below.

To examine the ability of the system to recognize lexical content, the algorithms were trained on the 600-token lexical corpus (80%) and then evaluated on the 150-token stress corpus (20%). Accuracy in recognizing lexical content was calculated via WRR as the mean percentage of words correctly identified in each phrase of the 150-token stress corpus.

To examine the ability of the sEMG-based SSR system in categorizing phrasal stress, the stress corpus was split into stress training (80%; 120 tokens) and stress test (20%; 30 tokens) sets. This split allotted 10 unique phrases of each stress type (first word, last word, no stress) to evaluate phrasal stress discriminability across eight speakers, totaling 80 samples. The classification algorithms were first trained to discriminate the three stress types using the 120-token stress training set. Classification performance was then assessed on the 30-token stress test set via metrics of phrasal stress discriminability and F1 score. Phrasal stress discriminability was evaluated as the percentage of tokens that were correctly categorized according to stress type (first word, last word, no stress). F1 score, on the other hand, was selected as a metric of discrimination accuracy to describe the precision and recall of the classifier in categorizing the tokens by stress type (see Sokolova & Lapalme, 2009; van Rijsbergen, 1979, for more details). In the current study, F1 score was computed using methodology described in Johner et al. (2012) for ease of comparison across studies. The output stress classifications were then used to provide the lexical and prosodic content for synthesizing personalized prosodic voices for subsequent perceptual evaluation, described in the Protocol II: Listener Perceptual Assessment of Synthetic Speech section.

sEMG-Based Text-to-Speech

Following the identification of lexical content and the categorization of phrasal stress type, the next step in the sEMG-based SSR pipeline includes synthesizing the

Table 2. Description of the corpus characteristics for the algorithm evaluation tasks.

Corpus characteristics	Algorithm evaluation tasks			
	Recognize lexical content		Categorize phrasal stress within stress corpus	
	Training set	Test set	Training set	Test set
Corpus name:	Lexical corpus	Stress corpus	Stress training set	Stress test set
Total no. of phrases	600	150	120	30
No. of unique phrases	600	50	40	10
Composition:	Unique phrases with no phrasal stress	50 phrases each with stress on the first word, last word, or no stress	40 phrases each with stress on the first word, last word, or no stress	10 phrases each with stress on the first word, last word, or no stress

recognized message into audible speech. To simulate this process, the stress corpus evaluated in the sEMG-Based Word and Prosody Recognition section was first pruned to remove tokens that were incorrectly classified by the recognition algorithms (words and/or stress condition) and/or consistently misarticulated by one or more speakers (e.g., one laryngectomized speaker reported problems producing /l/ due to a partial glossectomy). In doing so, a combination of 12 phrases (each produced with the three types of phrasal stress) was identified that balanced the incidence of monophones and triphones. These 36 tokens were considered as the "speech bank" to be synthesized by text-to-speech engines within the sEMG-based SSR system.

Eight personalized synthetic voices were trained using the Personifier speech engine by VocaliD, Inc. (Toman et al., 2018). The Personifier speech engine uses a deep learning, sequence-to-sequence architecture (Shen et al., 2018), coupled with a neural vocoder (Kalchbrenner et al., 2018), to synthesize text into audible speech. The synthetic voices for the speakers with typical voices were trained using speech data that the four speakers recorded via a web interface (VocaliD). Synthetic voices for the speakers with laryngectomy were created using previously recorded data from VocaliD's Human Voicebank of age-, gender- and geographic location–matched speakers. Static representations of pitch, timing, and loudness were derived with respect to the stress classification of each phrase to inform the Personifier speech engine of how phrasal stress was imparted on the first or last word. Tokens with phrasal stress were to be modified by increasing the f0 (Bolinger, 1958; Fry, 1955, 1958), duration (Fry, 1955), and intensity (Fry, 1955) of the intended word from the baseline characteristics of the synthetic voice. In turn, the Personifier speech engine imparted the desired stress postsynthesis using a combination of empirically derived gain modification algorithms, as well as pitch and segmental duration manipulation algorithms based on the pitch synchronous overlap and add algorithm (Moulines & Charpentier, 1990). By using these algorithms, phrasal stress was categorically, statically added to the 24 of 36 tokens of the speech bank to add emphasis to either the first or last word. This process was repeated for each of the eight synthetic voices, resulting in a total of 384 synthetic phrases (12 phrases × 3 phrasal stress types × 8 synthetic voices) to be analyzed in a series of perceptual experiments.

Method for Protocol II: Listener Perceptual Assessment of Synthetic Speech

Listeners

Twelve adults aged 23–37 years (seven men, five women; M = 26.7 years, SD = 4.9 years) were recruited as inexperienced listeners for the study. Listeners were healthy adults who reported no history of voice, speech, language, or hearing disorders. All listeners spoke English, were naïve to the purpose of the study, and were unfamiliar with alaryngeal speech. Prior to participation, listeners provided written consent in compliance with the Western Institutional Review Board.

Experimental Overview

Listeners were seated in a quiet room and wore headphones to perceptually evaluate four voice sources: natural voice (NV), synthetic voice matched to speakers with typical voices (SV_C), EL voice (EV), and synthetic voice matched to speakers with laryngectomy (SV_L). In the current study, the synthetic voice groups were separated by speaker type (control, laryngectomized) to ensure equal sample groups. Table 3 shows an overview of these voice sources. Each of the four voice sources was perceptually assessed in two experimental paradigms, which were presented to all listeners in the following order: a visual sort-and-rate task and a classification task. All listeners were requested to take rest breaks at 20-min intervals. Both paradigms were completed in an average of 1.6 hr (SD = 0.4 hr). An example of the phrase "chestnuts are starchy" is available in .wav format for one EL speaker (denoted by "EV") and his matched synthetic voice ("SV") when containing first word stress ("_FS"; see Supplemental Material S1 and S4, respectively), last word stress ("_LS"; see Supplemental Material S2 and S5, respectively), and no phrasal stress ("_NS"; see Supplemental Material S3 and S6, respectively).

Visual Sort-and-Rate Task

In the visual sort-and-rate task (Granqvist, 2003), listeners assessed the acceptability and intelligibility of NV, SV_C, EV, and SV_L speech tokens. Acceptability—an auditory-percept relating to how pleasant a voice sounds when considering pitch, quality, rate, and understandability (Bennett & Weinberg, 1973)—was chosen to evaluate

Table 3. Overview of four voice sources perceptually examined by listener visual sort-and-rate and phrasal stress classification tasks.

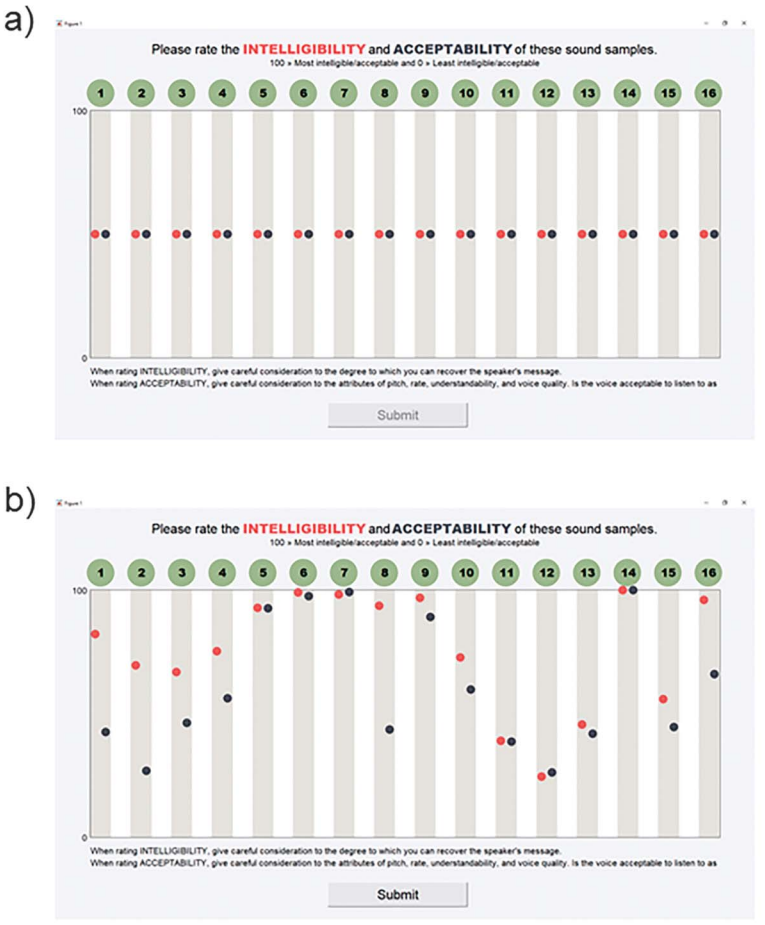
Speaker	Voice source	Abbreviation	No. of voices per source
Speaker with typical voice	Natural voice	NV	4
	Synthetic voice matched to control speaker	SV_C	4
Speaker with laryngectomy	EL voice	EV	4
	Synthetic voice matched to speaker with laryngectomy	SV_L	4

Note. NV = natural voice; SV_C = synthetic voice matched to speakers with typical voices; EL = electrolaryngeal; EV = electrolaryngeal voice; SV_L = synthetic voice matched to speakers with laryngectomy.

the "quality" of each speech token. Intelligibility, on the other hand, was selected to examine the "recognizability" of a speech token (Hustad et al., 1998; Lindblom, 1990; Meltzner & Hillman, 2005; Tjaden et al., 2014; Yorkston et al., 1996). This paradigm was chosen to allow listeners to hear multiple speech tokens in a single set for direct comparison against one another.

