INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Influence of Speaker-Specific Parameters on Speech Separation Systems

David Ditter, Timo Gerkmann

Signal Processing, Universität Hamburg, Germany
david.ditter@uni-hamburg.de, timo.gerkmann@uni-hamburg.de

Abstract

Recent studies have shown that Deep Learning based single-channel speech separation systems perform worse for same-gender mixtures than for different-gender mixtures. In this work, we provide for a more detailed analysis of the respective impact of the fundamental frequency and the vocal tract length on the system performance. While both parameters are correlated with gender, the vocal tract length is a fixed speaker-specific parameter, whereas the fundamental frequency can vary for different speaking styles. We show that the difference of the fundamental frequency medians of two speakers in a mixture is highly correlated with the SDR performance while the difference of the vocal tract lengths is not. Our analysis allows us to do performance predictions for given speakers based on measurements of their fundamental frequency. Furthermore we conclude that current systems separate (short-term) speaking styles rather than (long-term) speaker characteristics.

Index Terms: Chimera++, Speech Separation, Deep Clustering, Permutation Invariant Training, Fundamental Frequency, Vocal Tract Length

1. Introduction

Since the upcoming of Deep Learning techniques, speech separation systems have improved significantly such that they are now useful for real-world scenarios. A multitude of approaches have been proposed, such as Deep Clustering [1], Permutation Invariant Training (PIT) [2, 3], Deep Attractor Networks [4, 5] and End-to-End systems [6, 7]. While the average separation performance for a mixture of two speakers has risen to quite impressive levels, we see a large variance of the performance for different mixture examples. For instance, publications on speech separation have reported a performance gap between same-gender and different-gender mixtures [8]. For a pair of same-gender speakers, we know that systems will perform statistically below average and this constellation is not a rare corner case but instead something that will definitely occur frequently in the real-world.

For a speech separation system to be accepted in real-world speech communication systems, only optimizing for the average performance is thus not sufficient. Instead, it is of utmost importance to avoid constellations where the system fails, as indicated e.g. by low or even negative Source-to-Distortion Ratio (SDR)-improvement scores. To be able to improve systems to avoid such negative outliers, we need to understand which parameters of the competing speakers are dominant to predict the system performance. Thus, the goal of this paper is to provide for a deeper understanding of these limiting factors of state-of-the-art speech separation systems. In particular, we investigate if the dominant factor is the time-*in*variant speaker-specific Vocal Tract Length (VTL), or if it is rather the time-*variant* fundamental frequency. Note that this is an important difference, as the fundamental frequency changes (1) within a sentence (intonation), (2) in different environments (Lombard effect), but (3) can also voluntarily be changed and imitated. Thus, if the VTL were the dominant parameter for system performance, we would be able to conclude that the performance of the system is truly speaker dependent. However, if the fundamental frequency is dominant for performance, the system performance also strongly depends on the *speaking-style*. As both parameters can be estimated from signals, the results of this work also provide interesting insights to develop algorithms that predict the performance of a speech separation system in real-time. Such a prediction can for instance allow to turn the speech separation system off, if negative performance is likely.

We start our analysis with an overview of state-of-the art speech separation systems (Section 2) and the introduction to the considered speaker-specific parameters (Section 3). In Section 4, we analyze the influence of the VTL and the fundamental frequency on the performance of speech separation systems. We will then present and discuss results (Section 5) and draw conclusions in Section 6.

2. Deep Learning based Speech Separation

Our speech separation experiments are based on the Chimera++ network proposed in [9] which was introduced with slight variations under the name Chimera in [10] and separates speech mixtures via time-frequency masking [11]. We choose Chimera++, as it combines the strengths of Deep Clustering and PIT.

For both Deep Clustering and PIT, an additive mixture of sources in the Short-Time Fourier Transform (STFT) domain is assumed, i.e.

$$X(k,l) = \sum_{c=1}^{C} S_c(k,l) \tag{1}$$

where $X(k,l)$ and $S_c(k,l)$ denote the complex STFT values at frequency bin k and time index l for the mixture and C source signals. The goal is then to find so called STFT masks M_c for each source signal such that we have

$$\hat{S}_c(k,l) = X(k,l)\; M_c(k,l), \tag{2}$$

where $\hat{S}_c(k,l)$ is an estimation of the original source signal $S_c(k,l)$. The estimated time-domain signal of speaker c is reconstructed by applying the inverse STFT to $\hat{S}_c$.

