ACOUSTICALLY-DRIVEN PHONEME REMOVAL THAT PRESERVES VOCAL AFFECT CUES

Camille Noufi[*]*, Jonathan Berger*

Stanford University
Center for Computer Research
in Music and Acoustics
Stanford, CA, USA

Karen J. Parker, Daniel L. Bowling[†]

Stanford School of Medicine
Department of Psychiatry
and Behavioral Sciences
Stanford, CA, USA

ABSTRACT

In this paper, we propose a method for removing linguistic information from speech for the purpose of isolating paralinguistic indicators of affect. The immediate utility of this method lies in clinical tests of sensitivity to vocal affect that are not confounded by language, which is impaired in a variety of clinical populations. The method is based on simultaneous recordings of speech audio and electroglottographic (EGG) signals. The speech audio signal is used to estimate the average vocal tract filter response and amplitude envelop. The EGG signal supplies a direct correlate of voice source activity that is mostly independent of phonetic articulation. The dynamic energy of the speech audio and the average vocal tract filter are applied to the EGG signal create a third signal designed to capture as much paralinguistic information from the vocal production system as possible—maximizing the retention of bioacoustic cues to affect—while eliminating phonetic cues to verbal meaning. To evaluate the success of this method, we studied the perception of corresponding speech audio and transformed EGG signals in an affect rating experiment with online listeners. The results show a high degree of similarity in the perceived affect of matched signals, indicating that our method is effective.

Index Terms— speech, paralanguage, affect, voice transformation, electroglottagraphy, phoneme removal

1. INTRODUCTION

Much of the information conveyed by speech is transmitted through paralinguistic cues encoded in the audio signal. These paralinguistic cues are essential to the communication of emotions, intentions, and personality [1–3]. For the majority of our daily interactions, these paralinguistic cues are embedded among phonetic cues encoding linguistic meaning. Although most individuals have no problem parsing linguistic and paralinguistic cues in speech and responding appropriately, this ability is often impaired in clinical populations (e.g., in autism [4] and depression [5]). Focusing on autism, the impairment is assumed to pertain to the reception of paralinguistic cues to speaker affect. However, the tests on which this assumption is based use speech stimuli, and thus confound sensitivity to paralanguage with language functioning. Testing sensitivity to paralinguistic affect directly requires isolating it from speech. This is important for understanding the nature of auditory-vocal contributions to clinical dysfunction, particularly in mental health.

Existing methods that attempt to isolate paralinguistic cues from speech benefit from economy and efficiency, but they also lose significant amounts of paralinguistic information, particularly concerning affect. For example, one simple and efficient method is to remove phonologic content from speech audio by adaptively low-pass filtering the signal such that the filter roll-off occurs below the second formant peak, thus removing a critical cue to vowel identification (i.e., the ratio between the first and second formants). However, because this method removes high-frequency content ($>$ starting at approx. 500-2500 Hz, depending on the vowel [6]), it also destroys important affective content [7].

Another method is to discard phonetic cues by separating the vocal signal into two parts, the signal representing the laryngeal source, and the signal representing the supralaryngeal filter. Whereas the filter is more typically associated with linguistic articulation [8–10], the source is more associated with paralinguistic features that are essential to affect, such as voice pitch, breathiness, roughness, and other varieties of voice quality [1, 3, 11–13]. The most common method for separating the vocal source signal $e(t)$ from the vocal tract impulse response $h(t)$ is linear predictive coding (LPC) [14]. LPC uses a pth-order linear predictor to estimate speech signal $\tilde{s}(t)$ from p previous samples. This is done by solving for optimum predictor coefficients a_k of the filter for a pseudo-stationary frame of speech. Once the filter coefficients are found the "residual" $e(t)$ is calculated, representing a mixture of both glottal and noise-based phonetic content.