As depicted in Figure 3, the visual sort-and-rate paradigm was presented using a custom MATLAB (The MathWorks) graphical user interface (GUI) that enabled listeners to play a speech token produced in each of the 16 voices (see Table 3). Each listener was instructed to play the 16 speech tokens (labeled 1–16 in Figure 3) and then sort each of the tokens along the vertical axis (0–100) in terms of intelligibility and acceptability. Intelligibility was rated for each token by moving a red indicator along the vertical axis, with the most intelligible tokens placed at the top (100; anchored as "Most Intelligible") and the

Figure 3. An example of the visual sort-and-rate graphical user interface (a) at initialization and (b) after a listener rates the intelligibility and acceptability of each sample. Sound clips are denoted by the green circles (1 phrase × 16 voice sources). Red circles represent intelligibility ratings, and blue circles represent acceptability ratings.



least intelligible at the bottom (0; anchored as "Least Intelligible"). Instructions for judging speech intelligibility were adapted from methods by Kent et al. (1989) to consider "the degree to which a speaker's message can be recovered by a listener" and were provided at the bottom of the GUI for the duration of the study. Acceptability was rated for each token by moving a blue indicator along the vertical axis, with the most acceptable tokens placed at the top (100; anchored as "Most Acceptable") and the least acceptable at the bottom (0; anchored as "Least Acceptable"). Instructions for judging acceptability were based on methods from Bennett and Weinberg (1973) to "give careful consideration to the attributes of pitch, rate, understandability, and voice quality" and were described at the bottom of the GUI for the duration of the study.

In the visual sort-and-rate task, listeners could play each speech token as many times as needed and were not directed to the order in which to rate acceptability or intelligibility. After rating the intelligibility and acceptability of each of the 16 voices (4 voice sources × 4 speakers) of the same speech token, listeners were instructed to play the tokens again and rate the tokens against each other in order to make small adjustments as necessary. Once listeners were comfortable with their responses, they were instructed to lock in their answers and move to the next token.

The first three speech tokens were repeated across all listeners to act as a training module to familiarize listeners with the four voice sources, the concepts of acceptability and intelligibility, and the visual sort-and-rate paradigm. After completing the intelligibility and acceptability assessments for the training speech tokens, listeners were instructed to rate a series of 12 unique speech tokens (1 phrase × 4 voice sources × 3 phrasal stress types) that were pseudorandomly assigned to each listener from the 36-token speech bank described in the sEMG-Based Text-to-Speech section. After completing the 12 speech tokens, three tokens (25%; 1 token/stress type) were randomly repeated to assess intralistener reliability, for a total of 15 speech tokens.

Phrasal Stress Classification Task

In the phrasal stress classification task, listeners were instructed to play each speech token and determine the location of phrasal stress (first word, last word, no stress) as per Patel and Campellone (2009). Three multiple-choice options were presented to the listener: (a) "first word," indicating that the first word of the phrase was produced with the greatest stress; (b) "last word," indicating that the last word of the phrase was produced with the greatest stress; and (c) "no stress," indicating that none of the words in the phrase were stressed. Listeners were also presented with the transcript of the token to minimize the potential confound of intelligibility affecting listener responses. The transcript was always presented without any indication of phrasal stress so as not to bias listener responses (e.g., the token "JANE loves Bob"—wherein the first word of the phrase is stressed—would be shown as "Jane loves Bob"). Like the visual sort-and-rate task, listeners could play each speech token as many times as needed.

Each listener was pseudorandomly presented with 12 unique tokens produced in each of the 16 voices, for a total of 192 speech tokens (12 phrases × 16 voices). After completing the 192 speech tokens, 48 tokens (25%; 1 phrase × 3 stress types × 16 voices) were randomly repeated to assess intralistener reliability, for a total of 240 speech tokens. Tokens from each speaker were counterbalanced such that listeners were presented with four tokens per phrasal stress type.

Data Analysis

Three outcome measures were obtained from the two paradigms: Intelligibility and acceptability scores (0–100) were extracted directly from the listener visual sort-and-rate task, whereas phrasal stress classification accuracy (PSCA; 0 or 1) was calculated from the stress classification task. Intra- and interrater reliability measures were first computed for each task. Intralistener reliability was calculated on acceptability and intelligibility ratings using two-way mixed-effects intraclass correlation coefficients (ICCs) to assess absolute agreement on 25% of repeated data. Interrater reliability was computed on the intelligibility and acceptability ratings using a two-way random ICC to assess consistency of agreement. Listeners with poor reliability scores (ICC < .50) were removed from further analysis (Portney & Watkins, 2000). For acceptability, one listener was found to have poor reliability and was removed from further analyses; the average intrarater reliability across the remaining 11 listeners was calculated as ICC = .82 (*SD* = .07, range: .72–.91), and interrater reliability was calculated as ICC = .83 (*SD* = .05, range: .73–.89). For intelligibility, two listeners demonstrated poor reliability and were removed from further analysis. Mean intrarater reliability across the remaining 10 listeners for intelligibility was calculated as ICC = .75 (*SD* = .07, range: .65–.88), whereas interlistener reliability for the 10 listeners was calculated as ICC = .69 (*SD* = .07, range: .61–.81). Intralistener reliability was calculated on PSCA using Cohens' kappa, and three listeners with poor reliability scores ($\kappa < .41$) were removed from further analysis (McHugh, 2012). For PSCA, the average intrarater reliability across nine listeners was $\kappa = .64$ (*SD* = .14, range: .44–.90).

Welch's analysis of variance (ANOVA) tests were performed to evaluate overall differences in mean intelligibility and acceptability ratings between voice sources (i.e., NV, SV_C, EV, SV_L). For this analysis, a Welch's ANOVA was selected to overcome one-way ANOVA violations of homogeneity of variance across voice sources (Kohr & Games, 1974). An alpha level of .05 was used for significance testing, and effect sizes were calculated via adjusted omega-squared. Games–Howell post hoc tests were used to identify significant differences in mean intelligibility and acceptability ratings between voice sources. Chi-square tests of independence were then performed on PSCA data to examine the relationship between voice source (i.e., NV, SV_C,

Table 4. Word recognition rate and phrasal stress discriminability of speakers with typical voices (T) and speakers with laryngectomy (L).

Speaker	Word recognition rate (%)	Phrasal stress discriminability (%)
T1	95.4	99.0
T2	96.4	86.7
T3	96.3	90.0
T4	99.2	95.7
L1	96.2	86.8
L2	89.5	88.0
L3	99.0	90.0
L4	98.6	93.4

EV, SV_L) and phrasal stress discrimination. An alpha level of .05 was used for significance testing.

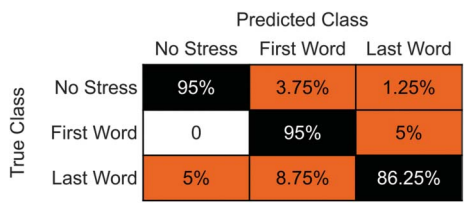
Results

Algorithmic Development

Mean WRR across the eight speakers was 97.4% (*SD* = 1.3%) for phrases without stress, 96.0% (*SD* = 4.8%) for phrases with first-word stress, and 95.7% (*SD* = 3.9%) for phrases with last-word stress. Table 4 shows the results of the sEMG-based word recognition algorithms for identifying lexical content and classifying phrasal stress. Cumulative WRR was 96.3% (*SD* = 3.1%) across the stress classes for all eight speakers, whereas mean phrasal stress discriminability was 91.2% (*SD* = 4.5%). Accuracy values were comparable between speakers with typical voices and speakers with laryngectomy for WRR (typical voice: 96.8%, laryngectomy: 95.8%) and phrasal stress discriminability (typical voice: 92.9%, laryngectomy: 89.6%).

