vibration

MDPI

Article

Prediction of Voice Fundamental Frequency and Intensity from Surface Electromyographic Signals of the Face and Neck

Jennifer M. Vojtech [1,2,*], Claire L. Mitchell [1,2], Laura Raiff [1,2,3], Joshua C. Kline [1,2] and Gianluca De Luca [1,2]

[1] Delsys, Inc., Natick, MA 01760, USA
[2] Altec, Inc., Natick, MA 01760, USA
[3] Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA
* Correspondence: jvojtech@delsys.com

Abstract: Silent speech interfaces (SSIs) enable speech recognition and synthesis in the absence of an acoustic signal. Yet, the archetypal SSI fails to convey the expressive attributes of prosody such as pitch and loudness, leading to lexical ambiguities. The aim of this study was to determine the efficacy of using surface electromyography (sEMG) as an approach for predicting continuous acoustic estimates of prosody. Ten participants performed a series of vocal tasks including sustained vowels, phrases, and monologues while acoustic data was recorded simultaneously with sEMG activity from muscles of the face and neck. A battery of time-, frequency-, and cepstral-domain features extracted from the sEMG signals were used to train deep regression neural networks to predict fundamental frequency and intensity contours from the acoustic signals. We achieved an average accuracy of 0.01 ST and precision of 0.56 ST for the estimation of fundamental frequency, and an average accuracy of 0.21 dB SPL and precision of 3.25 dB SPL for the estimation of intensity. This work highlights the importance of using sEMG as an alternative means of detecting prosody and shows promise for improving SSIs in future development.

Keywords: pitch; loudness; EMG; voice; speech; fundamental frequency; intensity

Citation: Vojtech, J.M.; Mitchell, C.L.; Raiff, L.; Kline, J.C.; De Luca, G. Prediction of Voice Fundamental Frequency and Intensity from Surface Electromyographic Signals of the Face and Neck. *Vibration* **2022**, *5*, 692–710. https://doi.org/10.3390/vibration5040041

Academic Editor: Aleksandar Pavic

Received: 19 August 2022
Accepted: 12 October 2022
Published: 13 October 2022

1. Introduction

Speech is the basis of human interaction. For many languages, spoken communication is not only governed by the words that make up a message, but also the relative emphasis of syllables within each word. Often conveyed by changes in prosody—including vocal characteristics of pitch, loudness, voice quality, and temporal variability—it is *how* the words are said that facilitates understanding, conveys meaning, and grants nuance to an interaction. Through unique modulations in these characteristics, individuals can develop their own speaking style and identity. However, people with a limited ability to produce speech, such as those who undergo laryngectomy due to trauma or disease, lack this natural method of self-expression. Consequentially, those affected often struggle with daily communication and tend to face psychosocial challenges, including difficulty integrating at work, social withdraw, depression, addiction, anxiety, and altered self-identity [1–5].

The development of assistive technologies known as silent speech interfaces (SSIs) attempts to bridge this gap in self-expression by providing an alternative method of communication that is independent of an acoustic signal. Instead, SSIs leverage other physiological signals to infer information about speech content and reconstruct this content as text or audible outputs [6]. Different approaches have included ultrasound and optical cameras [7–9], electropalatographic [10], or electromagnetic [11] devices for tracking tongue and lip movements; non-audible murmur microphones for detecting resonance in the vocal tract [12,13]; surface electromyography (sEMG) of articulatory muscles or the larynx (e.g., [14–18]); and motor cortex implants [19], electroencephalography [20] or electrocorticography (ECoG; [21]) to track speech-related brain activity.

Despite the advances in SSIs, the resulting synthesized speech often lacks prosody and, as a result, tends to sound monotone and unnatural. Recent work to overcome this shortcoming by Herff et al. [22] demonstrated that an SSI utilizing EcoG could preserve linguistic and conversational cues, wherein listeners found the synthesized speech to be intelligible 66% of the time. However, the system itself requires a craniotomy to operate, making it an invasive option that may not be ideal for those already suffering from trauma or disease. Another study conducted by Gonzalez et al. [23] also demonstrated the capability of an SSI to produce intelligible and natural speech using permanent magnetic articulography (PMA), but also suffers in usability due to the invasiveness of PMA and its current dependence on audio signals.

Using sEMG for alternative communication provides a noninvasive, easy-to-use alternative to EcoG- and PMA-based SSIs. Preliminary studies have shown the promise of sEMG-based SSIs to recognize a range of utterances including individual phonemes, isolated words, and even continuous speech with relatively high accuracy (e.g., [14,17,18,24,25]). Subsequent preliminary studies have begun to incorporate prosodic features in their sEMG-based SSI systems. By tracking articulatory muscle activity, sEMG-based SSIs from Johner et al. [26] and Vojtech et al. [18] were able to successfully distinguish emphasized words and questions from normal statements, demonstrating F1 scores of 0.68 and 0.92, respectively. While these studies demonstrated the ability of an sEMG-based SSI to detect prosodic features in speech, the metrics used may lack objectivity due to the large phonetic variation in how a word can be emphasized both within and across people [27]. As such, acoustic correlates of prosody could fulfill the unmet need to synthesize objective prosodic characteristics of speech more directly.

Past works have attempted to extract vocal pitch via estimates of fundamental frequency (f_o) from sEMG activity but encountered difficulties without the use of machine learning methods. This is likely because voice production is primarily modulated by the intrinsic laryngeal muscles, which are not detectable using surface electrodes [28]. Instead, sEMG-based estimates of f_o have largely been attributed to changes in extrinsic laryngeal muscles. Due to the small, interdigitated, and overlapping nature of the extrinsic laryngeal musculature, however, it has been postulated that some muscles that are not involved in the control of voice f_o still contribute to the sEMG signal [29]. In turn, more recent work has turned to machine learning to disentangle voice f_o from sEMG signals. Nakamura et al. [30] was first to extract the f_o contour from an sEMG signal via Gaussian mixture model-based voice conversion. Diener et al. [31] improved on this work by quantizing the f_o values instead of estimating the contour from a trained model, and by introducing a feed-forward neural network for f_o estimation. However, both studies resulted in relatively low model performance between observed and predicted f_o estimates ($r < 0.50$). On top of low performance, these works also focused on pitch as a sole prosodic feature even though modulations in pitch, loudness, timing, and voice quality are often interdependent [32] (i.e., a syllable that is perceived as stressed is often produced with simultaneous increases in f_o and intensity; [33]). Nevertheless, these studies provide an important first step toward introducing linguistic prosody into synthetic speech for sEMG-based SSIs.

The aim of our current study was to investigate the efficacy of using sEMG to recognize and track continuous estimates of voice f_o and intensity. To achieve this goal, a series of time-, cepstral-, and frequency-domain features derived from sEMG signals was used to train deep regression models to estimate f_o and intensity of a concurrently recorded acoustic signal. Model performance in generating continuous estimates of f_o and intensity was characterized using outcome measures of percent error, correlation (via Pearson's correlation and Lin's concordance correlation), accuracy (via mean bias error), and precision (via root-mean-square error). We hypothesized that our regression models would demonstrate prediction errors below perceptible ranges reported in the literature for f_o (0.20–0.30 semitones; [34–36]) and intensity (2–5 dB SPL; [32,37]).

2. Materials and Methods

2.1. Participants

Ten adults with typical voices (5 female, 5 male; *M* = 29.8 years, *SD* = 9.6 years, *range*: 21–53 years) participated in the study. All participants were fluent in English and reported no history of voice, speech, language, or hearing disorders. One participant spoke English with an Arabic accent. All participants provided informed, written consent in compliance with the Western Institutional Review Board.

2.2. Experimental Protocol

Participants were seated throughout the study in a quiet room. Surface EMG signals were collected using eight single-differential electrode pairs connected to either of two wireless Trigno Quattro sensors (Delsys, Natick, MA, USA). Each differential electrode pair was placed over a distinct region of the face or neck as described in Meltzner et al. [14,15] (Figure 1). Neck sensor placements included the anterior belly of the digastric, mylohyoid, and geniohyoid (sensor 1; [38]); platysma, mylohyoid and stylohyoid (sensor 2; [38]); and platysma, thyrohyoid, omohyoid, and sternohyoid (sensors 3 and 4; [39]). Face sensor placements [40] included the zygomaticus major and/or minor, levator labii superioris, and levator anguli oris (sensor 5); orbicularis oris (sensors 6 and 7); and mentalis (sensor 8). Just prior to sensor adhesion, the surface of the skin was prepared via alcohol wipe and tape peel exfoliation methods to remove excess hair and skin oils [41,42]. The eight sensors were then adhered to the skin using double-sided, hypoallergenic tape. Signals were recorded at 2222 Hz, bandpass filtered with roll-off frequencies of 20 Hz and 450 Hz, and amplified by a gain of 300.

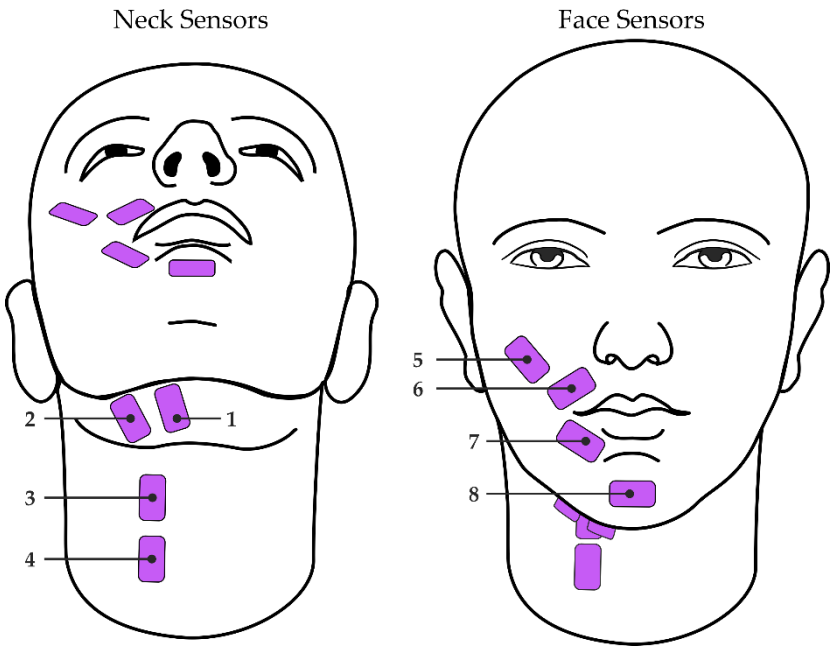

Figure 1. Configuration of sEMG sensors (pink) on the neck (**left**; sensors 1–4) and face (**right**; sensors 5–8).