To obtain the speaker masks, in PIT a Deep Neural network is trained which uses the mixture magnitudes $|X(k,l)|$ as the input and the ground truth STFT data of the source signals as targets to compute the training loss as

$$\mathcal{L}_{\mathrm{PIT}} = \min_{\pi\in\mathcal{P}} \sum_{c=1}^{C}\sum_{k=1}^{K}\sum_{l=1}^{L} \left|\,|X(k,l)|\; M_c(k,l) - |S_{\pi(c)}(k,l)|\,\right|^2, \tag{3}$$

where $\mathcal{P}$ is the set of permutations on $\{1, ..., C\}$, $|S_c|$ the magnitude of the c-th reference source, K is the total number of

Copyright © 2019 ISCA

http://dx.doi.org/10.21437/Interspeech.2019-2459

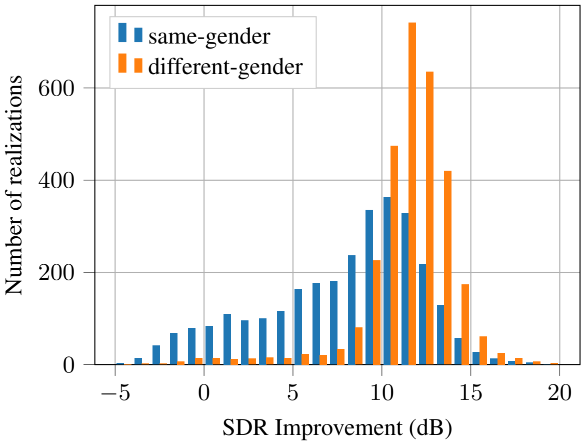

Figure 1: *SDR improvement comparison for same-gender mixtures vs. different-gender mixtures for the WSJ0-MIX2-0dB test set. Performance is significantly worse for same-gender mixtures.*

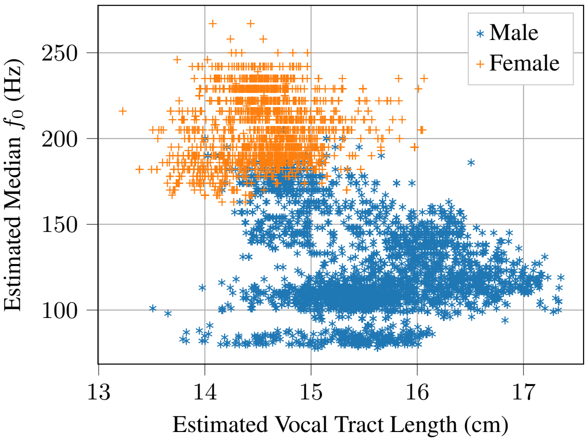

Figure 2: *Estimated Vocal Tract Length vs. Estimated Median Fundamental Frequency f_0 for 6000 utterances of the WSJ0-MIX2 test set. Samples are marked differently for male and female speakers.*

frequency bins and L is the number of time frames. The basic principle of this loss function is to minimize the error that the output masks produce independent of the permutation of the generated masks.

Deep Clustering uses a Deep Neural Network to map each STFT bin to an embedding space of dimensionality D. The loss function drives the network to map STFT bins of different speakers to orthogonal regions of the embedding space. Separation could now be done using a simple k-means clustering. Instead, the Chimera++ network, as proposed in [9], uses the PIT objective to obtain the speaker masks directly from the Neural Network while using the Deep Clustering training objective as a second target at training time. In this setup, the Deep Clustering objective works as a regularizer and improves the quality of the masks generated at the PIT output.

The performance of speech separation systems is commonly measured with the SDR [12] which compares a distorted source signal with the ground truth source signal. As part of the measurement, three types of distortion signals (interference, noise, artifacts) are computed and combined as

$$\mathrm{SDR} := 10 \log_{10} \frac{||s_{\text{target}}||^2}{||e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}||^2}. \qquad (4)$$

For a processed mixture signal, we measure the separation performance as the SDR improvement which compares the SDR of the unprocessed mixture signal to the SDR of the estimated source signal.