Figure 4 shows the breakdown of phrasal stress discrimination between stress classes: first word, last word, and no stress. A total of 5% of speech tokens without stress were misclassified as containing stress on the first (3.75% of tokens) or last (1.25% of tokens) word. Similarly, speech tokens with first-word stress were confused for last-word stress 5% of the time. These tokens were never misclassified as no-stress tokens, however. Finally, speech tokens with stress on the last word were misclassified most often, with 5% of tokens erroneously detected to have no stress and 8.75% of tokens classified as having stress on the first word instead of the last word. Within a similar vein, the (macro/weighted) F1-score results demonstrated mean value of 0.92 (*SD* = 0.03). These values were also equally distributed across speakers with typical voices (*M* = 0.93) and speakers with laryngectomy (*M* = 0.91).

Figure 4. Confusion matrix for phrasal stress discrimination (no stress, first word, last word). Correctly classified speech tokens are denoted in black, whereas incorrectly classified tokens are shown in orange.



Perceptual Experiment

The results of the Welch's ANOVAs revealed a significant, large main effect of voice source on acceptability ($F = 1219$, $p < .001$, $\omega^2 = 0.63$) and intelligibility ($F = 542.0$, $p < .001$, $\omega^2 = 0.77$). Post hoc Games–Howell pairwise comparisons revealed that acceptability and intelligibility ratings of all voice sources were statistically different from each other ($p_{adj} < .05$), except between SV_C and SV_L sources (acceptability: $p_{adj} = .21$, intelligibility: $p_{adj} = .051$).

Figure 5 shows the mean acceptability (see Figure 5a) and intelligibility (see Figure 5b) ratings of speech tokens of each voice source. On average, NVs were rated as the most acceptable (*M* = 85.6%) and intelligible (*M* = 91.8%) when compared to SV and EV counterparts. However, synthetic voices (SV_C, SV_L) were rated higher in acceptability ($M_{SVC} = 59.7\%$, $M_{SVL} = 57.1\%$) and intelligibility

Figure 5. Mean (a) acceptability and (b) intelligibility ratings for speech tokens produced by natural voices (NV), synthetic voices matched to speakers with typical voices (SV_C), electrolaryngeal voices (EV), and synthetic voices matched to speakers with laryngectomy (SV_L). Error bars represent 95% confidence intervals. *$p < .05$.

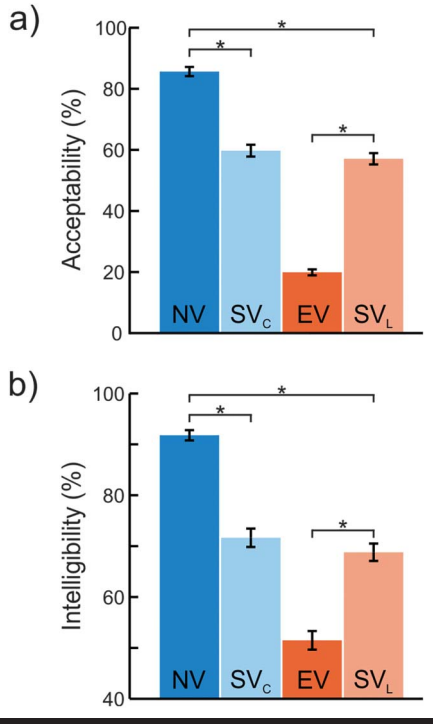
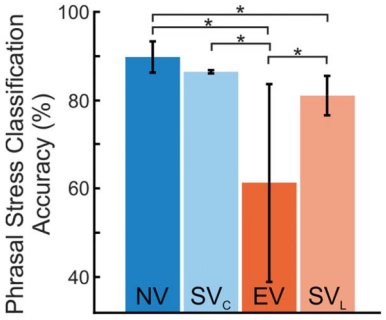


Figure 6. Mean phrasal stress classification accuracy across speakers for tokens produced by natural voices (NV), synthetic voices matched to speakers with typical voices (SV_C), electrolaryngeal voices (EV), and synthetic voices matched to speakers with laryngectomy (SV_L). Error bars are 95% confidence intervals. *$p < .05$.



(M_{SVC} = 71.7%, M_{SVL} = 68.8%) than EVs, which were rated as the least acceptable (M = 19.9%) and intelligible (M = 20.1%).

Chi-square tests of independence on PSCA revealed that NV and SV_C groups were not statistically significantly different from one another ($\chi^2 = 1.55$, $p = .213$). The PSCA of NV tokens was, however, significantly greater than that of EV ($\chi^2 = 88.2$, $p < .001$) and SV_L ($\chi^2 = 9.52$, $p = .002$) tokens. The chi-square analyses on PSCA also revealed that SV_C was significantly greater than EV ($\chi^2 = 68.8$, $p < .001$), but not statistically significantly different from SV_L ($\chi^2 = 3.45$, $p = .063$). Finally, SV_L was significantly greater than EL ($\chi^2 = 43.1$, $p < .001$).

Figure 6 displays the results of the multiple-choice task instructing listeners to classify phrasal stress in speech tokens (first word, last word, no stress). Mean PSCA was greatest when characterizing speech produced using an NV (M = 90.0%) compared to that of EL (M = 61.3%) and synthetic (M_{SVC} = 86.5%, M_{SVL} = 81.1%) voices.

Table 5 shows the average intelligibility, acceptability, and PSCA scores for each speaker when assessed in their primary voice source (natural, EL) and when synthesized into a personalized, prosodic voice. Intelligibility and acceptability were greatest for T1 when using their NV (intelligibility = 93.8%, acceptability = 88.3%). Of note, the speaker with laryngectomy who used dynamic pitch control (L2) elicited the lowest ratings of intelligibility (48.0%) and acceptability (17.5%), but the greatest PSCA score (95.1%).

Discussion

The purpose of the current study was to develop an sEMG-based SSR system that could identify words and classify basic manipulations in prosody (i.e., varying phrase-level stress), as well as to evaluate the acceptability, intelligibility, and phrasal stress discriminability of synthetic speech outputs. As prosody comprises numerous, complex interactions of pitch, timing, and intensity, the current study focused on evaluating "phrase-level stress" to determine initial feasibility of the proposed technology. Thus, the current study aimed to (a) improve existing word recognition algorithms described in prior works to identify phrasal stress and (b) synthesize correctly detected word content and stress location into audible voices capable of replicating the prosodic features of subvocal speech. Algorithms for recognizing words and categorizing phrasal stress were first assessed for accuracy. Accurately detected phrases were synthesized into audible speech using text-to-speech synthesis engines. Then, a perceptual listening experiment was performed to assess the acceptability, intelligibility, and PSCA of the synthetic speech. The results of this study serve as a proof of concept that sEMG-based speech recognition and synthesis can convey lexical content and distinguish phrase-level stress with greater fidelity than common EL devices.

Methodology for Eliciting Subvocal Speech

The cues used to elicit subvocal speech in the current study are not the only method of producing silent speech. Prior work has examined two alternative methods of subvocalizing, including mentally rehearsed (or "imagined") and silently mouthed speech. The former method consists of instructing speakers to mentally visualize the act of speaking aloud. Without speakers moving their articulators to produce the speech, however, the authors were unable to detect viable sEMG signals for use within the sEMG-based SSR

Table 5. Mean intelligibility, acceptability, and phrasal stress classification accuracy scores (PSCA) for speakers with typical voices (T) and speakers with laryngectomy (L) according to primary voice source (natural, electrolarynx, or electrolaryngeal [EL]) and the synthetic voice matched to the speaker.

	Primary voice source (natural, EL)				Synthetic voice		
ID	Voice source	Intelligibility	Acceptability	PSCA	Intelligibility	Acceptability	PSCA
T1	Natural	93.8	88.3	94.4	78.1	65.4	86.8
T2	Natural	92.3	88.0	86.1	73.7	60.8	86.8
T3	Natural	86.9	75.1	88.2	63.9	52.9	86.1
T4	Natural	93.2	85.7	90.3	70.9	56.3	86.1
L1	EL	54.3	21.8	53.5	72.7	58.9	78.5
L2	EL	48.0	17.5	95.1	72.5	90.0	77.1
L3	EL	51.3	21.6	45.8	66.6	51.5	81.3
L4	EL	52.4	19.3	50.7	63.3	50.1	87.5

system (Meltzner et al., 2008). The latter method, on the other hand, requires speakers to articulate the words of a message as they typically would, but without vocalizing. Using this method of eliciting silent speech, sEMG signals were detected that—when analyzed using the sEMG-based SSR system—led to an average WRR comparable to that of vocalized speech (98.3% using sEMG signals of vocalized speech vs. 96.7% using sEMG signals of silently mouthed speech; Meltzner et al., 2008).