Acoustic signals were recorded using an omnidirectional microphone (Movo LV-6C XLR) instrumented to a headset; for each participant, the microphone was positioned 45° from the midline and 4–7 cm from the lips. Microphone signals were pre-amplified (ART Tube MP Project Series) and digitized at 44.1 kHz (National Instruments USB NI-6251).

Time-aligned acoustic and sEMG signal acquisition was managed through a triggering setup within Delsys EMGworks software and involved a custom trigger module to connect the National Instruments DAQ board and sEMG base station trigger port.

To calculate sound pressure level (dB SPL) for all voice recordings, electrolaryngeal pulses were played at the lips while a sound pressure level meter (Check Mate CM-140) measured dB SPL at the microphone. The known sound pressure levels were later used to calibrate the microphone recordings.

From here, participants produced seven different types of voice and speech data to collect a heterogenous sample of vocal activity. A detailed overview of the voice and speech tasks can be found in Appendix A, and are listed in brief below:

1. **Tones**—Sustained vowels /a/, /i/, /u/, and /ae/ produced for 3–5 s at a constant pitch and loudness (normative, high/low pitch, high/low loudness)
2. **Legatos**—Continuous slide from one pitch to another for vowels /a/, /i/, /u/, or /ae/
3. **VCV Syllables**—Vowel-consonant-vowel sequences (e.g., /afa/) where both vowels are equally stressed or only one vowel is stressed
4. **Phrases**—Standard, short speech tokens uttered in a normal speaking voice
5. **Reading Passages**—Standard passages uttered in a normal speaking voice
6. **Questions**—Short segments of unstructured speech in response to a question
7. **Monologues**—Long segments of unstructured speech in response to a prompt

Tasks were presented to participants on printouts displayed on a weighted-base copyholder (Fellowes 21128). Participants were instructed to notify the experimenter (authors J.V. or C.M.) when ready to begin a task; the experimenter would then start a recording to collect concurrent sEMG and acoustic data. In this way, participants proceeded through each task at their own pace. For tasks in which participants were instructed to alter their pitch and/or loudness (i.e., tones, legatos, nonsense words; see Appendix A), the degree of change was not assigned a specific sound pressure level or f_0. Instead, it was determined by participants to fit within their comfortable conversational levels, similar to the recommended clinical instructions for instrumentally assessing voice [43]. An average of 2975.5 s of data was recorded for each participant (2501.9–3503.9 s), with recording duration by speech task shown in Table 1.

Table 1. Recording duration by speech task, shown as mean (range).

Speech Task	Recording Duration (s)
Tones	351.7 (232.2–620.7)
Legatos	132.1 (97.4–205.8)
VCV Syllables	284.4 (174.6–464.0)
Phrases	649.8 (523.5–790.9)
Reading Passages	1041.1 (888.9–1209.0)
Questions	241.6 (168.5–330.8)
Monologues	274.8 (214.5–374.9)

2.3. Data Processing

The sequence of data processing steps included: (1) signal alignment to align data recorded from the eight unique sEMG channels to the acoustic data recorded from the headset microphone, (2) voice f_0 and intensity contour extraction, (3) feature extraction, and (4) data splitting. Each processing step is described in detail below.

2.3.1. Signal Alignment

As each sEMG sensor was configured over distinct regions of the face or neck (with sensor configurations influenced by variable skin-electrode impedances and depth of the muscle from the skin surface, among other factors), a dynamic time warping (DTW) algorithm was implemented to capture the non-linear similarities between the acoustic data and the multiple, spatially distributed EMG sensors. For this procedure, the sEMG data from each sensor was first upsampled to 44.1 kHz to match the sampling rate of the acoustic data. An exact, memory-efficient algorithm for DTW was then employed using

the *linmdtw* package [44] in Python (v.3.8) to compute signal alignments using a hop value of 0.010 s.

2.3.2. Voice f_o and Intensity Contour Extraction

Two features were extracted from the acoustic data as outcome variables: voice f_o (Hz) and voice intensity (dB SPL). The f_o contour was extracted from each acoustic recording using the Praat autocorrelation-based algorithm [45] via the *Parselmouth* package [46] in Python. For this algorithm, minimum and maximum f_o values were set to 65 Hz and 475 Hz, respectively [47–49]. The time step for this algorithm was set to default (0.75/*minimum* f_o).

The intensity contour was extracted following methods used in Praat, wherein the amplitude of a signal was first squared, then convolved with a Gaussian analysis window (Kaiser-20 with sidelobes below −190 dB). The duration of the analysis window was set to the default used in the Praat algorithm (3.2/*minimum* f_o). Resulting intensity values were converted from units of dB to units of dB SPL using the known sound pressure levels acquired during data collection.

2.3.3. Feature Extraction

Acoustic (f_o and intensity contours) and sEMG signals were windowed at a frame size of 40 ms with a 20-ms step shift for f_o data and 150 ms with a 30-ms step shift for intensity data. The f_o and intensity data were represented per frame by mean values. The sEMG data were represented per channel and per frame by a set of 20 common EMG features, which are listed in Table 2. All listed features were extracted for each of the 8 sEMG channels, then 24 redundant channel-features (e.g., the cross-correlation of channels 3 and 8 vs. the cross-correlation of channels 8 and 3) were removed. All features were then cascaded into a final vector with a dimension of 593 per sEMG sample.

Table 2. Features used in sEMG data processing.

	Feature	Dimension per Channel	References
1	Beta coherence	8	[50,51]
2	Central frequency variance	1	[52,53]
3	Coherence	8	[50,51]
4	Cross-correlation	8	[54]
5	Daubechies 2 wavelet coefficients, maximum (peak)	4	[55]
6	Daubechies 2 wavelet coefficients, mean	4	[55]
7	Daubechies 2 wavelet coefficients, variance	4	[55]
8	Maximum (peak) frequency	1	[52,55]
9	Mean absolute value	1	[56–61]
10	Mean frequency	1	[59,60]
11	Mean power density	1	[53,62]
12	Median frequency	1	[60,61]
13	Mel-frequency cepstral coefficients	24	[14,15,18]
14	Power density wavelength	1	[57]
15	Root mean square	1	[56,57,60,61]
16	Slope sign change	1	[57,60]
17	Spectral moments	3	[52,57,61,62]
18	Variance	1	[56,57,60,61]
19	Waveform length	1	[57,59–61]
20	Zero crossings	1	[17,57,59,60]

Principal component analysis (PCA) was employed on the common set of 593 sEMG features from each participant to mitigate multicollinearity of features while constructing relevant features that capture most of the variance in the data. For each participant, the PCA criterion for the number of selected features was such that 90% of the variance in the data was explained [63–65]. This process yielded an average of 97.6 ± 2.1 features to characterize a given observation for intensity data and 106.0 ± 1.6 across participants for f_o.

2.3.4. Data Splitting

The amount of data available for model construction varied within and across participants due to differences in participant speech characteristics (e.g., speaking rate), task type (e.g., a sustained vowel vs. a long monologue), and outcome metric. For instance, there was substantially more data available for intensity than f_o since f_o could only be computed during voiced speech. Data splitting was therefore stratified across speech tasks to preserve the approximate proportions of the original dataset across models and to ensure an 80–20 (training-test) split.

Two methods were carried out to minimize overfitting: data augmentation and *k*-fold cross-validation. Data augmentation was applied as a regularization technique by injecting noise from a Gaussian distribution (based on the mean and standard deviations of the features) into the dataset [66,67]. Following, *k*-fold cross-validation with $k = 5$ folds was employed on the training data to quantify the variation in model performance [68]; resulting was a 60-20-20 split for training-validation-test sets.

2.4. Model Development

Model training was carried out using a Dell XPS 8950 desktop with the Windows 11 Pro 64-bit operating system. The processor was an Intel Core i7-12700 with 12 central processing unit cores. The computer was equipped with 32 GB random access memory, and the graphics processing unit of the computer was the NVIDIA GeForce RTX 3080.

Two types of f_o and intensity models were created: (1) single-speaker models, meaning that individual f_o and intensity models were trained for each participant, and (2) multi-speaker models, meaning that data from all 10 participants was used to train, validate, and test a single model for each outcome measure (f_o, intensity). The former scheme was implemented to account for variations in the sEMG signal that may occur across participants due to differences in exact electrode configuration, skin-electrode impedances, skin and adipose thickness, and muscle activation during speech. The latter scheme was implemented to determine feasibility in creating a generalized architecture for estimating f_o and intensity in spite of person-specific variations in sEMG activity. Importantly, data augmentation was not implemented for the multi-speaker models due to the large amount of available data (spanning 10 participants).

A schematic representation of the single-speaker models for f_o and intensity can be found in Figure 2. The hidden layers within both models use the GeLU activation function. Parameter optimization for the f_o (Figure 2a) and intensity (Figure 2b) models is performed at a learning rate of 0.001 (batch size: 1024) and 0.005 (batch size: 2048), respectively, using the ADAM optimizer. As the models are intended to solve a regression problem, mean squared error is used as a loss function. Accordingly, the output layer for each model comprises one unit with a linear activation function. In the models for f_o, all f_o values (predicted, observed) are standardized to semitones (ST) relative to a reference value based on the speaker's average f_o. Both models are deep regression neural networks that predict outcome values at a resolution of 0.01 ST (f_o) or 0.01 dB SPL (intensity).

A schematic of the multi-speaker models that were constructed for f_o and intensity are shown in Figure 3. As in the single-speaker models, the hidden layers within both models use the GeLU activation function, mean squared error is used as a loss function, and the output layer consists of one unit with linear activation. Parameter optimization for f_o (Figure 3a) and intensity (Figure 3b) models is performed at a learning rate of 0.001 (batch size: 1024) and 0.0005 (batch size: 4096), respectively, using the ADAM optimizer. Batch normalization is included before the first activation layer of the intensity model to normalize the inputs to the first GeLU activation function. Due to differences in habitual pitch and loudness, f_o values are standardized to ST using a reference value of 90 Hz rather than the speaker's average f_o and intensity values are normalized (0–1) within-participant across the available data. Both models are deep regression neural networks that predict outcome values at a resolution of 0.01 ST (f_o) or 0.01 dB (intensity).