Although source-filter separation is useful in many cases, the reality is that both linguistic and paralinguistic information are encoded across the entire range of the frequency spectrum [7] and produced by both source and filter [1, 11–13]. Given that existing methods operating on speech audio alone do not adequately separate linguistic and paralinguistic information, we designed a new method that leverages synchronized audio and electroglottagraphic (EGG) [15–18] recordings of speech to create a third transformed EGG ("tEGG") signal made by applying a speech-based spectro-temporal transform to the EGG signal. When played as audio, this stimulus lacks the speech signal's linguistic content but retains significant paralinguistic information across the frequency spectrum. In Section 2, we describe the algorithm for the transformation method. In Section 3, we describe the speech/EGG recording process and an implementation of the algorithm. In Section 4, we evaluate our algorithm in an online experiment comparing ratings of emotional affect in corresponding speech and tEGG signal pairs. We conclude by briefly summarizing future directions and applications of the method

---

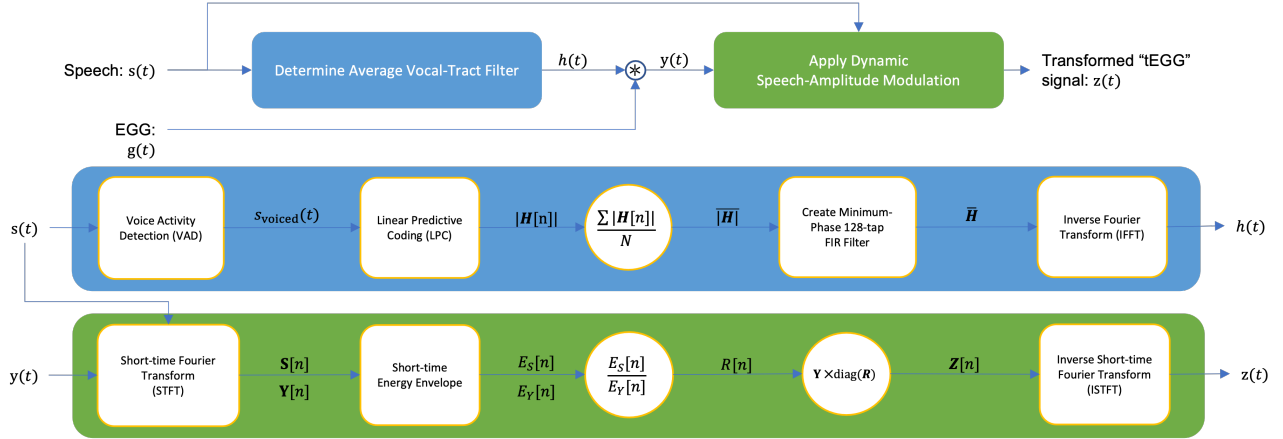[*]Author correspondence: cnoufi@ccrma.stanford.edu

Fig. 1: Proposed transformation method given speech signal $s(t)$ and EGG source signal $g(t)$. The two-stage method creates signal $z(t)$ that preserves paralinguistic content contained in the source, average vocal tract shape, and amplitude envelope, while removing linguistic content supplied by supralaryngeal articulation. The first stage (blue) extracts the LPC-based average resonant impulse response $h(t)$ representing the vocal tract during periods of vocal source activity. $h(t)$ is convolved (∗) with $g(t)$ in the second stage (green) and dynamically scaled via cross-filtering to produce $z(t)$.

introduced here for understanding and computing vocal affect.

2. TRANSFORMATION METHOD

The transformation algorithm operates in two stages (Figure 1). The first stage, represented by the blue block in Figure 1, aims to preserve the average frequency response profile of the speaker's vocal tract during voice speech production, up to, but not including, the supralaryngeal articulation. This involves iteratively estimating the filter representation of the vocal tract through LPC on short frames of audio, taking an average of the results, and then convolving this average filter representation in the time domain with the source signal captured by the EGG. The second stage, represented by the green block in Figure 1, aims to preserve the dynamic energy envelope of the original speech audio signal. This involves extracting dynamic amplitude modulation over time from the speech signal and applying it to the convolved signal.

Let a speech audio signal $s(t)$ and a corresponding EGG signal $g(t)$ both be monophonic digital audio signals with sampling rate f_s. Stage one of the algorithm proceeds as follows. An energy-based voice activity detection algorithm [19] is used to indicate if a sample at time t contains voice activity or not. Samples containing voice activity are concatenated and the remaining samples are discarded. We perform LPC on short frames of audio (assumed to be pseudo-stationary) to obtain the per-frame filter coefficients A_v, A_g, and A_l representing the vocal tract, glottal source, and lip radiation filters, respectively. This algorithm uses an Iterative Adaptive Inverse Filtering method based on a Glottal Flow Model to estimate the linear prediction coefficients of both vocal tract and glottis filters from a speech signal frame [20]. An adaptive filter with order $2 + f_s/1000$ is used to determine A_v and a 3rd-order filter to determine A_g. We apply a Hamming window to each frame of the voiced speech audio signal to obtain the short frame $s_n(t)$. To estimate the complex frequency response of the vocal tract, we set the the coefficients A_v as zeros of the complex polynomial $H_n(e^{j\omega})$ and set the polynomial representation of the poles to 1. For frames $1...N$, we apply LPC to each overlapping frame of voiced samples to obtain the vocal tract frequency response at each nth frame. We subsequently