3. Considered Speaker-Specific Parameters

[8] reports the average SDR improvements for same-gender mixtures versus mixtures of opposite genders for Deep Clustering. Results show a significant performance gap for same-gender mixtures of around 3.1 dB of average SDR improvement. In Figure 1 we plot the histogram of the SDR improvement for same-gender and different-gender mixtures for our implementation of Chimera++ which is described in detail in section 4.3. As reported in [8], it shows that the average SDR improvement differs for same-gender as compared to different-gender mixtures. It is notable that also the standard deviation is much higher for same-gender mixtures and we see more negative outliers. This observation is the core motivation for a detailed analysis of speaker-specific influencing factors on the performance of speech separation systems in this work. For the analysis on the influence of speaker-specific characteristics, we introduce two parameters which are correlated to gender, namely the fundamental frequency f_0 and the VTL.

The VTL is a physiological feature of a person and can be measured via Magnetic Resonance Imaging. It is measured as the length of the curved tube starting at the vocal chords and ending at the mouth entrance. A method to estimate the VTL from a speech signal is described in detail in section 4.4. Note that the VTL is a fixed time-invariant parameter per speaker.

The fundamental frequency is a time-varying parameter for a fixed speaker. It can only be measured during voiced frames of speech which are characterized by a periodic excitation signal in the source-filter model of speech (compare e.g. [13, p. 10ff]). In an ideal speech signal, f_0 is the inverse of the period in which the signal repeats itself on a small time scale. In any real-world speech signal, f_0 will show variations within phrases and even within single vowels. For example, usually at the end of a question f_0 will increase.

We choose to analyze the impact of VTL and f_0 on the SDR performance for the following reasons: First, the parameters are highly correlated with gender and we know that gender influences the SDR performance. In Figure 2 we plot the median vocal tract length and the fundamental frequency for 6000 utterances of the WSJ0-MIX2 test set [1]. It shows that the distribution of VTL is significantly different for male than for female speakers and the same is true for f_0. Second, within a gender category the two parameters show little correlation which is illustrated by the distribution of points for a certain color in Figure 2. Due to this the two parameters may show very different influences on the SDR performance. And thirdly, while VTL is a static parameter for a fixed speaker, f_0 in contrast is a time-varying parameter, so an influence of either one of them would have different implications as discussed in sections 1 and 5.

4. Analysis Framework

Now that we have discussed the theoretical basics and our motivation, in this section we give a detailed description of our setup for the analysis.

4.1. Chimera++ Setup

As training data, we use the 20 hour training set of the publicly available WSJ0-MIX2 data set [1] at 8 kHz sample rate which contains two speaker mixtures mixed at a power ratio uniformly distributed between 0 to 10 dB. For preprocessing, all utterances of the data set are analyzed with an STFT with a window size of 32 ms and a hop size of 8 ms using the square root Hann window. As the last step of preprocessing, we compute the log-magnitude values of the complex STFT data. The data is then cut into pieces of 400 frames which equals roughly 3.2 seconds per meta-frame, where each frame holds 129 frequency bins.

The network architecture is taken from [9] and begins with 4 Bidirectional LSTM (BLSTM) layers with 600 units for each layer and direction. The last BLSTM layer feeds into two separate fully connected layers, namely (1) the Deep Clustering output layer with $40 \times 129 \times 400$ units where 40 is the embedding dimension D and (2) the PIT output layer with $2 \times 129 \times 400$ units where 2 is the number of speakers. As the training target for PIT, we use the truncated phase-sensitive spectrum approximation technique [9]. For the Deep Clustering regularization objective we use ideal binary masks and a weighting scheme with magnitude ratio masks as proposed in [9]. The α parameter of Chimera++, which weights the two objectives, was determined empirically to weight the Deep Clustering objective function with $\alpha = 0.9975$ and the PIT objective function with $1 - \alpha = 0.0025$. The Deep Clustering output layer is discarded at test time.

This setup leads to an average SDR improvement of 10.0 dB on the WSJ0-MIX2 test set where, for comparison, [9] reports 11.2 dB improvement for a similar setup. We assume that this gap comes from minor implementational details and from [9] using the so-called whitened k-means objective which we have not utilized. Nonetheless, we strongly believe that the findings based on our implementation are transferable to other Chimera++ implementations.

4.2. SDR Evaluation

To evaluate the SDR performance we use a modified version of the test set of the WSJ0-MIX2 data set. The original set contains mixtures at power ratios from 0 to 10 dB, but this variable is an influencing factor on the SDR performance. For a more controlled experiment of influencing factors to the SDR performance we need to discard it. Our modified WSJ0-MIX2-0dB set is hence the exact same set as the original except that the power ratio within a mixture is always at 0 dB. By only modifying this parameter compared to the original WSJ0-MIX2 we ensure that our test set holds no utterances or speakers that were used during training.