As the goal of the current work was to use the sEMG signal to recognize lexical content (i.e., as in Meltzner et al., 2008, 2017, 2018) as well as categorize phrasal stress, a new approach to eliciting silent speech was considered. In particular, the MFCC features computed within the SSR algorithms were trained to detect phonemic content and stress content on a subject-specific, frame-by-frame basis. To reduce the impacts of coarticulation that may occur within sEMG signal frames, instructions to subvocally recite phrases were refined from prior instructions to silently mouth words to include cues to deliberately reduce their speaking rate and hyperarticulate. These specific techniques were chosen as they reflect similar instructions that are typically given to persons learning to use an EL following laryngectomy to improve speech clarity.

sEMG-Based Voice Synthesis

It was hypothesized that adapting existing speech recognition models to be robust to variations in phrasal stress (instead of only focusing on the phonemic content of the message) would enable the sEMG-based SSR system to accurately detect lexical content recited using prosodic features of speech. When tested within different speakers, the sEMG-based SSR algorithms were able to recognize subvocal speech tokens produced with and without phrasal stress with a mean word recognition accuracy of 96.3%. Word recognition accuracy was comparable across phrases produced with stress (96.0% for first-word stress, 95.7% for last-word stress) and without stress (97.4%). Prior work examining sEMG-based SSR in the absence of prosody reports differing results. Using a data set of 2,200–2,500 words, Meltzner et al. achieved an accuracy of 91.1% accuracy in 19 typical speakers (Meltzner et al., 2018) and an accuracy of 89.7% in eight laryngectomized speakers (Meltzner et al., 2017). In smaller data set of 108 words, however, Jou et al. (2006) achieved an accuracy of 68%. Few sEMG-based subvocal studies have examined WRR in the face of prosodic variations.

Thus far, the current study is the first to report > 95% WRR for both speakers with typical voices and speakers with laryngectomy. Improvements in average WRR may be thought of as a byproduct of the chosen speech corpus. In particular, the corpus examined here was derived from prior works investigating the ability of the sEMG-based SSR system to recognize lexical content within a large corpus of 2,200–2,500 words (approximately 1,200 phrases; Meltzner et al., 2017, 2018). Since the goal of this work was to modify the algorithms to effectively adapt to and classify variations in word production (i.e., phrase-level stress), the size of the speech corpus was reduced to minimize recording time. It is unlikely that the reduced corpus size led to increases in WRR, as previous work using the sEMG-based SSR system for lexical recognition within the same base corpus showed that WRR "decreased" with smaller corpus sizes (Meltzner et al., 2017). These increases in WRR may instead be the result of algorithmic modifications to include feature space transformation: By incorporating a fMLLR and training a speaker-adaptive model, the lexical recognition algorithms may have accounted for variations in word production from individual speakers to, in turn, increase WRR.

A fundamental advantage of this approach, when compared to other SSR systems described in the literature, is that it is based on sEMG signals recorded from muscles of the neck and face. These muscles play a fundamental role not only in speech production but also in conveying prosodic information through unique activation patterns. Similar technology developed by Janke and Diener (2017) demonstrated an average WRR of 92.7% across three speakers with typical voices. Other investigators have developed SSR using other modalities such as ultrasound of the articulators. For instance, previous work examining ultrasound of the lips and tongue demonstrated a WRR of 86.1% in a speech corpus of 20,000 words collected from one typical speaker (Cai et al., 2011). SSR has also been examined using EMA. Kim et al. (2017) examined the 132 EMA recordings collected from 12 typical speakers and two speakers with laryngectomy using a neural network, resulting in a WRR of 44.8%. More recently, Kim et al. (2018) achieved a WRR of 67.9% through EMA recordings of 65 words and 25 phrases. The systems examined in this study, which were trained on independent training (n = 650) and test (n = 150) sets in four speakers with typical voices and four speakers with laryngectomy, demonstrated a mean WRR of 96.3%. While direct comparisons are difficult to make since discrepancies in algorithmic accuracy could be a result of the different recording modalities (e.g., EMA vs. sEMG) and/or recognition algorithms, the results of this study lend support to the notion that sEMG-based SSR systems are capable of achieving good lexical recognition performance across speakers with typical voices and speakers with laryngectomy.

To discriminate phrase-level changes in stress, it was hypothesized that the spectral and temporal characteristics of the sEMG recordings during stressed words spoken would be different from the recordings in which words were spoken without phrasal stress. In line with this hypothesis, a three-class phrasal stress (first word, last word, no stress) classifier was designed with quantitative temporal and spectral features extracted on a window-based segmentation of the recorded phrases. The features included in this classifier comprised 10 MFCC features originally used to recognize lexical content, as well as two TWL features, two SWL features, and one PCC feature. The results of this classifier demonstrated success in discerning whether phrase-level stress was included within the token, and if so, whether the first or last

word was stressed. Using this classifier, phrasal stress was categorized at an average accuracy of 91.2%.

It is difficult to directly compare the phrasal stress classification results from the sEMG-based SSR system with those of other works due to differences in recording and analysis methods. For instance, Gonzalez et al. (2017) used audible speech signals obtained in conjunction with PMA recordings to validate PMA-based measures of voice f0. A similar evaluation technique was carried out by Diener et al. (2019) to evaluate their sEMG-based SSR system, wherein audible speech recordings were used to validate sEMG-derived metrics of voicing. Audible speech recordings were not utilized in this study since acoustic training data could not be collected from the speakers with laryngectomy, as all speakers were enrolled at least 1 year after their operation. Of note, however, a study performed by Johner et al. (2012) for sEMG-based detection of emphasized words in a complete sentence achieved an F1 score of .68, with large variations over different speakers. When translated in the current study, the F1 score for phrasal stress detection resulted in a value of .92. Although this work shows promise for discriminating prosodic manipulations in subvocal speech, the method of introducing these manipulations into the subvocal speech corpus used here remains primitive (i.e., eliciting phrase-level stress from speakers). As such, future work is needed to investigate the ability of the sEMG-based SSR algorithm to identify more complex prosodic manipulations in prosody.

Listener Perceptual Assessment of Synthetic Speech

The purpose of the listener perceptual experiment was to evaluate the reception of synthetic speech outputs produced by the sEMG-based SSR system. Specifically, correctly recognized phrases were synthesized into audible speech using text-to-speech engines; this synthesis was performed by introducing static representations of pitch, loudness, and duration derived from the SSR algorithms to inform the Personifier speech engine (VocaliD) as to whether phrasal stress should be imparted on the first or last word. This process resulted in sets of speech tokens with and without phrasal stress, from which it was possible to assess how naïve listeners perceived the synthetic speech as compared to natural and EL speech. Three performance metrics were chosen to assess the listener perceptions of synthetic speech: acceptability, intelligibility, and PSCA.

Speech Acceptability and Intelligibility

One of the goals of the current study was to synthesize speech that effectively conveyed lexical content and incorporated prosodic cues found in natural speech. As such, it was hypothesized that the acceptability and intelligibility of synthetic speech would be significantly greater than that of speech produced using an EL. Even though EL speech aids are a common communication method for laryngectomized speakers, one aim of the current study was to synthesize speech that would be more similar in its listener reception to natural speech, including quality and clarity. The results obtained in this work support this hypothesis, showing that EL speech was rated as significantly less acceptable and intelligible than natural and synthetic counterparts. These findings are consistent with others who have shown poor intelligibility and/or acceptability of EL speech compared to natural speech (Bennett & Weinberg, 1973) and other forms of alaryngeal communication (e.g., tracheoesophageal speech; Bennett & Weinberg, 1973; Eadie et al., 2013, 2016; Law et al., 2009; Williams & Watson, 1985), likely due to the mechanical nature of the EL. Most importantly, these results suggest that the method of text-to-speech synthesis implemented in the current study may be a promising alternative to using an EL to produce recognizable speech that is "acceptable" to the listener.