calculate the average of the magnitudes of the complex frequency responses, $|\overline{H[n]}| = \frac{\sum_{n=1}^{N} |H_n(e^{j\omega})|}{N}$. A minimum-phase finite impulse response (FIR) filter is created given $|\overline{H[n]}|$. The phase of the FIR filter is the imaginary component of the negative of Hilbert transform of the magnitude of the full spectrum frequency response. Finally, the estimated average impulse response of the vocal tract $h(t)$ is convolved with the glottal source signal $g(t)$ in the time domain to create filtered source signal $y(t)$.

In stage two of the algorithm, we then dynamically modulate the amplitude of the convolved signal $y(t)$ to match the amplitude of the original speech signal $s(t)$. We apply a short-time Fourier transform (STFT) to obtain time-frequency representations $S[n]$ and $Y[n]$. We obtain the short-time energy envelopes $E_S[n]$ and $E_Y[n]$ of each representation by taking the sum of the magnitudes across the frequency spectrum at each time step n. We determine the dynamic modulation vector $R[n]$ to be the ratio of the short-time energy envelopes $E_S[n]$ and $E_Y[n]$. This dynamic modulation vector is then applied to $Y[n]$ via a right-hand-side multiplication to yield $Z[n]$. Equations 1 and 2 describe this cross-filtering procedure:

$$\mathbf{Z} = \mathbf{Y} \times \text{diag}(\mathbf{R}) \tag{1}$$

where $\mathbf{Z}^{F \times N}$ and $\mathbf{Y}^{F \times N}$ are real-valued matrices and $\mathbf{R}^{1 \times N}$ is a real-valued column vector. F represents the number of frequency bins f determining the resolution of the STFT. Each nth element of $\mathbf{R}$ is determined by

$$R[n] = \frac{E_S[n]}{E_Y[n]} = \frac{\sum_{f=1}^{F} |S_f[n]|}{\sum_{f=1}^{F} |Y_f[n]|} \tag{2}$$

$Z[n]$ now represents the EGG source signal that has been both filtered by the average vocal tract response and the dynamic energy envelope of the speech signal. The inverse STFT is applied to bring back $Z[n]$ into the time domain, resulting in the tEGG signal $z(t)$ in which phonetic variation has been systematically removed.

3. IMPLEMENTATION

To test the transformation method, we used a series of parallel audio and EGG recordings of affective speech made by our group as part of a related project [21]. Here we provide only the details of the protocol that apply to implementing our transformation method.

In brief, "voice actors" from the local Stanford community were recruited to produce a set of audio and EGG recordings of affective speech. Each actor was tasked with expressing 16 varieties of affect using a short emotionally neutral sentence (representative examples include "He stands on the dock" and "A bag is in the room" [22]). Speech recording sessions took place inside a sound-attenuating chamber at the Center for Computer Research in Music and Acoustics at Stanford University. After a short vocal warm-up and EGG calibration, the experimenter sounded a clapperboard (as a synchronization pulse) and proceeded to lead the actor through the 16 affect targets in a fixed order. Actors were recorded standing.

Audio recordings were made with a microphone (DPA, 466-OC-R-B00) and a digital audio recorder (Zoom H4n PRO) set to sample at 48 kHz with 24 bit-depth. The microphone was located 30 cm in front of the actor's mouth. Parallel EGG recordings were made with an electroglottagraph (Glottal Enterprises, model EG2-PCX2) and a pair of skin surface electrodes placed on either side of the larynx and held in place by an elastic strap. The EGG signal was A-to-D converted by the electroglottagraph and transmitted over USB to a computer (Apple Macmini 9,1) running Adobe Audition (v 13.0.13.46) set to sample at 48 kHz at 4 bit depth. The sound intensity level was recorded in dB using a sound level meter (NTi XL2 with M4261 microphone; frequency weighting = A, sample rate = 10 Hz).