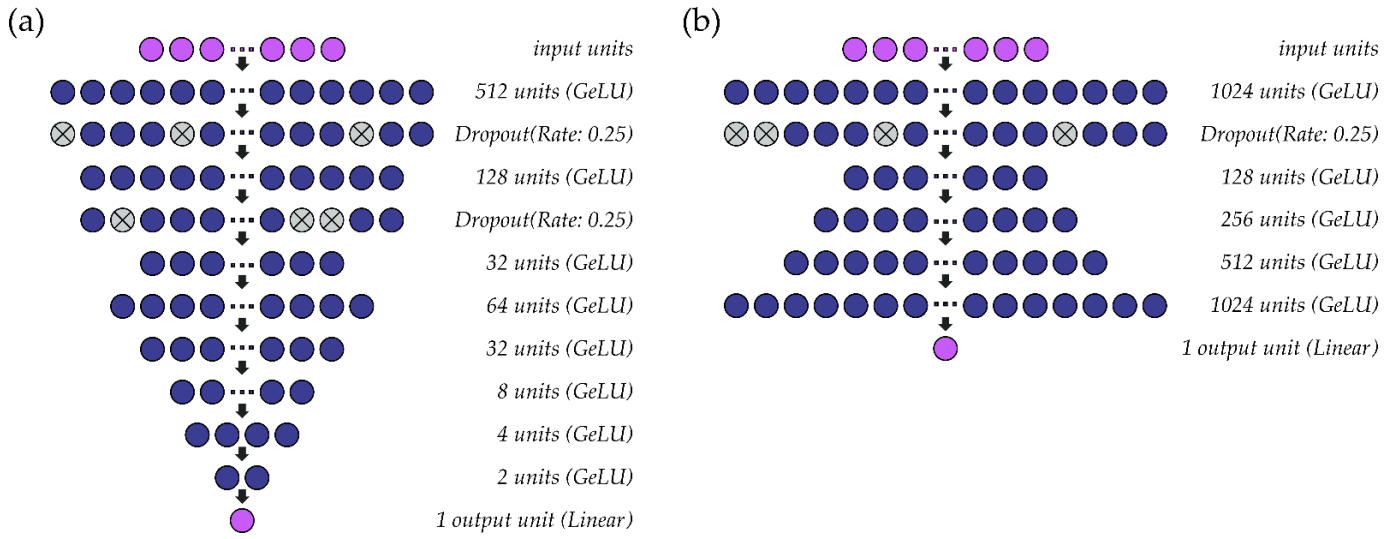

Figure 2. Structure of the single-speaker deep regression neural networks used to estimate (**a**) f_o and (**b**) intensity from sEMG signals.

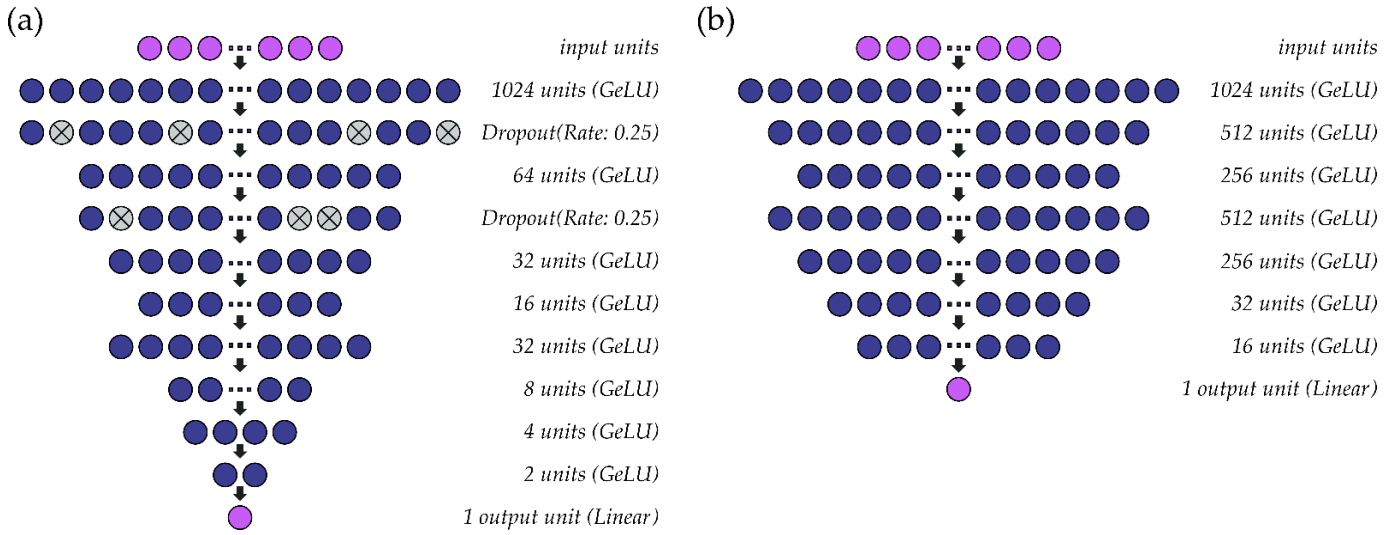

Figure 3. Structure of the multi-speaker deep regression neural networks used to estimate (**a**) f_o and (**b**) intensity from sEMG signals.

2.5. Model Performance

Model performance was quantified using metrics of mean absolute percent error (MAPE) as well as Pearson product-moment correlation coefficients (r) and Lin concordance correlation coefficients (CCC) to enable comparisons to the literature. Model performance was also quantified as the root-mean-square error (RMSE) and mean bias error (MBE) between observed and predicted estimates to provide insight into the precision and accuracy of f_o or intensity estimates. Performance for the training (60%) and validation (20%) data was compared across $k = 5$ folds. The fold that yielded the highest *CCC* value for validation data was identified as the final model for f_o or intensity. Final f_o and intensity models were then evaluated using the unseen test data (20%), and model performance was quantified per participant via MAPE, r, *CCC*, RMSE, and MBE.

3. Results

3.1. Single-Speaker Models

3.1.1. Training and Validation Set Performance

Mean outcomes from both models (f_o, intensity) were of the same magnitude between training and validation datasets, with validation results exhibiting slightly larger standard deviation values across the $k = 5$ cross-validation folds. Average model performance across cross-validation folds is shown by participant in Table A2 for f_o and Table A3 for intensity as well as summarized below.

Model performance in estimating f_o was comparable across cross-validation folds for training and validation datasets. Results for MAPE were, on average, 1.58% (*SD* = 0.24%) for the training data and 2.39% (*SD* = 0.72%) for the validation data. Findings were of similar magnitude for *r* and *CCC*, demonstrating average values of $r = 0.98$ (*SD* = 0.01) and *CCC* = 0.97 (*SD* = 0.01) for training data and $r = 0.92$ (*SD* = 0.05) and *CCC* = 0.92 (*SD* = 0.06) for validation data. Average training RMSE values were 0.34 ST (*SD* = 0.05 ST) and 0.52 ST (*SD* = 0.15 ST) for validation. Finally, MBE results were 0.27 ST (*SD* = 0.04 ST) and 0.41 ST (*SD* = 0.12 ST) for training and validation data, respectively.

Performance in estimating intensity demonstrated similar errors between training and validation datasets. Across the cross-validation folds, average training MAPE was 1.87% (*SD* = 0.41%) whereas validation MAPE was 3.31% (*SD* = 0.94%). Pearson's *r* and Lin's *CCC* values were above 0.90 for both datasets, averaging at $r = 0.98$ (*SD* = 0.01) and *CCC* = 0.98 (*SD* = 0.01) for training data with $r = 0.92$ (*SD* = 0.04) and *CCC* = 0.91 (*SD* = 0.04) for validation data. Average training RMSE was 2.38 dB SPL (*SD* = 0.96 dB SPL) whereas validation RMSE was 4.81 dB SPL (*SD* = 1.89 dB SPL). Results demonstrated an average MBE of 1.82 dB SPL (*SD* = 0.73 dB SPL) and 3.15 dB SPL (*SD* = 1.22 dB SPL) for training and validation data, respectively.

3.1.2. Test Set Performance

Within-participant performance on the test set is shown in Table 3. In the model for f_o, MAPE was under 5% for all participants (*M* = 2.54%, *SD* = 0.72%). Pearson's *r* and Lin's *CCC* values demonstrated mean values of $r = 0.92$ (*SD* = 0.05) and *CCC* = 0.91 (*SD* = 0.07). The mean ST error between observed and predicted values was 0.01 ST (*SD* = 0.08 ST), with precision estimates averaging at 0.56 ST (*SD* = 0.16 ST). An example of observed and predicted contours is shown for f_o in Figure 4b.

Table 3. Single-speaker f_o and intensity model performance on the test set for 10 participants.