Speech tokens produced by typical speakers were also assessed in the current study to provide a gold standard of intelligibility and acceptability for speech. The results of this work demonstrated that natural speech was significantly more intelligible and acceptable than synthetic speech, suggesting that speech performance when using a synthetic source was lower in both clarity and quality compared to natural speech. The methods used to extract measures of intelligibility and acceptability should be discussed. Although the gold standard practice for measuring intelligibility is through orthographic transcription (Miller, 2013; Weismer, 2006; Yorkston & Beukelman, 1981), visual analog scales within a visual sort-and-rate task were implemented in the current study. This method was selected to minimize the potential for ceiling effects to occur in listener orthographic transcriptions, as each of the 16 voices produced the same 12 speech tokens. Prior work has shown a strong relationship between visual analog scale ratings and orthographic transcription measures of intelligibility in Parkinson's disease (Abur et al., 2019; Stipancic et al., 2016), but it is unclear how differences in methodology affect the intelligibility of the voice sources analyzed in the current study. Moreover, it is possible that ratings of speech acceptability were confounded by intelligibility in the current study. The definition of acceptability provided to listeners encompassed multiple aspects of speech quality, including pitch, quality, rate, and "understandability." Despite efforts to distinguish intelligibility and acceptability in the current study, similar trends were observed between intelligibility and acceptability results, with NV tokens ranking higher than SV, followed by EL. As such, it may be speculated that listeners perceived some tokens as less acceptable than others as a result of an interaction between speech quality and recognizability.

PSCA

In assessing the ability of listeners to classify the presence and location of phrasal stress, it was hypothesized that classification accuracy of synthetic speech tokens would not be statistically significantly different from natural speech tokens but would be significantly greater that EL speech tokens. The PSCA of natural and synthetic speech tokens was not significantly different, but only when the selected synthetic voices were age- and sex-matched to speakers with

typical voices; it is possible that these results were confounded by unequally distributed age and sex demographics across the groups. In particular, the group of speakers with typical voices comprised three women and one man who averaged to be 25.7 years of age, whereas the group of speakers with laryngectomy comprised three male speakers, of whom averaged 70.3 years in age. The discrepancy in these demographics may have biased listener perceptions of the selected synthetic voices; however, it is important to note that the small sample sizes included within the current study make it difficult to identify factors that may have contributed to these detriments. As such, a follow-up study should be conducted to identify potential relationships between listener perceptions of these performance metrics and features that comprise the synthetic speech signal as compared to the natural speech signal (e.g., concatenation of phonemes).

Listener accuracy in classifying phrasal stress (first word, last word, no stress) was lowest on average for EL speech, suggesting that this method of communication was less effective in conveying stress than natural or synthetic speech. This is not surprising given the monotonic nature of EL speech produced by three of the four speakers with laryngectomy in this study and previous research demonstrating that vocal stress patterns, like those used in this study, are extremely difficult to convey when using an EV prosthesis (Gandour & Weinberg, 1984). Taken together, these findings provide preliminary evidence to suggest that the sEMG-based synthetic speech is more effective in conveying phrase-level stress than EL alternatives and, moreover, that it is more similar to natural speech than EL speech.

Effect of Electrolarynx on Listener Reception of Speech

Of the four speakers who used an EL in the current study, only one speaker used an EL set to provide noticeable dynamic pitch control for their daily communication. Likewise, this was the only speaker who felt comfortable using an EL with dynamic pitch capabilities during the study. Although this speaker conveyed phrasal stress to an accuracy that was greater than the other speakers with laryngectomy and all speakers with typical voices, this speaker also performed the worst in terms of perceived intelligibility and acceptability. These results would suggest that there are artificial means for conveying phrasal stress (e.g., hand-controlled), yet the goal of this work was to do so in a way that is reflective of the intelligibility and acceptability of natural vocalizations using articulation-related muscle activity. On account of this, it is not surprising then the three speakers who were accustomed to monotonic EL speech elected not to use an EL with a more dynamic pitch range, despite being shown how pitch control can be achieved through activation-button-pressure variation. These speakers instead indicated a preference for using a monopitch EL due to reported issues regarding the dynamic pitch modulation feature being too unwieldly, too cognitively complex to control in real time, and/or not representative of how the speaker envisions their voice.

Limitations and Future Directions

The current work provides preliminary evidence of detecting prosody from subvocal speech. Rather than specifically instructing speakers on how to modify the suprasegmental characteristics of their subvocal speech, however, the prosodic changes examined in this study were based on each speaker's interpretation of how to convey phrasal stress. Nonspecific changes in prosody were examined across phrases to determine the feasibility of detecting suprasegmental characteristics that differ from typical (nonstressed) productions using sEMG activity. It is possible that the method used to induce phrasal stress may have differed not only among speakers but also across productions of the same phrase. Nonetheless, the results of this study provide a foundation for further investigations into segmental and suprasegmental variations in subvocalized phrases.

Contrary to the complex manipulations in linguistic prosody that are used in the English language, only basic manipulations to prosody were assessed in the current study. The sEMG features used to recognize lexical content and categorize phrasal stress were not implemented with the purpose of identifying continuous changes in suprasegmental attributes such as pitch, duration, and loudness; instead, static representations of how these attributes may change with the introduction of phrasal stress were used to inform the text-to-speech engine to alter the prosodic characteristics of synthetic speech. Although exploratory in nature, this study serves as a proof of concept that SSR is possible in the absence of an acoustic signal when coupled with basic prosodic manipulations. The findings of this work support further endeavors that expand the scope of SSR and synthesis to include a variety of prosodic contrasts.

The corpus examined in this work was substantially smaller in size than those previously examined in the development of the sEMG-based SSR system. In particular, Meltzner et al. (2018) assessed word recognition accuracy in a test set comprising 430 unique tokens, whereas that of the current work examined word recognition accuracy in a test set of 150 tokens (50 unique) and PSCA in a test set of 30 tokens (10 unique). The difference in size of the test sets from our prior work and the current study were driven by practical considerations of the experimental methodology, such that a number and combination of phrases with varied lexical content and phrasal stress could be evaluated without causing fatigue and associated degradations to subvocal performance. Even with the reduced data set, the relatively large number of combinations of phonemic content, phrasal stress, voice sources, and speakers per voice source precluded the use of orthographic transcription to evaluate speech intelligibility within a reasonable amount of time while avoiding possible listener fatigue. Instead, we chose to implement a visual sort-and-rate paradigm—based, in part, on prior reports of visual analog scales providing comparable outcomes to orthographic transcriptions in certain scenarios (e.g., Abur et al., 2019)—to provide ratings of speech intelligibility within a reasonable amount of time. The intelligibility results in the current study were

therefore evaluated for each voice on a relative scale with respect to 15 additional voice samples (four NV, eight SV, four EV).

The differential contributions of individual muscle sites to recognizing lexical content and categorizing phrasal stress were not examined in this work. Although it is possible to draw conclusions as to how sEMG signals from different sensors contributed to the manifestation of subvocalized, phrasal stress within each speaker, the current study on a nonreduced sensor set to discriminate phrasal stress across the data set. By demonstrating the proof of concept that discriminability is possible by leveraging sEMG activity across the full eight-sensor set, future work should aim to assess the effects of sensor reduction on recognizing both lexical and prosodic content.

It is also worth noting that the populations examined in the current study were not necessarily representative of all speakers with or without laryngectomy. All individuals who relied on alaryngeal speech within the current study were survivors of laryngeal cancer, yet the need for augmentative and alternative methods of communication is not specific to those who have laryngeal cancer. Moreover, none of the speakers with laryngectomy in the current study were female due to the relatively small sample size and the greater prevalence of laryngeal cancer in men (3:1 male-to-female ratio; Lewin, 2004). Future investigations could be generalized to include female speakers and those with other cancers of the head and neck that impair voice/speech.

Finally, while speakers with laryngectomy could speak in real time when using their EL speech aid, the sEMG-based ASR techniques examined in the current study were not yet capable of real-time functionality. Specifically, this study was a proof of concept to determine the feasibility of developing a system, wherein the two components necessary for sEMG-based ASR (i.e., word/prosody recognition and text-to-speech processing) were decoupled in the current study. As such, EL speech may be preferred to a better sounding yet delayed synthetic proxy of natural speech. Regardless, the results of the current study are still promising. Future work should continue to work toward a real-time sEMG-based ASR system.

Conclusions

sEMG-based speech recognition and synthesis is a feasible technique for communicating lexical content and basic phrase-level stress of subvocal (i.e., silently mouthed) speech. Algorithms for recognizing subvocal speech show promise for not only identifying the words in a message but also for distinguishing whether the message contains phrasal stress at the beginning or end of the message. Upon synthesizing these messages into speech via personalized text-to-speech, lexical and phrasal stress characteristics were found to be more intelligible and acceptable than speech produced using current, state-of-the-art EL devices. These findings lay a solid foundation for investigating the detection and reception of complex prosodic interactions. As such, future work should aim to control prosodic variations to differentially examine the impacts of pitch, loudness, and timing variations to SSR and synthesis.