The audio, EGG, and sound intensity tracks from each recording were synchronized in post-processing. Then, for each actor, the experimenters listened and selected a single rendition basis of its perceived conveyance of the intended affect and being free of artifacts (e.g., coughing or failed electrode contact). Next, electrical noise was removed from each selected rendition's EGG channel using Audition's "Effects→DeNoise" function and a proximate voice-free sample of the noise. EGG channels which contained substantial low-frequency energy below the fundamental frequency of the voice (due, e.g., to laryngeal movement) were additionally high-pass filtered with an adaptive cut-off set approximately 20 Hz below the fundamental. Finally, the amplitudes of the audio and EGG tracks were adjusted to accord with measured sound intensity level (at 60% scale, to avoid clipping), and the mean (calculated across all recordings) was centered at -23 LUF using the EBU R-128 loudness standard [23, 24]. The results were set between two 50 ms buffers of silence and saved as the left and right channels of a two-track .wav file.

Applying the transformation method to these tracks first required creating a copy of the speech audio signal $s(t)$ down-sampled to 16 kHz. This was necessary because the adaptive filtering procedure used in LPC is not guaranteed to produce reliable results on high frequency content. All steps to compute the average vocal tract filter response via LPC were performed on this 16 kHz signal. We used a 20 ms Hamming window, 50% overlap, and adjusted endpoints to adhere to constant-overlap-add requirements [25]. The resulting impulse response $h(t)$ was then upsampled to the original sampling rate of 48 kHz using 4th-order low-pass filter interpolation. This upsampled impulse response was then convolved with $g(t)$ and dynamic amplitude modulation was performed at 48 kHz. The two-stage transformation process runs approximately 30% faster than real-time.

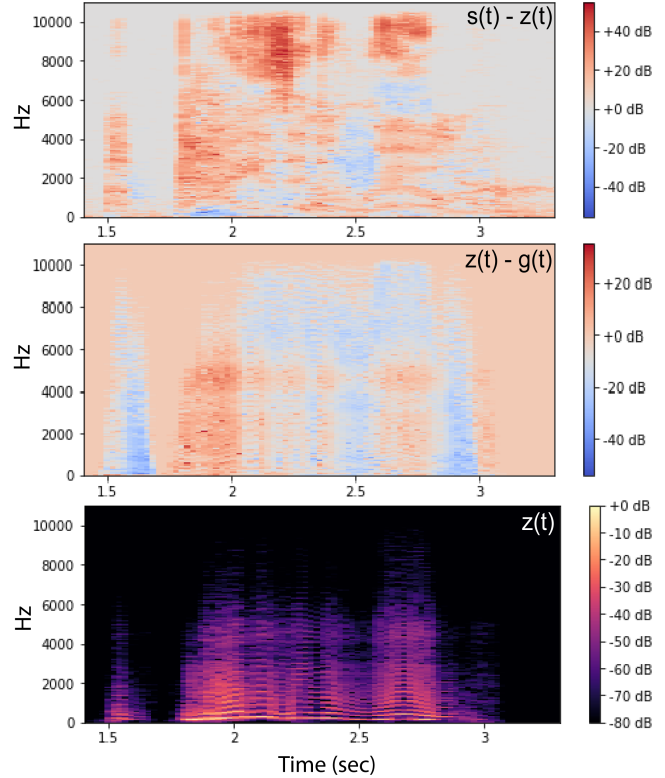The tEGG signal $z(t)$ is the resulting audio signal that retains

Fig. 2: Residual difference spectrograms for affective speech sample $s(t)$. The sample uses "A bag is in the room" as linguistic material. (Top) speech audio signal $s(t)-$ the tEGG signal $z(t)$; (Middle) $z(t)-$ EGG signal $g(t)$. (Bottom) transformed tEGG signal $z(t)$. These spectrograms detail the signal content removed and augmented through the transformation process.

the energy levels, rhythmic cadence, pitch prosody and voice quality of the speech signal at a 20 ms resolution. The residual spectrograms in Figure 2 show the acoustic content removed from $s(t)$ and added to $g(t)$ to obtain $z(t)$. As shown in Figure 2, the average resonant filter $h(t)$ applied to the source signal $g(t)$ allows the signal to retain the speaker's average "tone of voice" as dictated by anatomy of the vocal tract, but removes dynamic variations in the formants that result in phonologic variation. Instead, the result is an average of all the vowels and voiced consonants uttered in the sentence.[1]