ID	f_o					Intensity				
	MAPE (%)	***r***	***CCC***	**RMSE (ST)**	**MBE (ST)**	**MAPE (%)**	***r***	***CCC***	**RMSE (dB SPL)**	**MBE (dB SPL)**
1	1.75	0.96	0.95	0.38	0.05	2.21	0.98	0.98	2.06	−0.27
2	1.82	0.95	0.94	0.40	0.09	2.44	0.98	0.98	2.23	−0.03
3	2.49	0.94	0.94	0.55	−0.02	1.60	0.97	0.97	4.17	0.88
4	2.33	0.95	0.94	0.51	0.03	1.46	0.98	0.98	3.53	−0.69
5	2.36	0.94	0.94	0.52	−0.04	3.40	0.96	0.95	3.35	0.85
6	2.26	0.96	0.95	0.50	−0.01	1.36	0.97	0.97	3.33	−0.35
7	2.25	0.96	0.96	0.49	0.00	4.60	0.94	0.94	6.17	2.04
8	3.79	0.79	0.74	0.86	−0.16	2.75	0.98	0.98	2.41	0.10
9	2.41	0.95	0.94	0.53	0.07	1.54	0.97	0.97	3.11	−1.00
10	4.03	0.83	0.80	0.90	0.12	2.50	0.98	0.98	2.17	0.52

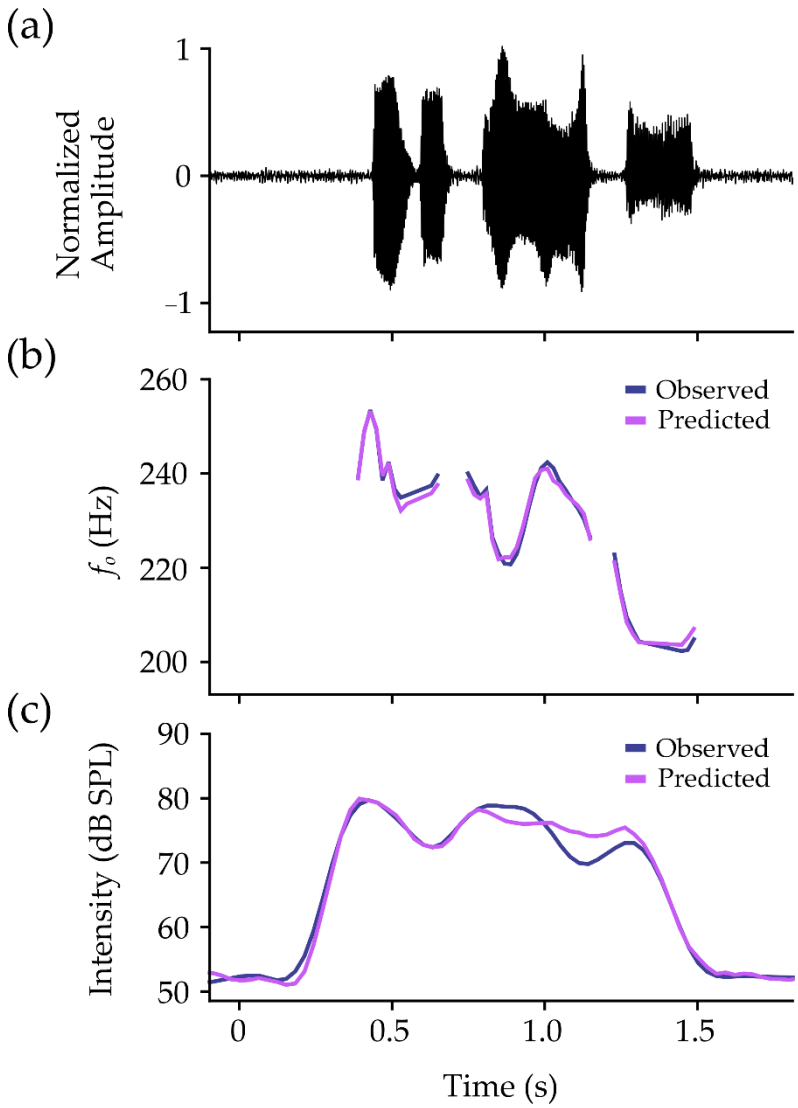

Figure 4. Example data for one participant from the phrase "Easy for you to say". The normalized microphone signal is shown (**a**), with observed (navy lines) and predicted (pink lines) contours for (**b**) f_o and (**c**) intensity. Contours for f_o have been converted from semitones to Hertz (Hz) for visualization purposes.

Results for intensity also showed MAPE values under 5% for all participants (M = 2.38%, SD = 0.97%). Pearson's r and Lin's CCC values were over 0.94 for all participants, showing mean values of r = 0.97 (SD = 0.01) and CCC = 0.97 (SD = 0.01). The RMSE between observed and predicted values was 3.25 dB SPL (SD = 1.18 dB SPL), with MBE averaging at 0.21 dB SPL (SD = 0.85 dB SPL). An example of observed and predicted contours is shown for intensity in Figure 4c.

3.2. Multi-Speaker Models

Results for the multi-speaker f_o model is shown for the training, validation, and test datasets in Table 4. The multi-speaker f_o model demonstrated similar trends across outcome metrics, wherein performance was worst on the validation data, followed by the training data. Performance in the test set was comparable to the training and validation data. Specifically, MBE which was lowest (most accurate) for the test dataset (0.13 ST). Average MAPE values were below 10% across all three dataset types, with poor validation correlations (r = 0.25, CCC = 0.10) and moderate training (r = 0.41, CCC = 0.17) and test (r = 0.36, CCC = 0.25) correlations.

Table 4. Multi-speaker f_0 model performance on training, validation, and test datasets.

Dataset	MAPE (%)	*r*	*CCC*	RMSE (ST)	MBE (ST)
Training *	8.15 (0.53)	0.41 (0.04)	0.17 (0.12)	1.67 (0.12)	1.42 (0.11)
Validation *	8.16 (0.62)	0.25 (0.10)	0.10 (0.07)	1.66 (0.11)	1.42 (0.11)
Test	7.95	0.36	0.25	1.65	0.13

* Results are shown across cross-validation folds as mean (standard deviation) for training and validation datasets.

Results for the multi-speaker intensity model is shown for the training, validation, and test datasets in Table 5. As the multi-speaker model was evaluated on normalized SPL values, results for RMSE and MBE are shown in units of decibels (dB). The multi-speaker intensity model showed the best performance on the test dataset in terms of correlation (r = 0.56, CCC = 0.48) and accuracy (−0.02 dB). MAPE was under 15% for all datasets, with poor-to-moderate training (r = 0.51, CCC = 0.44) and validation (r = 0.32, CCC = 0.24) correlations. Finally, the precision of intensity estimates was comparable across the three datasets (0.11–0.12 dB).

Table 5. Multi-speaker intensity model performance on training, validation, and test datasets.

Dataset	MAPE (%)	*r*	*CCC*	RMSE (dB)	MBE (dB)
Training *	12.67 (2.27)	0.51 (0.10)	0.44 (0.10)	0.12 (0.03)	0.10 (0.02)
Validation *	13.18 (1.97)	0.32 (0.10)	0.24 (0.09)	0.12 (0.01)	0.10 (0.01)
Test	12.36	0.56	0.48	0.11	−0.02

* Results are shown across cross-validation folds as mean (standard deviation) for training and validation datasets.

4. Discussion

The goal of this study was to determine the feasibility of using sEMG signals of the face and neck to predict two primary attributes of linguistic prosody: voice f_0 and intensity. This study builds on our primary work in using sEMG activity for silent speech recognition (i.e., identifying the words in a message; [14,15]) and for classifying basic manipulations in prosody (i.e., identifying how the words in a message are conveyed; [18]). Taking this past work into account, the current study successfully demonstrates efficacy in using sEMG as an alternative method for detecting prosody via continuous estimates of f_0 and intensity.

4.1. Single-Speaker vs. Multi-Speaker Models

Single- and multi-speaker models were examined in this work. The single-speaker models were trained and tested on data recorded for an individual participant, whereas the multi-speaker models were trained and tested from the data of 10 participants. The motivation for examining both single- and multi-speaker models stems from the reliance of each model on the acoustic signal. Both models rely on audio data for training, but the multi-speaker models could, in theory, be used by other individuals without an inherent reliance on their specific audio data. Applications for this latter model include situations in which the individual cannot supply acoustic data to train a model (e.g., those who cannot voice due to trauma or disease, such as laryngectomees).

Unsurprisingly, our single-speaker models performed better than the multi-speaker counterparts, as sEMG signals are speaker-dependent due to skin-electrode impedances, skin and adipose thickness, as well as differences in muscle activation during speech. Indeed, most prior works in this area focus on single-speaker models for this very reason (e.g., [18,25,31,69]). We argue that the overall performance of the multi-speaker models is still promising, as our results provide preliminary evidence of predicting f_0 and intensity within 10% and 15% error, respectively. Additional work is still necessary to extend this approach toward a robust system that is independent of the user's acoustic information. Moreover, the multi-speaker models examined here included data from all 10 participants with each dataset (training, validation, test), such that model performance on unseen participants was not evaluated. This was done to determine the feasibility of using a

single model to capture sEMG and acoustic variability across individuals to estimate f_o or intensity prior to doing so in unseen individuals. However, future work should aim to train and test such models on independent participants to determine the generalizability of our approach (e.g., for those who cannot contribute acoustic information to model training). Future work should also consider acquiring more data from individuals across a wide range of vocal function as one potential method of increasing the generalizability of our multi-speaker models, as a small sample size of only ten individuals with typical voices was included here.

4.2. Significance of Single-Speaker Model Performance

4.2.1. Comparisons to Model Performance in the Literature

We investigated the ability of deep regression models to predict discrete estimates of voice f_o and intensity from sEMG data of the face and neck musculature. This work expands on studies from the literature that utilize different machine learning approaches for estimating prosodic information from EMG data alone. Our results notably surpass values reported in the literature for f_o estimation while also detailing one of the first accounts (to our knowledge) of predicting vocal intensity (loudness) from sEMG signals.

The use of sEMG for estimating voice f_o is a concept that has been scarcely explored over the past decade, resulting in a limited number of comparative works. A pioneering study by Nakamura et al. [30] sought to use a Gaussian mixture model-based approach to estimate f_o from five sEMG sensors, demonstrating an average correlation between observed and predicted f_o values of $r = 0.49$ across three speakers. De Armas et al. [69] sought to predict f_o using support vector machine regression and classification from sEMG traces. In estimating f_o from tones, the authors reported an average correlation of $r = 0.96$; however, this correlation decreased to $r = 0.88$ when estimating f_o from phrases. Making use of a similar protocol, Ahmadi et al. [70] aimed to achieve better correlations in predicting f_o values from sEMG data as compared to De Armas et al. [69]. As anticipated, the authors reported an average correlation of $r = 0.93$ when estimating f_o from phrases from a small sample of three participants.