4.3. Fundamental Frequency Estimation

The fundamental frequency was measured using an auto-correlation based method that additionally reduces octave errors by introducing an octave jump cost function [14]. It was evaluated with a step size of 10 ms using the Praat Speech Analyzer Software [15]. For a final estimation for a full utterance we use the median f_0 over all voiced frames of the utterance.

Table 1: *Formant weighting coefficients for VTL estimation as proposed in [16]*

β_1	β_2	β_3	β_4
0.022	0.136	0.254	0.637

4.4. Vocal Tract Length Estimation

To estimate the VTL we make use of methods described in [16] which are based on formant measurements and a physical model of the vocal tract. For voiced speech frames, we can get an estimate of the vocal tract length with the help of the lowest resonance frequency of a lossless uniform vocal tract Φ, which depends on the n-th formant F_n as

$$\Phi = \frac{F_n}{2n-1}. \tag{5}$$

To map multiple formants to a single quantity, a weighted sum is used as

$$\hat{\Phi} = \sum_{n=1}^{N} \frac{\beta_n F_n}{2n-1}. \tag{6}$$

In practice, we consider the first four formants ($N = 4$) and set β_n as proposed in [16] and shown in Table 1.

Once we have computed a resonance frequency estimate $\hat{\Phi}$ we can infer the VTL estimate using

$$L = \frac{c}{4\hat{\Phi}}, \tag{7}$$

where $c = 350\,\mathrm{m/s}$ denotes the speed of sound within the vocal tract.

To get a VTL estimate for a full utterance, we first detect signal frames with voiced speech, then estimate the formants at voiced speech frames, infer the VTL from measurements with help of equations (6) and (7), and finally take the median value over all voiced frames. To detect the voiced speech frames and formants we again use Praat [14, 15] where the formant analysis is based on a short-term Linear Predictive Coding (LPC) analysis [17]. To measure the formants F_1 to F_4 we use a version of the WSJ0-MIX-0dB data set at 16000 Hz sample rate. This sample rate is necessary because formant F_4 usually lies in the region of 3000 to 4300 Hz and therefore cannot be tracked correctly in a signal with 8000 Hz sample rate.

5. Results and Discussion

We now analyze how differences in the VTL and f_0 impact the SDR performance for a two-speaker mixture. If we denote the median fundamental frequency of speaker n by $f_{0,n}$ and the respective vocal tract length by VTL_n, these differences are defined by $\Delta f_0 = |f_{0,1} - f_{0,2}|$ and $\Delta\mathrm{VTL} = |\mathrm{VTL}_1 - \mathrm{VTL}_2|$.

In Figure 3, we plot the SDR improvement versus Δf_0 for 3000 mixtures of the WSJ0-MIX2-0dB test set. It shows that Δf_0 is an important influencing factor to the SDR performance. When the Δf_0 is above 60 Hz, the SDR performance of the system is at its best, with a mean value of 12.0 dB SDR improvement and a relatively low standard deviation. There are no far negative outliers for this region of Δf_0 and all mixtures show an SDR improvement of 5.6 dB or higher. For Δf_0 values below 60 Hz we see a steady decline of the mean SDR improvement and it is lowest for Δf_0 close to 0 Hz. The standard deviation goes up when the Δf_0 gets smaller and in this

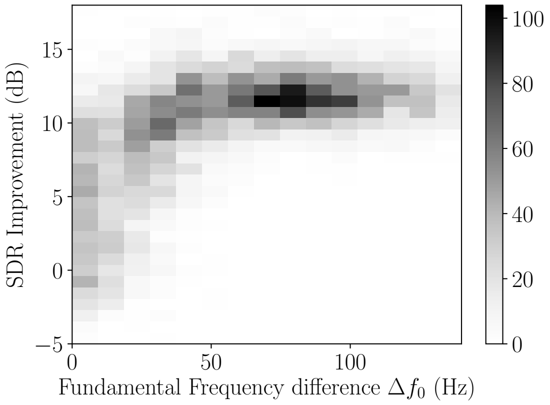

Figure 3: *2-D Histogram of the number of realizations of a certain SDR improvement against Δf_0 for the 3000 mixtures of the WSJ0-MIX2-0dB test set. A strong decline of the SDR performance can be seen for the Δf_0 below* 60 Hz. *Above* 60 Hz *there are no far outliers. Correlation factor: 0.58.*