4. EVALUATION

To determine the effects of our transformation method on perception, we conducted an online affect rating experiment. We hypothesized that affect ratings for the transformed (tEGG) signals $z(t)$ will be practically the same as those for matching original speech audio signals $s(t)$. Statistically, we tested this by comparing variation in ratings within speech-tEGG signal pairs $[s(t), z(t)]$ to variation in ratings of the same speech audio signal. We reasoned that, if variation within signal pairs is $\leq$ variation in ratings of the speech audio signal, it would imply that our transformation method successfully

---

[1]Audio examples of the transformation process, alongside comparative baselines and code repository, can be found online: `https://ccrma.stanford.edu/~cnoufi/demos/TAVA-PhonemeRemoval/`

preserves the major paralinguistic cues to affect in speech.

For this experiment, a total of 61 participants were recruited from Amazon's Mechanical Turk, via Cloud Research (mean age=40, SD =12; 31 male, 28 female, 1 unknown). Each participant rated 128 stimuli (64 $[s(t), z(t)]$ pairs). These 64 pairs were derived from a set of 64 speech audio stimuli selected from a larger set on the basis of being consistently rated for affect across listeners in a different study [21]. They included 21 different voices (10 male), with four 4 instances of each of the 16 targets.

For each trial within an experiment, a single stimulus was rated using three visual analogue scales: valence (V; from "negative" to "positive"), arousal (A, from "low-energy" to "high-energy"), and dominance (D; from "submissive" to "dominant") [26]. Stimulus order was fully randomized and participants were not shown the 16 categorical affect labels. The experiment began with 10 practice trials (half speech, half tEGG, not paired) designed to familiarize participants with the rating interface. Practice stimuli did not overlap with the 128 test stimulus set, no feedback was given, and these data are not analyzed here. Finally, a subset of 10 stimuli from the 128 test stimulus set (half speech, half tEGG) were repeated throughout the experiment. This allowed us to assess intra-rater reliability as an indicator of whether or not participants performed the task in good faith. Data from 7 of the 61 participants were discarded on these grounds to poor correlation across repeated ratings (mean intra-rater correlation coefficient < 0.5). Thus, 20,352 ratings (53 raters $\times$ 128 stimuli $\times$ 3 dimensions) were used in subsequent analyses. The average affect ratings for each signal are shown in Figure 3a.

For every speech-tEGG stimulus pair, we calculate absolute differences in affect ratings along each affect dimension (d_V, d_A, and d_D), resulting in a set of 3,392 differences (64 stimulus pairs $\times$ 53 raters) for each dimension. As a basis for comparison with these data, we estimated "baseline" variation in speech-affect perception by randomly selecting 53 pairs of ratings for each of the 64 speech signals, and calculating the differences along each dimension as above. This random sampling procedure was repeated 20 times to guard against possible sampling bias. Figure 3b shows the probability distribution of the observed speech-tEGG differences for each dimension (in gray) overlaid with a representative probability distribution of differences in affect ratings for the same speech stimuli (in color). A close correspondence between distributions in apparent in every case, as it was across all 20 repetitions.