Acknowledgments

This work was supported in part by a grant from the National Institutes of Health under Grant R43 DC017097 (awarded to Altec, Inc.) and by the De Luca Foundation.

References

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

Adler, J. J., & Zeides, J. (1986). Evaluation of the electrolarynx in the short-term hospital setting. *Chest, 89*(3), 407–409. https://doi.org/10.1378/chest.89.3.407

Ahsan, M. R., Ibrahimy, M. I., & Khalifa, O. O. (2011). Neural network classifier for hand motion detection from EMG signal. In N. A. A. Osman, W. A. B. W. Abas, A. K. A. Wahab, & H.-N. Tin (Eds.), *5th Kuala Lumpur International Conference on Biomedical Engineering 2011 (BIOMED 2011)* (pp. 536–541). Springer.

Anastasakos, T., McDonough, J., Schwartz, R., & Makhoul, J. (1996). A compact model for speaker-adaptive training. In H. T. Bunnell, & W. Idsardi (Eds.), *Proceedings of 4th International Conference on Spoken Language Processing* (pp. 1137–1140). IEEE.

Bennett, S., & Weinberg, B. (1973). Acceptability ratings of normal, esophageal, and artificial larynx speech. *Journal of Speech and Hearing Research, 16*(4), 608–615. https://doi.org/10.1044/jshr.1604.608

Betts, B. J., & Jorgensen, C. (2005). *Small vocabulary recognition using surface electromyography in an acoustically harsh environment* (Vol. 16). NASA's Ames Research Center.

Bolinger, D. L. (1958). A theory of pitch accent in English. *WORD, 14*(1–2), 109–149. https://doi.org/10.1080/00437956.1958.11659660

Boucher, V. J., Ahmarani, C., & Ayad, T. (2006). Physiologic features of vocal fatigue: Electromyographic spectral-compression in laryngeal muscles. *The Laryngoscope, 116*(6), 959–965. https://doi.org/10.1097/01.MLG.0000216824.07244.00

Broniatowski, M., Sonies, B. C., Rubin, J. S., Bradshaw, C. R., Spiegel, J. R., Bastian, R. W., & Kelly, J. H. (1999). Current evaluation and treatment of patients with swallowing disorders. *Otolaryngology—Head & Neck Surgery, 120*(4), 464–473. https://doi.org/10.1053/hn.1999.v120.a93228

Brumberg, J. S., Guenther, F. H., & Kennedy, P. R. (2013). An auditory output brain–computer interface for speech communication. In C. Guger, B. Allison, & G. Edlinger (Eds.), *Brain-computer interface research* (pp. 77–114). Springer: Briefs in Electrical and Computer Engineering.

Cai, J., Denby, B., Roussel-Ragot, P., Dreyfus, G., & Crevier-Buchman, L. (2011). Recognition and real time performance of a lightweight ultrasound based silent speech interface employing a language model. In P. Cosi, R. De Mori, G. Di Fabbrizio, & R. Pieraccini (Eds.), *INTERSPEECH 2011* (pp. 1005–1008). ISCA Archive.

Costello, J. M. (2014). *Message banking, voice banking and legacy messages.* Boston Children's Hospital.

Cox, S. R., & Doyle, P. C. (2018). The influence of clear speech on auditory-perceptual judgments of electrolaryngeal speech. *Journal of Communication Disorders, 75,* 25–36. https://doi.org/10.1016/j.jcomdis.2018.06.003

Crevier-Buchman, L., Gendrot, C., Denby, B., Pillot-Loiseau, C., Roussel, P., Colazo-Simon, A., & Dreyfus, G. (2011). Articulatory strategies for lip and tongue movements in silent versus vocalized speech. In W. S. Lee, & E. Zee (Eds.), *17th International Congress of Phonetic Science* (pp. 532–536). International Phonetic Association.

De Luca, C. J., Kuznetsov, M., Gilmore, L. D., & Roy, S. H. (2012). Inter-electrode spacing of surface EMG sensors: Reduction of crosstalk contamination during voluntary contractions. *Journal of Biomechanics, 45*(3), 555–561. https://doi.org/10.1016/j.jbiomech.2011.11.010

De Luca, C. J., & Mambrito, B. (1987). Voluntary control of motor units in human antagonist muscles: Coactivation and reciprocal activation. *Journal of Neurophysiology, 58*(3), 525–542. https://doi.org/10.1152/jn.1987.58.3.525

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J. M., & Brumberg, J. S. (2010). Silent speech interfaces. *Speech Communication, 52*(4), 270–287. https://doi.org/10.1016/j.specom.2009.08.002

Diedrich, W. M., & Youngstrom, K. A. (1966). *Alaryngeal speech*. Charles C. Thomas.

Diener, L., Umesh, T., & Schultz, T. (2019). Improving fundamental frequency generation in emg-to-speech conversion using a quantization approach. In H. Li (Ed.), *Proceedings of the ASRU 2019 IEEE Automatic Speech Recognition and Understanding Workshop* (pp. 682–689). IEEE.

Ding, R., Larson, C. R., Logemann, J. A., & Rademaker, A. W. (2002). Surface electromyographic and electroglottographic studies in normal subjects under two swallow conditions: Normal and during the Mendelsohn manuever. *Dysphagia, 17*(1), 1–12. https://doi.org/10.1007/s00455-001-0095-3

Doyle, P. C., & Eadie, T. L. (2005). The perceptual nature of alaryngeal voice and speech. In P. C. Doyle, & R. L. Keith (Eds.), *Contemporary considerations in the treatment and rehabilitation of head and neck cancer: Voice, speech, and swallowing* (pp. 113–140). Pro-Ed.

Drager, K. D., Reichle, J., & Pinkoski, C. (2010). Synthesized speech output and children: A scoping review. *American Journal of Speech-Language Pathology, 19*(3), 259–273. https://doi.org/10.1044/1058-0360(2010/09-0024)

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification*. Wiley.

Eadie, T. L., Day, A. M. B., Sawin, D. E., Lamvik, K., & Doyle, P. C. (2013). Auditory-perceptual speech outcomes and quality of life after total laryngectomy. *Otolaryngology—Head & Neck Surgery, 148*(1), 82–88. https://doi.org/10.1177/0194599812461755

Eadie, T. L., Otero, D., Cox, S., Johnson, J., Baylor, C. R., Yorkston, K. M., & Doyle, P. C. (2016). The relationship between communicative participation and postlaryngectomy speech outcomes. *Head & Neck, 38*(S1), E1955–E1961. https://doi.org/10.1002/hed.24353

Englehart, K., & Hudgins, B. (2003). A robust, real-time control scheme for multifunction myoelectric control. *IEEE Transactions on Biomedical Engineering, 50*(7), 848–854. https://doi.org/10.1109/TBME.2003.813539

Eskes, M., Van Alphen, M. J. A., Balm, A. J. M., Smeele, L. E., Brandsma, D., & Van Der Heijden, F. (2017). Predicting 3D lip shapes using facial surface EMG. *PLOS ONE, 12*(4), e0175025. https://doi.org/10.1371/journal.pone.0175025

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. https://doi.org/10.1044/1092-4388(2006/099)

Fagan, M. J., Ell, S. R., Gilbert, J. M., Sarrazin, E., & Chapman, P. M. (2008). Development of a (silent) speech recognition system for patients following laryngectomy. *Medical Engineering & Physics, 30*(4), 419–425. https://doi.org/10.1016/j.medengphy.2007.05.003

Fry, D. B. (1955). Duration and intensity as acoustic correlates of linguistic stress. *The Journal of the Acoustical Society of America, 35,* 765–769. https://doi.org/10.1121/1.1908022

Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech, 1,* 126–152. https://doi.org/10.1177/002383095800100207

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. https://doi.org/10.1080/07434619512331277209

Gandour, J., & Weinberg, B. (1984). Production of intonation and contrastive stress in electrolaryngeal speech. *Journal of Speech and Hearing Research, 27*(4), 605–612. https://doi.org/10.1044/jshr.2704.605

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. https://doi.org/10.1016/S0021-9924(02)00064-3

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., & Zue, V. (1993). *TIMIT acoustic–phonetic continuous speech corpus*. Linguistic Data Consortium.