Although the average correlations in Nakamura et al. [30], De Armas et al. [69], and Ahmadi et al. [70] are lower than or comparable to those observed in the current study ($r = 0.92$), it must be noted that it is difficult to directly compare model performance across studies. There are substantial differences in methodology across these works, ranging from experimental setup (e.g., sEMG hardware), protocol (e.g., vocal tasks), and model construction (e.g., support vector machine vs. deep regression models) that complicate interpretations for why a given model may have performed better than another. For instance, our study utilized bipolar sEMG sensors sampled at 2222 Hz whereas that of Nakamura et al. [30] acquired sEMG activity via a mix of bipolar and monopolar sEMG sensors sampled at 600 Hz. Nakamura et al. [30] recorded participants as they produced phrases and De Armas et al. [69] and Ahmadi et al. [70] recorded participants as they produced tones, legatos, and phrases, whereas the current study incorporated these three vocal tasks as well as additional types of continuous (i.e., reading passages) and spontaneous (i.e., monologues and questions) speech. Thus, we caution readers to consider the differences in methodology across sEMG-based SSI studies rather than taking the correlative results presented here at face value.

Still, it must be considered that developing an SSI that estimates f_o from basic speech units like tones or legatos may be a necessary first step to demonstrate the proof of principle; however, the introduction of continuous and spontaneous speech tasks as in the current study is important to consider for ensuring ecological validity. In fact, these tasks represented more than 52% of the total data recorded in the study. Without such tasks, the SSI is inherently constrained in requiring basic f_o manipulations (in the case of tones or legatos) and pauses (in the case of phrases) to decipher f_o. Moreover, De Armas et al. [69] observed an average RMSE of 2.81 ST for f_o estimation, which is about 5-fold greater than the average

RMSE obtained in the current work of 0.56 ST. These results show the importance of using multiple outcome metrics to provide comprehensive insight into model performance.

More recently, Diener et al. [31] examined the relationship between acoustic (observed) and sEMG-derived (predicted) speech features when using electrode arrays. The authors opted to build upon their prior work by deriving "quantized" estimates of f_o rather than continuous estimates; however, the authors still observed poor correlative performance (r = 0.27). A shift from direct f_o estimation can be observed in Janke et al. [69] and Botelho et al. [70], wherein algorithmic performance did not specifically include f_o as an outcome. Instead, the authors sought to determine the output quality of the speech (via mel-cepstral distortion and mel-frequency cepstral coefficients) rather than the quality of specific prosodic attributes (e.g., f_o, intensity). Though outside the scope of the current study, future work could incorporate these speech quality features in addition to the prosodic features examined here.

4.2.2. Comparisons to Meaningful Changes in f_o and Intensity

Our results show a high degree of agreement between acoustic and sEMG-derived estimates of f_o and intensity within each participant. Within this analysis, RMSE and MBE were calculated as an estimate of prediction precision and accuracy, respectively. For multi-speaker f_o models, our results indicate a mean MBE of 0.03 ST. This suggests that our models will, on average, generate a positively biased systematic error (i.e., overestimated) of approximately 0.03 ST. The average RMSE across participants was 0.56 ST, indicating that the average spread of errors will approach 0.56 ST when using our models to estimate f_o. For single-speaker intensity models, our findings indicate an average MBE of 0.21 dB SPL and RMSE of 3.25 dB SPL. These results suggest that using our models to estimate intensity from sEMG signals will generate a positively biased error of 0.21 dB SPL, with the precision of intensity estimates approaching 3.25 dB SPL.

It is important to consider how these errors between observed and predicted f_o values compare to meaningful differences in the literature. For instance, the average vocal pitch discrimination ability of an adult has been reported to be within the range of 0.20 to 0.30 ST [34–36]. The average accuracy of our f_o estimations was found to be 0.01 ST, meaning that the MBE associated with using our single-speaker f_o models is on the order of one magnitude smaller than the pitch discrimination abilities of a typical adult reported in the literature. This suggests that erroneous f_o values predicted by our model will, on average, not be perceived by the typical adult.

The average errors obtained for vocal intensity can also be compared to meaningful values reported in the literature. Specifically, the mean short-term variation in vocal intensity has been reported to be approximately 2–5 dB SPL for adults [37,71]. With an average MBE of 0.21 dB SPL, our results suggest that average erroneous intensity estimates predicted by the single-speaker intensity models will be within the bounds of typical, short-term variations in vocal intensity.

4.3. Physiological Interprations of Model Performance

The results of the current study suggest that f_o and intensity can be sufficiently estimated on a per-individual basis from sEMG activity of the face and neck. The notion that these prosodic attributes—f_o, in particular—can be estimated from relatively surface-level muscles is interesting when considering the orofacial and laryngeal muscles necessary for voicing, as voice production is primarily modulated by the intrinsic laryngeal muscles. Specifically, the primary function of the cricothyroid is to lengthen and stretch the vocal folds to, in turn, increase the vibratory rate of the vocal folds (and thus, increase f_o; [72]). The thyroarytenoid, on the other hand, stabilizes the onset of phonation and contributes to increases in the vibratory rate of the vocal folds [71,73]. Taken together, the contraction force of these muscles has been shown to jointly increase with increases in voice f_o and intensity [74].

Due to the relatively deep location of muscles within the larynx, however, it is unlikely that the activity of the cricothyroid or thyroarytenoid contributes to the detected signal when using surface electrodes [75]. Instead, it is more likely that activity from the extrinsic laryngeal muscles—which induce changes in laryngeal elevation to *indirectly* affect the vibratory rate of the vocal folds [76]—along with muscles of the face contributed to the detected sEMG signals. Indeed, prior work examining the thyrohyoid, sternothyroid, and sternohyoid ("strap muscles") during different vocal tasks suggests that these extrinsic laryngeal muscles are involved in the dynamic modulation of voice production (i.e., rising or falling frequency) rather than in the specific f_o itself [77]. It has also been reported that the strap muscles are differentially active during high and low f_o productions [78–80], as well as during voice productions at varying loudness levels [81]. In addition to the extrinsic laryngeal muscles, changes in vocal intensity from habitual loudness to either softer or louder levels has been shown to significantly alter average sEMG amplitude of the lip muscles [82]. Increases in voice f_o have also been associated with differential changes in surface electromyographic activity of the face [83].

Taking these prior works into account, it is likely that our models were able to learn from the sEMG activity from the sensors placed over the extrinsic laryngeal muscles (i.e., sensors 1–4 in Figure 1) and the orofacial muscles (i.e., sensors 5–8 in Figure 1) to understand how a given participant's dynamic patterns used to modulate their voice, including f_o and intensity. It is also important to note that these past studies examined the amplitude of the sEMG signal relative to voice f_o and intensity, whereas the current study leveraged a combination of 57 time-, frequency-, and cepstral-domain features from the sEMG signal. Our results suggest that this combination of features can effectively detect changes in extrinsic laryngeal and orofacial muscle activity in a way that is associated with changes in voice f_o and intensity. Additional investigations should be undertaken to examine these voice attributes relative to specific sEMG sensor sites (e.g., over the strap muscles vs. over the lip muscles) to further elucidate the relationship between extrinsic laryngeal or orofacial muscle activity and f_o or intensity.

4.4. Limitations and Future Directions

Although the current study details favorable results regarding the performance of deep regression neural networks for predicting voice f_o and intensity, further investigation is warranted to continue to enhance the accuracy and accessibility of the models. For instance, voice f_o is relatively position-independent whereas voice intensity may vary based on the distance from the microphone to the mouth. Though outside the scope of this study—which sought to demonstrate the proof-of-concept that f_o and intensity could be estimated from sEMG activity of the face and neck—future work should investigate normalization methods to account for differences in microphone distance that may occur within and across individuals who use the system. Within this vein, our multi-speaker models did not perform as well as single-speaker models for f_o and intensity predictions. As a result, the current methods must rely on an individual's acoustic signal to train a model, hampering usability in the target population of individuals who are unable to voice (due to trauma or disease). As discussed in Section 4.2, future work is needed to increase the accuracy and precision of multi-speaker f_o and intensity models possibly by expanding the number of participants as is done for acoustic speech recognition models (e.g., [84–86]); in this way, the models could be trained using sEMG and acoustic data from individuals with typical voices and then tested (used) by those without a voice.

Voice f_o and intensity are important as suprasegmental characteristics of speech but are not the only two attributes of linguistic prosody. Though outside the scope of the current study, future investigations should incorporate attributes of timing (e.g., word duration) and voice quality into the models for f_o and intensity estimation. Within a similar vein, the current study aimed to examine suprasegmental characteristics of speech separately from segmental characteristics, such as word or phoneme prediction. Subsequent efforts will be

undertaken to combine our approach with the word recognition methods detailed in our prior works toward developing a prosodic, sEMG-based SSI.

5. Conclusions

Surface EMG is a promising modality for SSIs due to its noninvasive nature and ease of application; however, most sEMG-based SSIs fail to convey the expressive attributes of prosody, including pitch and loudness. This work details the construction and evaluation of deep regression neural networks for predicting continuous estimates of voice f_o and intensity from sEMG recordings from muscles of the face and neck. When evaluated in ten participants, model estimation of f_o yielded an average accuracy of 0.01 ST and precision of 0.56 ST while model estimation of intensity provided a mean accuracy of 0.21 dB SPL and precision of 3.25 dB SPL. The average accuracy of f_o estimation was approximately one order of magnitude smaller than the pitch discrimination abilities of a typical adult, suggesting that erroneous f_o values predicted by our model will, on average, not be perceived by the typical adult. Moreover, our results suggest that erroneous model estimates of intensity will, on average, be within the bounds of typical, short-term variations in vocal intensity. This study is a critical first step toward introducing linguistic prosody into synthetic speech for sEMG-based SSIs.

Author Contributions: Conceptualization, J.M.V. and J.C.K.; methodology, J.M.V.; software, J.M.V. and C.L.M.; validation, J.M.V.; formal analysis, J.M.V. and C.L.M.; investigation, J.M.V.; resources, G.D.L. and J.C.K.; data curation, C.L.M. and J.M.V.; writing—original draft preparation, J.M.V. and L.R.; writing—review and editing, J.M.V., C.L.M., L.R., G.D.L. and J.C.K.; visualization, J.M.V. and L.R.; supervision, J.M.V. and J.C.K.; project administration, J.M.V.; funding acquisition, J.M.V., J.C.K. and G.D.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the De Luca Foundation and by the National Institutes of Health under Grant No. R44DC017097 (G.D.L.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Western Institutional Review Board (Protocol #20182089, approved 9 March 2021).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the identifiable nature of voice acoustic recordings.