As a basic evaluation of the statistical significance of these correspondences, we compared the speech-tEGG differences to the baseline difference using two-tailed Mann Whitney U-tests for each repetition [27]. The results were not significant for valence ($U_\mu = 5.86e6, n_1 = n_2 = 3392, P_\mu = 0.209$) or arousal ($U_\mu = 5.91e6, n_1 = n_2 = 3392, P_\mu = 0.141$), indicating that variation in affect perception between paired speech and tEGG stimuli was statistically indistinguishable from baseline variation in speech-affect perception along these dimensions. For dominance, there was a significant difference ($U_\mu = 5.93e6, n_1 = n_2 = 3392, P_\mu < 1e - 3$), but its direction was such that variation in affect perception between paired speech and tEGG stimuli was less than baseline variation in speech-affect perception (medians = 0.23 vs. 0.33, respectively). More telling than these tests is a measure of absolute effect size [28]. For valence, Cohen's d varied from -0.065 to 0.009 ($\mu = -0.036$) across the 20 repetitions; for arousal, it varied from -0.07 to -0.03 ($\mu = -0.047$); for dominance, it varied from -0.08 to -0.038 ($\mu = -0.058$). These very small effect sizes (all less than one tenth of a standard deviation) indicate that variation in affect perception between paired speech and tEGG signals is practically indistinguishable from baseline variation in speech-affect perception.
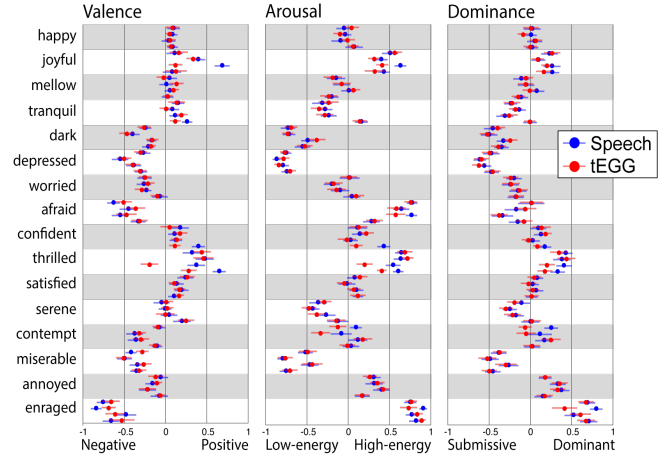
(a) Mean ratings and 95% confidence intervals for 64 speech-tEGG signal pairs.
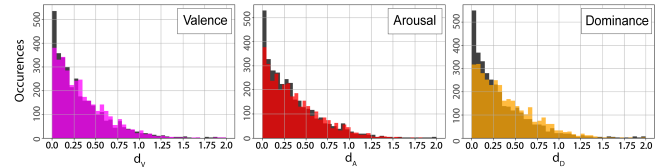
(b) Comparison of rating difference scores to speech-only baseline. The dark gray histogram displays the distribution of difference scores calculated for speech-tEGG pairs, e.g. $d_V = |V_{s(t)} - V_{z(t)}|$ for valence. The overlaid color histogram displays the distribution of difference scores for pairs of ratings (e.g. $d_{V_{\text{baseline}}} = |V1_{s(t)} - V2_{s(t)}|$ of the same speech signal $s(t)$).

Fig. 3: Affect ratings of valence (V), arousal (A) and dominance (D) given to speech signals $s(t)$ and transformed signals $z(t)$ by listeners ($N = 53$ raters).

5. CONCLUSION

In this paper, we proposed a novel signal processing method based on LPC and EGG to systematically remove phonetic content from a vocal signal while retaining as much of the signal as possible to preserve paralinguistic cues. Empirical comparison of perceived affect in response to original speech audio signals and corresponding transformed EGG signals were shown to be highly similar, indicating that little affective information is lost as a result of our transformation method. As such, the algorithm we describe provides a method for generating a synchronized dataset of affective speech, EGG, and phoneme-agnostic audio signals that can be used to further research acoustic correlates of paralinguistic vocal cues.

Finally, we note that an important limitation on the generalizability of the method described here is the necessity of having parallel speech audio and EGG recordings, the latter being relatively rare. For this reason, we are currently using a dataset of speech, EGG, and transformed EGG stimuli (only part of which has been described here) toward training an artificial neural network model to isolate paralinguistic cues from speech audio without the need for EGG.

7. REFERENCES

[1] Jody Kreiman and Diana Sidtis, "Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception," *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception*, 4 2011.

[2] Björn W. Schuller and Anton M. Batliner, *Computational paralinguistics: Emotion, affect and personality in speech and language processing*, John Wiley & Sons, Ltd, first edit edition, 2014.

[3] Alex B. Van Zant and Jonah Berger, "How the voice persuades," *Journal of Personality and Social Psychology*, vol. 118, no. 4, pp. 661–682, 4 2020.

[4] K. O'Connor, "Auditory processing in autism spectrum disorder: A review," *Neuroscience and Biobehavioral Reviews*, vol. 36, no. 2, pp. 836–854, 2 2012.

[5] C. Naranjo, C. Kornreich, S. Campanella, X. Noël, Y. Vandriette, B. Gillain, X. De Longueville, B. Delatte, P. Verbanck, and E. Constant, "Major depression is associated with impaired processing of emotion in music as well as in facial and vocal stimuli," *Journal of Affective Disorders*, vol. 128, no. 3, pp. 243–251, 2 2011.