Goldstein, E. A., Heaton, J. T., Kobler, J. B., Stanley, G. B., & Hillman, R. E. (2004). Design and implementation of a hands-free electrolarynx device controlled by neck strap muscle electromyographic activity. *IEEE Transactions on Biomedical Engineering, 151*(2), 325–332. https://doi.org/10.1109/TBME.2003.820373

Goldstein, E. A., Heaton, J. T., Stepp, C. E., & Hillman, R. E. (2007). Training effects on speech production using a hands-free electromyographically controlled electrolarynx. *Journal of Speech, Language, and Hearing Research, 50*(2), 335–351. https://doi.org/10.1044/1092-4388(2007/024)

Gonzalez, J. A., Cheah, L. A., Gomez, A. M., Green, P. D., Gilbert, J. M., Ell, S. R., Moore, R. K., & Holdsworth, E. (2017). Direct speech reconstruction from articulatory sensor data by machine learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25*(12), 2362–2374. https://doi.org/10.1109/TASLP.2017.2757263

Granqvist, S. (2003). The visual sort and rate method for perceptual evaluation in listening tests. *Logopedics, Phoniatrics, Vocology, 128*(3), 109–116. https://doi.org/10.1080/14015430310015255

Guenther, F. H., Brumberg, J. S., Wright, E. J., Nieto-Castanon, A., Tourville, J. A., Panko, M., Law, R., Siebert, S. A., Bartels, J. L., & Andreasen, D. S. (2009). A wireless brain-machine interface for real-time speech synthesis. *PLOS ONE, 4*(12), e8218. https://doi.org/10.1371/journal.pone.0008218

Harris, K. S. (1978). Vowel duration change and its underlying physiological mechanisms. *Language and Speech, 21,* 354–361. https://doi.org/10.1177/002383097802100410

Heaton, J. T., Goldstein, E. A., Kobler, J. B., Walsh, M. J., Zeitels, S. M., Gooey, J. E., Randolph, G. W., & Hillman, R. E. (2004). Surface electromyographic activity in total laryngectomy patients following laryngeal nerve transfer to neck strap muscles. *Annals of Otology, Rhinology & Laryngology, 113*(9), 754–764. https://doi.org/10.1177/000348940411300915

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

Hermens, H. J., Freriks, B., Disselhorst-Klug, C., & Rau, G. (2000). Development of recommendations for SEMG sensors and sensor placement procedures. *Journal of Electromyography and Kinesiology, 10*(5), 361–374. https://doi.org/10.1016/S1050-6411(00)00027-4

Hillman, R. E., Walsh, M. J., & Heaton, J. T. (2005). Laryngectomy speech rehabilitation: A review of outcomes. *Contemporary considerations in the treatment and rehabilitation of head and neck cancer: Voice, speech, and swallowing* (pp. 75–90). Pro-Ed.

Hočevar-Boltežar, I., & Žargi, M. (2001). Communication after laryngectomy. *Radiology and Oncology, 35*(4), 249–254.

Honda, K., Hirai, H., Masaki, S., & Shimada, Y. (1999). Role of vertical larynx movement and cervical lordosis in F0 control. *Language and Speech, 42*(4), 401–411. https://doi.org/10.1177/00238309990420040301

Hudgins, B., Parker, P., & Scott, R. N. (1993). A new strategy for multifunction myoelectric control. *IEEE Transactions on Biomedical Engineering, 40*(1), 82–94.

Hueber, T., Benaroya, E.-L., Chollet, G., Denby, B., Dreyfus, G., & Stone, M. (2010). Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. *Speech Communication, 52*(4), 288–300. https://doi.org/10.1016/j.specom.2009.11.004

Hustad, K. C., Beukelman, D. R., & Yorkston, K. M. (1998). Functional outcome assessment in dysarthria. *Seminars in Speech and Language, 19*(3), 291–302. https://doi.org/10.1055/s-2008-1064051

Janke, M., & Diener, L. (2017). EMG-to-speech: Direct generation of speech from facial electromyographic signals. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25*(12), 2375–2385. https://doi.org/10.1109/TASLP.2017.2738568

Johner, C., Janke, M., Wand, M., & Schultz, T. (2012). Inferring prosody from facial cues for emg-based synthesis of silent speech. In *Proceedings of the 4th International Conference on Applied Human Factors & Ergonomics* (pp. 5317–5326). Applied Human Factors and Ergonomics.

Jorgensen, C., Lee, D. D., & Agabont, S. (2003). Sub auditory speech recognition based on EMG signals. In *Proceedings of the International Joint Conference on Neural Networks* (pp. 3128–3133). IEEE.

Jou, S.-C., Schultz, T., Walliczek, M., Kraft, F., & Waibel, A. (2006). Towards continuous speech recognition using surface electromyography, In R. M. Stern (Ed.), *INTERSPEECH 2006—ICSLP* (pp. 573–576). International Speech Communication Association.

Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., van den Oord, A., Dieleman, S., & Kavukcuoglu, K. (2018). Efficient neural audio synthesis. arXiv: 180208435.

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. https://doi.org/10.1044/jshd.5504.751

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

Kim, M., Cao, B., Mau, T., & Wang, J. (2017). Speaker-independent silent speech recognition from flesh-point articulatory movements using an LSTM neural network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25*(12), 2323–2336. https://doi.org/10.1109/TASLP.2017.2758999

Kim, M., Sebkhi, N., Cao, B., Ghovanloo, M., & Wang, J. (2018). Preliminary test of a wireless magnetic tongue tracking system for silent speech interface. In P. Mohseni, & D. Tyler (Eds.), *2018 IEEE Biomedical Circuits and Systems Conference* (pp. 1–4). IEEE.

Kingma, D. P., & Ba, J. (2014). ADAM: A method for stochastic optimization. arXiv: 1412.6980.

Kohr, R. L., & Games, P. A. (1974). Robustness of the analysis of variance, the Welch procedure and a box procedure to heterogeneous variances. *The Journal of Experimental Education, 43*(1), 61–69. https://doi.org/10.1080/00220973.1974.10806305

Kramp, B., & Dommerich, S. (2009). Tracheostomy cannulas and voice prosthesis. *GMS Current Topics Otorhinolaryngology Head and Neck Surgery, 8*. Doc05

Krause, J. C., & Braida, L. D. (2002). Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility. *The Journal of the Acoustical Society of America, 112*(5), 2165–2172. https://doi.org/10.1121/1.1509432

Kubert, H. L., Stepp, C. E., Zeitels, S. M., Gooey, J. E., Walsh, M. J., Prakash, S. R., Hillman, R. E., & Heaton, J. T. (2009). Electromyographic control of a hands-free electrolarynx using neck strap muscles. *Journal of Communication Disorders, 42*(3), 211–225. https://doi.org/10.1016/j.jcomdis.2008.12.002

Lam, J., & Tjaden, K. (2013). Intelligibility of clear speech: Effect of instruction. *Journal of Speech, Language, and Hearing Research, 56*(5), 1429–1440. https://doi.org/10.1044/1092-4388(2013/12-0335)

Law, I. K. Y., Ma, E. P. M., & Yiu, E. M. L. (2009). Speech intelligibility, acceptability, and communication-related quality of life in Chinese alaryngeal speakers. *Otolaryngology—Head & Neck Surgery, 135*(7), 704–711.

Lee, K.-S. (2008). EMG-based speech recognition using hidden Markov models with global control variables. *IEEE Transactions on Biomedical Engineering, 55*(3), 930–940. https://doi.org/10.1109/TBME.2008.915658

Lewin, J. S. (2004). Advances in alaryngeal communication and the art of tracheoesophageal voice restoration. *The ASHA Leader, 9*(1), 6–21. https://doi.org/10.1044/leader.FTR1.09012004.6

Lindblom, B. (1990). On the communication process: Speaker-listener interaction and the development of speech. *Augmentative and Alternative Communication, 6*(4), 220–230. https://doi.org/10.1080/07434619012331275504

Liu, Z. T., Xie, Q., Wu, M., Cao, W. H., Mei, Y., & Mao, J. W. (2018). Speech emotion recognition based on an improved brain emotion learning model. *Neurocomputing, 309,* 145–156. https://doi.org/10.1016/j.neucom.2018.05.005

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. https://doi.org/10.1590/S2317-17822013.05000011

Maier-Hein, L., Metze, F., Schultz, T., & Waibel, A. (2005). Session independent non-audible speech recognition using surface electromyography. In J. Glass, & R. Rose (Eds.), *Proceedings of the ASRU 2005 IEEE Automatic Speech Recognition and Understanding Workshop* (pp. 331–336). IEEE.

Manabe, H., & Zhang, Z. (2004). Multi-stream HMM for EMG-based speech recognition. In D. Hudson (Ed.), *EMBC 2004: 26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Conference Proceedings* (pp. 4389–4392). IEEE.