Acknowledgments: The authors would like to thank Bhawna Shiwani for assistance in data processing.

Conflicts of Interest: J.M.V., C.L.M., L.R., G.D.L. and J.C.K. are employed by Delsys, Inc., a commercial company that manufactures and markets sensor and software technologies for human movement, and Altec, Inc., an R&D company that performs research to reimagine human potential.

Appendix A

Table A1. Overview of speech tasks.

Task	Description	Subtasks
Tones	Sustained vowels /a/, /i/, /u/, and /ae/ produced at a constant pitch and loudness, repeated three times for each variation	1. Typical pitch and loudness 2. High pitch 3. Low pitch 4. High intensity 5. Low intensity
Legatos	Continuous slide from one pitch to another using the vowels /a/, /i/, /u/, and /ae/	1. Low pitch 2. Mid pitch 3. High pitch

Table A1. *Cont.*

Task	Description	Subtasks
VCV [a] Syllables	Bisyllabic productions repeated three times for each variation	1. Equal stress 2. Stress on first vowel 3. Stress on second vowel
Phrases	Standard, short speech tokens that introduce various stress placements	1. UNL Phrases 2. RFF Phrases
Reading Passages	Standard reading passages that introduce various stress placements	1. The Caterpillar Passage 2. My Grandfather Passage 3. Rainbow Passage 4. Golf Passage 5. Pronunciation Reading Passage 6. Please Call Stella 7. Comma Gets a Cure 8. Frog and Toad 9. Excerpt from Harry Potter and the Chamber of Secrets 10. Excerpt from The Little Prince 11. Excerpt from the Boston University Radio Speech Corpus
Questions	Short (<30 s) segment of unstructured, conversational speech	1. If you could live in any decade, what would it be and why? 2. What is your favorite time of day and why? 3. If you were to make a movie about your life, what genre would you choose and why? 4. How did you get here today? 5. Do you have any vacation or travel plans? 6. Tell me about how the weather has been recently. 7. What did you do last weekend?
Monologues	Long (>60 s) segment of unstructured, conversational speech	1. Tell me how to make a peanut butter and jelly sandwich. 2. Tell me how you do your laundry. 3. Tell me how you get ready for work. 4. Tell me how you make your bed.

[a] VCV = vowel-consonant-vowel, with vowels /a/, /i/, or /u/ and consonants /f/, /v/, /p/, or /b/.

Appendix B

Table A2. Performance for the f_0 models across k = 5 cross-validation training (Train) and validation (*Valid*) datasets for N = 10 participants. Results are shown as mean (standard deviation).

ID	MAPE (%)		*r*		*CCC*		RMSE (ST)		MBE (ST)	
	Train	***Valid***	**Train**	***Valid***	**Train**	***Valid***	**Train**	***Valid***	**Train**	***Valid***
1	1.36 (0.04)	*2.21 (0.71)*	0.98 (0.00)	*0.92 (0.06)*	0.97 (0.00)	*0.90 (0.08)*	0.29 (0.01)	*0.49 (0.16)*	0.23 (0.01)	*0.38 (0.13)*
2	1.27 (0.03)	*2.07 (0.43)*	0.98 (0.00)	*0.93 (0.04)*	0.97 (0.00)	*0.92 (0.04)*	0.28 (0.01)	*0.46 (0.10)*	0.22 (0.01)	*0.36 (0.07)*
3	1.60 (0.02)	*1.69 (0.08)*	0.98 (0.00)	*0.98 (0.00)*	0.98 (0.00)	*0.97 (0.00)*	0.34 (0.00)	*0.36 (0.02)*	0.28 (0.00)	*0.29 (0.01)*
4	1.79 (0.04)	*2.69 (0.75)*	0.97 (0.00)	*0.92 (0.06)*	0.97 (0.00)	*0.91 (0.06)*	0.38 (0.01)	*0.60 (0.18)*	0.31 (0.01)	*0.47 (0.13)*
5	1.66 (0.02)	*3.67 (2.63)*	0.97 (0.00)	*0.80 (0.29)*	0.97 (0.00)	*0.78 (0.32)*	0.36 (0.01)	*0.77 (0.51)*	0.29 (0.00)	*0.62 (0.43)*
6	1.79 (0.02)	*2.77 (0.80)*	0.97 (0.00)	*0.93 (0.06)*	0.97 (0.00)	*0.92 (0.06)*	0.39 (0.01)	*0.61 (0.20)*	0.31 (0.00)	*0.48 (0.14)*
7	1.63 (0.02)	*2.66 (0.84)*	0.98 (0.00)	*0.93 (0.06)*	0.98 (0.00)	*0.92 (0.06)*	0.35 (0.00)	*0.59 (0.20)*	0.28 (0.00)	*0.46 (0.15)*
8	1.13 (0.01)	*1.23 (0.09)*	0.98 (0.00)	*0.98 (0.00)*	0.98 (0.00)	*0.98 (0.00)*	0.25 (0.00)	*0.27 (0.02)*	0.19 (0.00)	*0.21 (0.02)*
9	1.80 (0.03)	*3.03 (1.18)*	0.97 (0.00)	*0.89 (0.11)*	0.97 (0.00)	*0.88 (0.13)*	0.38 (0.01)	*0.68 (0.29)*	0.31 (0.00)	*0.53 (0.21)*
10	1.78 (0.01)	*1.86 (0.02)*	0.97 (0.00)	*0.97 (0.00)*	0.97 (0.00)	*0.97 (0.00)*	0.38 (0.00)	*0.40 (0.00)*	0.31 (0.00)	*0.32 (0.00)*

Table A3. Performance for the intensity models across $k = 5$ cross-validation training (Train) and validation (*Valid*) datasets for $N = 10$ participants. Results are shown as mean (standard deviation).

ID	MAPE (%)		*r*		*CCC*		RMSE (dB SPL)		MBE (dB SPL)	
	Train	***Valid***	**Train**	***Valid***	**Train**	***Valid***	**Train**	***Valid***	**Train**	***Valid***
1	1.96 (0.31)	*3.88 (2.77)*	0.99 (0.00)	*0.90 (0.15)*	0.99 (0.00)	*0.89 (0.17)*	1.7 (0.25)	*3.94 (3.00)*	1.28 (0.21)	*2.60 (1.95)*
2	2.30 (1.01)	*3.80 (1.58)*	0.99 (0.00)	*0.95 (0.05)*	0.99 (0.01)	*0.94 (0.05)*	1.72 (0.67)	*3.33 (1.38)*	1.36 (0.59)	*2.28 (0.93)*
3	1.68 (0.36)	*3.44 (2.13)*	0.97 (0.01)	*0.81 (0.23)*	0.97 (0.01)	*0.81 (0.24)*	3.90 (0.88)	*9.30 (6.42)*	2.94 (0.62)	*5.96 (3.70)*
4	1.78 (0.81)	*2.39 (0.80)*	0.96 (0.03)	*0.93 (0.05)*	0.96 (0.04)	*0.92 (0.05)*	4.16 (1.85)	*5.91 (2.01)*	3.20 (1.41)	*4.31 (1.41)*
5	2.11 (0.16)	*4.27 (1.75)*	0.99 (0.00)	*0.92 (0.08)*	0.99 (0.00)	*0.91 (0.08)*	1.68 (0.09)	*4.21 (1.99)*	1.29 (0.09)	*2.67 (1.15)*
6	1.27 (0.16)	*2.07 (0.64)*	0.98 (0.01)	*0.93 (0.06)*	0.98 (0.01)	*0.92 (0.07)*	2.83 (0.35)	*5.06 (1.91)*	2.17 (0.27)	*3.54 (1.12)*
7	2.46 (0.09)	*4.65 (0.96)*	0.99 (0.00)	*0.94 (0.03)*	0.99 (0.00)	*0.94 (0.03)*	2.33 (0.09)	*5.87 (1.47)*	1.75 (0.08)	*3.41 (0.72)*
8	2.11 (0.12)	*3.55 (1.50)*	0.99 (0.00)	*0.95 (0.06)*	0.99 (0.00)	*0.94 (0.06)*	1.72 (0.08)	*3.25 (1.51)*	1.32 (0.07)	*2.21 (0.91)*
9	1.20 (0.21)	*1.83 (0.84)*	0.98 (0.01)	*0.91 (0.12)*	0.98 (0.01)	*0.91 (0.12)*	2.28 (0.46)	*4.39 (3.00)*	1.74 (0.30)	*2.62 (1.18)*
10	1.88 (0.05)	*3.18 (1.29)*	0.99 (0.00)	*0.96 (0.04)*	0.99 (0.00)	*0.96 (0.04)*	1.50 (0.03)	*2.88 (1.34)*	1.15 (0.03)	*1.93 (0.76)*