[6] Peter Ladefoged, *Elements of Acoustic Phonetics*, University of Chicago Press, 2019.

[7] Marco Guzman, Soledad Correa, Daniel Muñoz, and Ross Mayerhoff, "Influence on Spectral Energy Distribution of Emotional Expression," *Journal of Voice*, vol. 27, no. 1, pp. 1–129, 1 2013.

[8] Kenneth C. Hill and Peter Ladefoged, "A Course in Phonetics," *Language*, vol. 53, no. 4, pp. 911, 1977.

[9] Adamantios Gafos and Pascal van Lieshout, "Editorial: Models and Theories of Speech Production," 6 2020.

[10] Zhaoyan Zhang, "Mechanics of human voice production and control," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 2614–2635, 10 2016.

[11] Jody Kreiman, Yoonjeong Lee, Marc Garellek, Robin Samlan, and Bruce R Gerratt, "Validating a psychoacoustic model of voice quality," *The Journal of the Acoustical Society of America*, vol. 149, no. 1, pp. 457–465, 2021.

[12] Klaus R. Scherer, "Personality inference from voice quality: The loud voice of extroversion," *European Journal of Social Psychology*, vol. 8, no. 4, pp. 467–487, 1978.

[13] Andrey Anikin, "A Moan of Pleasure Should Be Breathy: The Effect of Voice Quality on the Meaning of Human Nonverbal Vocalizations," *Phonetica*, vol. 77, no. 5, pp. 327–349, 2020.

[14] Douglas O. Shaughnessy, "Linear predictive coding," *IEEE Potentials*, vol. 7, no. 1, pp. 29–32, 1988.

[15] F. L.E. Lecluse, M. P. Brocaar, and J. Verschuure, "The electroglottography and its relation to glottal activity," *Folia Phoniatrica*, vol. 27, no. 3, pp. 215–224, 1975.

[16] D. G. Childers, A. M. Smith, and G. P. Moore, "Relationships between electroglottograph, speech, and vocal cord contact," *Folia Phoniatrica et Logopaedica*, vol. 36, no. 3, pp. 105–118, 1984.

[17] Ki Hwan Hong and Hyun Ki Kim, "Electroglottography and laryngeal articulation in speech," *Folia Phoniatrica et Logopaedica*, vol. 49, no. 5, pp. 225–233, 1997.

[18] Christian T. Herbst, "Electroglottography – An Update," *Journal of Voice*, vol. 34, no. 4, pp. 503–526, 7 2020.

[19] Mike Brookes, "Voicebox: Speech Processing Toolbox for Matlab," 2011.

[20] Olivier Perrotin and Ian McLoughlin, "A Spectral Glottal Flow Model for Source-filter Separation of Speech," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2019, pp. 7160–7164.

[21] Daniel L. Bowling, Camille Noufi, Jonathan Berger, Michael Frank, and Karen Parker, "The test of auditory-vocal affect (tava): a clinical assessment of sensitivity to non-verbal cues in spoken communication," *Stanford University School of Medicine*, In preparation.

[22] Boaz M. Ben-David, Pascal H.H.M. Van Lieshout, and Talia Leszcz, "A resource of validated affective and neutral sentences to assess identification of emotion in spoken language after a brain injury," *Brain injury*, vol. 25, no. 2, pp. 206–220, 2011.

[23] The MathWorks Inc., "Matlab audio toolbox," 2020a.

[24] "Loudness Normalization and Permitted Maximum Level of Audio Signals," Tech. Rep. August, EBU, Geneva, 2020.

[25] Julius O. Smith, *Spectral Audio Signal Processing*, http://ccrma.stanford.edu/~jos/sasp/, 2011, Online Book, 2011 Edition.

[26] Petri Laukka, Patrick N. Juslin, and Roberto Bresin, "A dimensional approach to vocal expression of emotion," *Cognition and Emotion*, vol. 19, no. 5, pp. 633–653, 8 2005.

[27] The SciPy community, "scipy.stats.mannwhitneyu," 2022.

[28] Jacob Cohen, *Statistical Power Analysis for the Behavioral Sciences*, Routledge, 5 2013.