McCall, F., Markova, 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. https://doi.org/10.1080/13682829709177099

McClean, M. D. (2000). Patterns of orofacial movement velocity across variations in speech rate. *Journal of Speech, Language, and Hearing Research, 43*(1), 205–216. https://doi.org/10.1044/jslhr.4301.205

McClean, M. D., & Tasko, S. M. (2002). Association of orofacial with laryngeal and respiratory motor output during speech. *Experimental Brain Research, 146*(4), 481–489. https://doi.org/10.1007/s00221-002-1187-5

McClean, M. D., & Tasko, S. M. (2003). Association of orofacial muscle activity and movement during changes in speech rate and intensity. *Journal of Speech, Language, and Hearing Research, 46*(6), 1387–1400. https://doi.org/10.1044/1092-4388(2003/108)

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemical Medicine, 22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

Meltzner, G. S., Heaton, J. T., Deng, Y., De Luca, G., Roy, S. H., & Kline, J. C. (2018). Development of sEMG sensors and algorithms for silent speech recognition. *Journal of Neural Engineering, 15*(4), 046031. https://doi.org/10.1088/1741-2552/aac965

Meltzner, G. S., Heaton, J. T., Deng, Y., De Luca, G., Roy, S. H., & Kline, J. C. (2017). Silent speech recognition as an alternative communication device for persons with laryngectomy. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25*(12), 2386–2398. https://doi.org/10.1109/TASLP.2017.2740000

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. https://doi.org/10.1044/1092-4388(2005/053)

Meltzner, G. S., Hillman, R. E., Heaton, J. T., Houston, K. M., Kobler, J. B., & Qi, Y. (2005). Electrolaryngeal speech: The state of the art and future directions for development. In *Contemporary considerations in the treatment and rehabilitation of head and neck cancer: Voice, speech, and swallowing* (pp. 571–590). Pro-Ed.

Meltzner, G. S., Sroka, J., Heaton, J. T., Gilmore, L. D., Colby, G., Roy, S. H., Chen, N., & De Luca, C. J. (2008). Speech recognition for vocalized and subvocal modes of production using surface EMG signals from the neck and face. In H. Fujisaki (Ed.), *INTERSPEECH 2008* (pp. 2667–2670). ISCA.

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

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication, 9*(5), 453–467. https://doi.org/10.1016/0167-6393(90)90021-Z

Palmer, P. M., Luschei, E. S., Jaffe, D., & McCulloch, T. M. (1999). Contributions of individual muscles to the submental surface electromyogram during swallowing. *Journal of Speech, Language, and Hearing Research, 42*(6), 1378–1391. https://doi.org/10.1044/jslhr.4206.1378

Patel, R., & Campellone, P. (2009). Acoustic and perceptual cues to contrastive stress in dysarthria. *Journal of Speech, Language, and Hearing Research, 52*(1), 206–222. https://doi.org/10.1044/1092-4388(2008/07-0078)

Patel, R., & Threats, T. (2016). One's voice: A central component of personal factors in augmentative and alternative communication. *Perspectives off the ASHA Special Interest Groups, 1*(12), 94–98. https://doi.org/10.1044/persp1.SIG12.94

Phinyomark, A., Phukpattaranont, P., & Limsakul, C. (2012). Feature reduction and selection for EMG signal classification. *Expert Systems With Applications, 39*(8), 7420–7431. https://doi.org/10.1016/j.eswa.2012.01.102

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1985). Speaking clearly for the hard of hearing I. *Journal of Speech and Hearing Research, 28*(1), 96–103. https://doi.org/10.1044/jshr.2801.96

Porbadnigk, A., Wester, M., Calliess, J., & Schultz, T. (2009). EEG-based speech recognition impact of temporal effects. In P. Encarnação, & A. Veloso (Eds.), *BIOSIGNALS 2009* (pp. 376–3781). DBLP.

Portney, L. G., & Watkins, M. P. (2000). *Foundations of clinical research: Applications to practice* (Vol. 2). Prentice Hall.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., & Schwarz, P. (2011). The Kaldi speech recognition toolkit. In D. Nahamoo, & M. Picheny (Eds.), *Proceedings of the ASRU 2019 IEEE Automatic Speech Recognition and Understanding Workshop* (pp. 1–4). IEEE.

Pullin, G., Treviranus, J., Patel, R., & Higginbotham, J. (2017). Designing interaction, voice and inclusion in AAC research. *Augmentative and Alternative Communication, 33*(3), 139–148. https://doi.org/10.1080/07434618.2017.1342690

Roy, S. H., De Luca, G., Cheng, M. S., Johansson, A., Gilmore, L. D., & De Luca, C. J. (2007). Electro-mechanical stability of surface EMG sensors. *Medical & Biological Engineering & Computing, 45*(5), 447–457. https://doi.org/10.1007/s11517-007-0168-z

Sataloff, R. T., Heman-Ackah, Y. D., & Hawkshaw, M. J. (2007). Clinical anatomy and physiology of the voice. *Otolaryngologic Clinics of North America, 40*(5), 909–929. https://doi.org/10.1016/j.otc.2007.05.002

Searl, J., & Knollhoff, S. (2018). Sense of effort and fatigue associated with talking after total laryngectomy. *American Journal of Speech-Language Pathology, 27*(4), 1434–1444. https://doi.org/10.1044/2018_AJSLP-17-0218

Shen, J., Pang, R., Weiss, R., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R., Agiomvrgiannakis, Y., & Wu, Y. (2018). Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018* (pp. 4779–4783). IEEE.

Smiljanić, R., & Bradlow, A. R. (2009). Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Language and Linguistics Compass, 3*(1), 236–264. https://doi.org/10.1111/j.1749-818X.2008.00112.x

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management, 45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Stepp, C. E. (2012). Surface electromyography for speech and swallowing systems: Measurement, analysis, and interpretation. *Journal of Speech, Language, and Hearing Research, 55*(4), 1232–1246. https://doi.org/10.1044/1092-4388(2011/11-0214)

Stepp, C. E., Hillman, R. E., & Heaton, J. T. (2010). Use of neck strap muscle intermuscular coherence as an indicator of vocal hyperfunction. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 18*(3), 329–335. https://doi.org/10.1109/TNSRE.2009.2039605

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

Suárez-Quintanilla, J., Fernández Cabrera, A., & Sharma, S. (2019). Anatomy, head and neck, larynx. In *StatPearls* [Internet]. StatPearls. https://www.ncbi.nlm.nih.gov/books/NBK538202

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. https://doi.org/10.1044/2014_JSLHR-S-13-0086

Toman, M., Meltzner, G., & Patel, R. (2018). Data requirements, selection and augmentation for DNN-based speech synthesis from crowdsourced data. In B. Yegnanarayana (Ed.), *INTERSPEECH 2018* (pp. 2878–2882). ISCA.

Tuller, B., Harris, K. S., & Kelso, J. S. (1981). Phase relationships among articulator muscles as a function of speaking rate and syllable stress. *The Journal of the Acoustical Society of America, 69*(S1), S55. https://doi.org/10.1121/1.386217

Ueda, N., Oyama, M., Harvey, J. E., & Ogura, J. H. (1972). Influence of certain extrinsic laryngeal muscles on artificial voice production. *The Laryngoscope, 82*(3), 468–482. https://doi.org/10.1288/00005537-197203000-00016

van Rijsbergen, C. J. (1979). In *Information retrieval* (2nd ed.). Butterworth-Heinemann.

Vojtech, J. M., Cler, G. J., & Stepp, C. E. (2018). Prediction of optimal facial electromyographic sensor configurations for human–machine interface control. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 26*(8), 1566–1576. https://doi.org/10.1109/TNSRE.2018.2849202

Weismer, G. (2006). *Motor speech disorders: Essays for Ray Kent*. Plural.

Welham, N. V., & Maclagan, M. A. (2003). Vocal fatigue: Current knowledge and future directions. *Journal of Voice, 17*(1), 21–30. https://doi.org/10.1016/S0892-1997(03)00033-X

Williams, S. E., & Watson, J. B. (1985). Differences in speaking proficiencies in three laryngectomee groups. *Archives of Otolaryngology, 111*(4), 216–219. https://doi.org/10.1001/archotol.1985.00800060040003

Yorkston, K. M., & Beukelman, D. R. (1981). *Assessment of intelligibility of dysarthric speech*. Pro-Ed.

Yorkston, K. M., Strand, E. A., & Kennedy, M. R. T. (1996). Comprehensibility of dysarthric speech. *American Journal of Speech-Language Pathology, 5*(1), 55–66. https://doi.org/10.1044/1058-0360.0501.55

Zhang, Y., Li, P., Zhu, X., Su, S. W., Guo, Q., Xu, P., & Yao, D. (2017). Extracting time-frequency feature of single-channel vastus medialis EMG signals for knee exercise pattern recognition. *PLOS ONE, 12*(7), e0180526. https://doi.org/10.1371/journal.pone.0180526