References

1. Keszte, J.; Danker, H.; Dietz, A.; Meister, E.F.; Pabst, F.; Vogel, H.-J.; Meyer, A.; Singer, S. Mental disorders and psychosocial support during the first year after total laryngectomy: A prospective cohort study. *Clin. Otolaryngol.* **2013**, *38*, 494–501. [CrossRef] [PubMed]
2. Terrell, J.E.; Fisher, S.G.; Wolf, G.T. Long-term Quality of Life After Treatment of Laryngeal Cancer. *Arch. Otolaryngol. Head Neck Surg.* **1998**, *124*, 964–971. [CrossRef]
3. Bickford, J.M.; Coveney, J.; Baker, J.; Hersh, D. Self-expression and identity after total laryngectomy: Implications for support. *Psycho-Oncology* **2018**, *27*, 2638–2644. [CrossRef]
4. Lúcio, G.D.S.; Perilo, T.V.D.C.; Vicente, L.C.C.; Friche, A.A.D.L. The impact of speech disorders quality of life: A questionnaire proposal. *CoDAS* **2013**, *25*, 610–613. [CrossRef] [PubMed]
5. Garcia, S.M.; Weaver, K.; Moskowitz, G.B.; Darley, J.M. Crowded minds: The implicit bystander effect. *J. Pers. Soc. Psychol.* **2002**, *83*, 843–853. [CrossRef] [PubMed]
6. Denby, B.; Schultz, T.; Honda, K.; Hueber, T.; Gilbert, J.; Brumberg, J. Silent speech interfaces. *Speech Commun.* **2009**, *52*, 270–287. [CrossRef]
7. Fabre, D.; Hueber, T.; Girin, L.; Alameda-Pineda, X.; Badin, P. Automatic animation of an articulatory tongue model from ultrasound images of the vocal tract. *Speech Commun.* **2017**, *93*, 63–75. [CrossRef]
8. Hueber, T.; Benaroya, E.-L.; Chollet, G.; Denby, B.; Dreyfus, G.; Stone, M. Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. *Speech Commun.* **2009**, *52*, 288–300. [CrossRef]
9. Crevier-Buchman, L.; Gendrot, C.; Denby, B.; Pillot-Loiseau, C.; Roussel, P.; Colazo-Simon, A.; Dreyfus, G. Articulatory strategies for lip and tongue movements in silent versus vocalized speech. In Proceedings of the 17th International Congress of Phonetic Science, Hong Kong, China, 17–21 August 2011; pp. 1–4.
10. Kimura, N.; Gemicioglu, T.; Womack, J.; Li, R.; Zhao, Y.; Bedri, A.; Su, Z.; Olwal, A.; Rekimoto, J.; Starner, T. SilentSpeller: Towards mobile, hands-free, silent speech text entry using electropalatography. In Proceedings of the CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April–5 May 2022; pp. 1–19. [CrossRef]
11. Fagan, M.; Ell, S.; Gilbert, J.; Sarrazin, E.; Chapman, P. Development of a (silent) speech recognition system for patients following laryngectomy. *Med. Eng. Phys.* **2008**, *30*, 419–425. [CrossRef] [PubMed]
12. Hirahara, T.; Otani, M.; Shimizu, S.; Toda, T.; Nakamura, K.; Nakajima, Y.; Shikano, K. Silent-speech enhancement using body-conducted vocal-tract resonance signals. *Speech Commun.* **2010**, *52*, 301–313. [CrossRef]
13. Nakajima, Y.; Kashioka, H.; Shikano, K.; Campbell, N. Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin. In Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, China, 6–10 April 2003; pp. 708–711. [CrossRef]

14. Meltzner, G.S.; Heaton, J.T.; Deng, Y.; De Luca, G.; Roy, S.H.; Kline, J.C. Development of sEMG sensors and algorithms for silent speech recognition. *J. Neural Eng.* **2018**, *15*, 046031. [CrossRef]
15. Meltzner, G.S.; Heaton, J.T.; Deng, Y.; De Luca, G.; Roy, S.H.; Kline, J.C. Silent Speech Recognition as an Alternative Communication Device for Persons With Laryngectomy. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 2386–2398. [CrossRef]
16. Maier-Hein, L.; Metze, F.; Schultz, T.; Waibel, A. Session independent non-audible speech recognition using surface electromyography. In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, Cancun, Mexico, 27 November–1 December 2005; pp. 331–336. [CrossRef]
17. Jou, S.-C.; Schultz, T.; Walliczek, M.; Kraft, F.; Waibel, A. Towards continuous speech recognition using surface electromyography. In Proceedings of the Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, 17–21 December 2006. [CrossRef]
18. Vojtech, J.M.; Chan, M.D.; Shiwani, B.; Roy, S.H.; Heaton, J.T.; Meltzner, G.S.; Contessa, P.; De Luca, G.; Patel, R.; Kline, J.C. Surface Electromyography–Based Recognition, Synthesis, and Perception of Prosodic Subvocal Speech. *J. Speech Lang. Hear. Res.* **2021**, *64*, 2134–2153. [CrossRef]
19. Brumberg, J.S.; Guenther, F.H.; Kennedy, P.R. An Auditory Output Brain–Computer Interface for Speech Communication. In *Briefs in Electrical and Computer Engineering*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 7–14. [CrossRef]
20. Porbadnigk, A.; Wester, M.; Calliess, J.; Schultz, T. EEG-based speech recognition impact of temporal effects. In Proceedings of the International Conference on Bio-Inspired Systems and Signal Processing, Porto, Portugal, 14–17 January 2009; pp. 376–381.
21. Angrick, M.; Herff, C.; Mugler, E.; Tate, M.C.; Slutzky, M.W.; Krusienski, D.J.; Schultz, T. Speech synthesis from ECoG using densely connected 3D convolutional neural networks. *J. Neural Eng.* **2019**, *16*, 036019. [CrossRef] [PubMed]
22. Herff, C.; Diener, L.; Angrick, M.; Mugler, E.; Tate, M.C.; Goldrick, M.A.; Krusienski, D.J.; Slutzky, M.W.; Schultz, T. Generating Natural, Intelligible Speech From Brain Activity in Motor, Premotor, and Inferior Frontal Cortices. *Front. Neurosci.* **2019**, *13*, 1267. [CrossRef]
23. Gonzalez, J.A.; Cheah, L.A.; Gilbert, J.M.; Bai, J.; Ell, S.R.; Green, P.D.; Moore, R.K. A silent speech system based on permanent magnet articulography and direct synthesis. *Comput. Speech Lang.* **2016**, *39*, 67–87. [CrossRef]
24. Lee, K.-S. EMG-Based Speech Recognition Using Hidden Markov Models With Global Control Variables. *IEEE Trans. Biomed. Eng.* **2008**, *55*, 930–940. [CrossRef] [PubMed]
25. Diener, L.; Bredehöft, S.; Schultz, T. A comparison of EMG-to-Speech Conversion for Isolated and Continuous Speech. In *ITG-Fachbericht 282: Speech Communication*; ITG: Oldenburg, Germany, 2018; pp. 66–70.
26. Johner, C.; Janke, M.; Wand, M.; Schultz, T. Inferring Prosody from Facial Cues for EMG-based Synthesis of Silent Speech. In *Advances in Affective and Pleasurable Design*; CRC: Boca Raton, FL, USA, 2013; pp. 634–643.
27. Kohler, K.J. What is Emphasis and How is it Coded? In Proceedings of the Speech Prosody Dresden, Dresden, Germany, 2–5 May 2006; pp. 748–751.
28. Nakamura, K.; Janke, M.; Wand, M.; Schultz, T. Estimation of fundamental frequency from surface electromyographic data: EMG-to-F0. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Prague, Czech Republic, 22–27 May 2011; pp. 573–576. [CrossRef]
29. Diener, L.; Umesh, T.; Schultz, T. Improving Fundamental Frequency Generation in EMG-To-Speech Conversion Using a Quantization Approach. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019—Proceedings, Singapore, 15–18 December 2019; pp. 682–689. [CrossRef]
30. Gramming, P. Vocal loudness and frequency capabilities of the voice. *J. Voice* **1991**, *5*, 144–157. [CrossRef]
31. Anderson, C. Transcribing Speech Sounds. In *Essentials of Linguistics*; McMaster University: Hamilton, ON, USA, 2018.
32. Moore, R.E.; Estis, J.; Gordon-Hickey, S.; Watts, C. Pitch Discrimination and Pitch Matching Abilities with Vocal and Nonvocal Stimuli. *J. Voice* **2008**, *22*, 399–407. [CrossRef] [PubMed]
33. Nikjeh, D.A.; Lister, J.J.; Frisch, S.A. The relationship between pitch discrimination and vocal production: Comparison of vocal and instrumental musicians. *J. Acoust. Soc. Am.* **2009**, *125*, 328–338. [CrossRef]
34. Murray, E.S.H.; Stepp, C.E. Relationships between vocal pitch perception and production: A developmental perspective. *Sci. Rep.* **2020**, *10*, 3912. [CrossRef]
35. Hunter, E.J.; Titze, I.R. Variations in Intensity, Fundamental Frequency, and Voicing for Teachers in Occupational Versus Nonoccupational Settings. *J. Speech Lang. Hear. Res.* **2010**, *53*, 862–875. [CrossRef]
36. Palmer, P.M.; Luschei, E.S.; Jaffe, D.; McCulloch, T.M. Contributions of Individual Muscles to the Submental Surface Electromyogram During Swallowing. *J. Speech Lang. Hear. Res.* **1999**, *42*, 1378–1391. [CrossRef] [PubMed]
37. Ding, R.; Larson, C.R.; Logemann, J.A.; Rademaker, A.W. Surface Electromyographic and Electroglottographic Studies in Normal Subjects Under Two Swallow Conditions: Normal and During the Mendelsohn Manuever. *Dysphagia* **2002**, *17*, 1–12. [CrossRef] [PubMed]
38. Eskes, M.; van Alphen, M.; Balm, A.J.M.; Smeele, L.E.; Brandsma, D.; van der Heijden, F. Predicting 3D lip shapes using facial surface EMG. *PLoS ONE* **2017**, *12*, e0175025. [CrossRef]
39. Hermens, H.J.; Freriks, B.; Disselhorst-Klug, C.; Rau, G. Development of recommendations for SEMG sensors and sensor placement procedures. *J. Electromyogr. Kinesiol.* **2000**, *10*, 361–374. [CrossRef]
40. Roy, S.H.; De Luca, G.; Cheng, M.S.; Johansson, A.; Gilmore, L.D.; De Luca, C.J. Electro-mechanical stability of surface EMG sensors. *Med. Biol. Eng. Comput.* **2007**, *45*, 447–457. [CrossRef] [PubMed]

41. Patel, R.R.; Awan, S.N.; Barkmeier-Kraemer, J.; Courey, M.; Deliyski, D.; Eadie, T.; Paul, D.; Svec, J.G.; Hillman, R. Recommended Protocols for Instrumental Assessment of Voice: American Speech-Language-Hearing Association Expert Panel to Develop a Protocol for Instrumental Assessment of Vocal Function. *Am. J. Speech Lang. Pathol.* **2018**, *27*, 887–905. [CrossRef]
42. Tralie, C.J.; Dempsey, E. Exact, Parallelizable Dynamic Time Warping Alignment with Linear Memory. In Proceedings of the 21st International Society for Music Information Retrieval Conference, Montréal, QC, Canada, 11–16 October 2020; pp. 462–469.
43. Boersma, P.; Weenink, D. Praat: Doing Phonetics by Computer. 2013. Available online: http://www.praat.org (accessed on 19 August 2022).
44. Jadoul, Y.; Thompson, B.; de Boer, B. Introducing Parselmouth: A Python interface to Praat. *J. Phon.* **2018**, *71*, 1–15. [CrossRef]
45. Coleman, R.F.; Markham, I.W. Normal variations in habitual pitch. *J. Voice* **1991**, *5*, 173–177. [CrossRef]
46. Baken, R.J. *Clinical Measurement of Speech and Voice*; College-Hill Press: Worthing, UK, 1987.
47. Awan, S.N.; Mueller, P.B. Speaking fundamental frequency characteristics of centenarian females. *Clin. Linguist. Phon.* **1992**, *6*, 249–254. [CrossRef] [PubMed]
48. Stepp, C.E.; Hillman, R.E.; Heaton, J.T. Modulation of Neck Intermuscular Beta Coherence During Voice and Speech Production. *J. Speech Lang. Hear. Res.* **2011**, *54*, 836–844. [CrossRef]
49. Stepp, C.E.; Hillman, R.E.; Heaton, J.T. Use of Neck Strap Muscle Intermuscular Coherence as an Indicator of Vocal Hyperfunction. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2010**, *18*, 329–335. [CrossRef] [PubMed]
50. Phinyomark, A.; Phukpattaranont, P.; Limsakul, C. Feature reduction and selection for EMG signal classification. *Expert Syst. Appl.* **2012**, *39*, 7420–7431. [CrossRef]
51. Malvuccio, C.; Kamavuako, E.N. The Effect of EMG Features on the Classification of Swallowing Events and the Estimation of Fluid Intake Volume. *Sensors* **2022**, *22*, 3380. [CrossRef]
52. Joshi, D.; Bhatia, D. Cross-correlation evaluated muscle co-ordination for speech production. *J. Med. Eng. Technol.* **2013**, *37*, 520–525. [CrossRef]
53. Abbaspour, S.; Lindén, M.; Gholamhosseini, H.; Naber, A.; Ortiz-Catalan, M. Evaluation of surface EMG-based recognition algorithms for decoding hand movements. *Med. Biol. Eng. Comput.* **2019**, *58*, 83–100. [CrossRef]
54. Soon, M.W.; Anuar, M.I.H.; Abidin, M.H.Z.; Azaman, A.S.; Noor, N.M. Speech recognition using facial sEMG. In Proceedings of the 2017 IEEE International Conference on Signal and Image Processing Applications, ICSIPA, Sarawak, Malaysia, 12–14 September 2017; pp. 1–5. [CrossRef]
55. Fraiwan, L.; Lweesy, K.; Al-Nemrawi, A.; Addabass, S.; Saifan, R. Voiceless Arabic vowels recognition using facial EMG. *Med. Biol. Eng. Comput.* **2011**, *49*, 811–818. [CrossRef]
56. Srisuwan, N.; Phukpattaranont, P.; Limsakul, C. Three steps of Neuron Network classification for EMG-based Thai tones speech recognition. In Proceedings of the 2013 10th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, ECTI-CON, Krabi, Thailand, 15–17 May 2013; pp. 1–6. [CrossRef]
57. Jong, N.S.; Phukpattaranont, P. A speech recognition system based on electromyography for the rehabilitation of dysarthric patients: A Thai syllable study. *Biocybern. Biomed. Eng.* **2018**, *39*, 234–245. [CrossRef]
58. Phinyomark, A.; Limsakul, C.; Phukpattaranont, P. A novel feature extraction for robust EMG pattern recognition. *J. Comput.* **2020**, *1*, 71–80.
59. Srisuwan, N.; Phukpattaranont, P.; Limsakul, C. Feature selection for Thai tone classification based on surface EMG. *Procedia Eng.* **2012**, *32*, 253–259. [CrossRef]
60. Du, S.; Vuskovic, M. Temporal vs. spectral approach to feature extraction from prehensile EMG signals. In Proceedings of the 2004 IEEE International Conference on Information Reuse and Integration, Las Vegas, NV, USA, 8–10 November 2004; pp. 344–350. [CrossRef]
61. Enders, H.; Maurer, C.; Baltich, J.; Nigg, B.M. Task-Oriented Control of Muscle Coordination during Cycling. *Med. Sci. Sports Exerc.* **2013**, *45*, 2298–2305. [CrossRef] [PubMed]
62. Matrone, G.C.; Cipriani, C.; Secco, E.L.; Magenes, G.; Carrozza, M.C. Principal components analysis based control of a multi-dof underactuated prosthetic hand. *J. Neuroeng. Rehabil.* **2010**, *7*, 16. [CrossRef] [PubMed]
63. Soechting, J.F.; Flanders, M. Sensorimotor control of contact force. *Curr. Opin. Neurobiol.* **2008**, *18*, 565–572. [CrossRef]
64. Shorten, C.; Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
65. Maharana, K.; Mondal, S.; Nemade, B. A review: Data pre-processing and data augmentation techniques. *Glob. Transit. Proc.* **2022**, *3*, 91–99. [CrossRef]
66. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer Science & Business Media: New York, NY, USA, 2013.
67. De Armas, W.; Mamun, K.A.; Chau, T. Vocal frequency estimation and voicing state prediction with surface EMG pattern recognition. *Speech Commun.* **2014**, *63–64*, 15–26. [CrossRef]
68. Ahmadi, F.; Araujo Ribeiro, M.; Halaki, M. Surface electromyography of neck strap muscles for estimating the intended pitch of a bionic voice source. In Proceedings of the IEEE 2014 Biomedical Circuits and Systems Conference, BioCAS 2014—Proceedings, Lausanne, Switzerland, 22–24 October 2014; pp. 37–40. [CrossRef]
69. Janke, M.; Diener, L. EMG-to-Speech: Direct Generation of Speech From Facial Electromyographic Signals. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 2375–2385. [CrossRef]

70. Botelho, C.; Diener, L.; Küster, D.; Scheck, K.; Amiriparian, S.; Schuller, B.W.; Trancoso, I. Toward silent paralinguistics: Speech-to-EMG—Retrieving articulatory muscle activity from speech. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Brno, Czech Republic, 30 August–3 September 2020; pp. 354–358. [CrossRef]
71. Choi, H.-S.; Ye, M.; Berke, G.S.; Kreiman, J. Function of the Thyroarytenoid Muscle in a Canine Laryngeal Model. *Ann. Otol. Rhinol. Laryngol.* **1993**, *102*, 769–776. [CrossRef]
72. Chhetri, D.K.; Neubauer, J.; Sofer, E.; Berry, D.A. Influence and interactions of laryngeal adductors and cricothyroid muscles on fundamental frequency and glottal posture control. *J. Acoust. Soc. Am.* **2014**, *135*, 2052–2064. [CrossRef]
73. Chhetri, D.K.; Neubauer, J. Differential roles for the thyroarytenoid and lateral cricoarytenoid muscles in phonation. *Laryngoscope* **2015**, *125*, 2772–2777. [CrossRef] [PubMed]
74. Lindestad, P.; Fritzell, B.; Persson, A. Quantitative Analysis of Laryngeal EMG in Normal Subjects. *Acta Oto-Laryngol.* **1991**, *111*, 1146–1152. [CrossRef]
75. Vojtech, J.M.; Stepp, C.E. Electromyography. In *Manual of Clinical Phonetics*, 1st ed.; Ball, M., Ed.; Routledge: London, UK, 2021; pp. 248–263. [CrossRef]
76. Ueda, N.; Ohyama, M.; Harvey, J.E.; Ogura, J.H. Influence of certain extrinsic laryngeal muscles on artificial voice production. *Laryngoscope* **1972**, *82*, 468–482. [CrossRef] [PubMed]
77. Roubeau, B.; Chevrie-Muller, C.; Guily, J.L.S. Electromyographic Activity of Strap and Cricothyroid Muscles in Pitch Change. *Acta Oto-Laryngol.* **1997**, *117*, 459–464. [CrossRef]
78. Hollien, H.; Moore, G.P. Measurements of the Vocal Folds during Changes in Pitch. *J. Speech Hear. Res.* **1960**, *3*, 157–165. [CrossRef]
79. Collier, R. Physiological correlates of intonation patterns. *J. Acoust. Soc. Am.* **1975**, *58*, 249–255. [CrossRef]
80. Andersen, K.F.; Sonninen, A. The Function of the Extrinsic Laryngeal Muscles at Different Pitch. *Acta Oto-Laryngol.* **1960**, *51*, 89–93. [CrossRef]
81. Goldstein, E.; Heaton, J.; Kobler, J.; Stanley, G.; Hillman, R. Design and Implementation of a Hands-Free Electrolarynx Device Controlled by Neck Strap Muscle Electromyographic Activity. *IEEE Trans. Biomed. Eng.* **2004**, *51*, 325–332. [CrossRef]
82. Wohlert, A.B.; Hammen, V.L. Lip Muscle Activity Related to Speech Rate and Loudness. *J. Speech Lang. Hear. Res.* **2000**, *43*, 1229–1239. [CrossRef]
83. Zhu, M.; Wang, X.; Deng, H.; He, Y.; Zhang, H.; Liu, Z.; Chen, S.; Wang, M.; Li, G. Towards Evaluating Pitch-Related Phonation Function in Speech Communication Using High-Density Surface Electromyography. *Front. Neurosci.* **2022**, *16*, 941594. [CrossRef]
84. Li, J.; Lavrukhin, V.; Ginsburg, B.; Leary, R.; Kuchaiev, O.; Cohen, J.M.; Gadde, R.T. Jasper: An End-to-End Convolutional Neural Acoustic Model. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Graz, Austria, 5–19 September 2019; pp. 71–75. [CrossRef]
85. Post, M.; Kumar, G.; Lopez, A.; Karakos, D.; Callison-Burch, C.; Khudanpur, S. Improved speech-to-text translation with the Fisher and Callhome Spanish-English speech translation corpus—ACL Anthology. In Proceedings of the 10th International Workshop on Spoken Language Translation: Papers, Heidelberg, Germany, 5–6 December 2013.
86. Rao, K.; Sak, H.; Prabhavalkar, R. Exploring Architectures, Data and Units For Streaming End-to-End Speech Recognition with RNN-Transducer. In Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2017—Proceedings, Okinawa, Japan, 16–20 December 2017; pp. 193–199. [CrossRef]