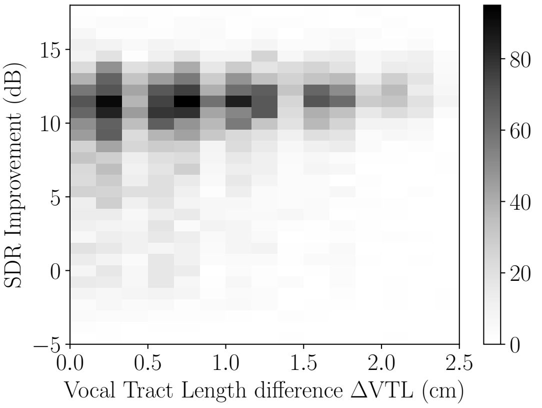

Figure 4: *2-D Histogram of the number of realizations of a certain SDR improvement against vocal tract length differences for the 3000 mixtures of the WSJ0-MIX2-0dB test set. A strong decline as in Fig. 3 cannot be seen here. Correlation factor: 0.23.*

region there are hardly any mixtures with a separation performance that reaches the average performance for mixtures with Δf_0 above 60 Hz. The correlation factor between Δf_0 and the SDR improvement is at 0.58. If we only take into account mixtures with $\Delta f_0 < 60$ Hz, we have an even higher correlation of 0.66. The limit at 60 Hz is reasonable as there the system seems to reach an upper performance bound.

If we do the same analysis for the VTL difference we can see that this gender-dependent parameter is much less an influence to the SDR performance than the Δf_0. In Figure 4, we plot the SDR improvement versus ΔVTL under the same conditions as in Figure 3. We see that throughout all regions of ΔVTL the median of the distribution does not significantly change. We can furthermore observe that for lower ΔVTL we have a higher number of negative outliers but the negative influence of similar vocal tract lengths (small ΔVTL) is significantly smaller than in the case of f_0. The correlation factor of ΔVTL and SDR improvement is only at 0.23 and thus much lower than for the Δf_0.

Our results provide for a deeper understanding of the known performance differences for same-gender and different-gender mixtures. We have shown that while the parameter f_0 that is correlated with gender has an important influence on the performance, the gender-correlated parameter VTL influences the performance very weakly. Thus, we are now able to make predictions about the SDR performance for a mixture if we know the median f_0 of the respective speakers. This is an improvement to the gender category which should be illustrated with a short example: Assume that we have a male speaker with a relatively high-pitched voice and a female speaker with a relatively low-pitched voice. From the formerly known performance differences for same-gender and different-gender we may expect that the separation system should perform above its average performance for this mixture. But given the set of speakers as described, our new findings suggest that the system will perform below its average performance, because we have two speakers with a similar median f_0. In the opposite case where we have two speakers with a largely different median f_0 (i.e. $\Delta f_0 > 60$ Hz) we now expect that the system makes no grave errors as we have not seen any negative outliers in our measurements. Again, for the same-gender/different-gender categorization we are not able to make such strong predictions.

Our results show furthermore that the dominant factor is the time-*variant* fundamental frequency and not the time-*in*variant speaker-specific VTL. As mentioned above, the fundamental frequency of a speaker may change within a sentence (intonation), in different environments (e.g. due to the Lombard effect) and can also be varied on purpose. Thus, we conclude that the system performance depends not only on the speakers within a mixture but also on the (time-varying) *speaking-style* of these speakers. By means of imitating the fundamental frequency this also implies that separation systems can be willingly undermined.

6. Conclusions

With our analysis we have shown that the speaker-specific median f_0 can be utilized to make predictions about the SDR performance of a speech separation system while the VTL can not. By knowing the f_0 of speakers in a mixture we are able to make stronger predictions on the SDR performance than the predictions based on same-gender/different-gender categories. We have also shown that the dominant influencing factor is the time-varying fundamental frequency and we conclude that the performance of speech separation systems does not only depend on the speakers but also on the current speaking-style of the speakers. In future work, it should be explored why f_0 has this strong influence on the performance and ideally we find methods that can exploit this knowledge to improve the weaknesses of current systems. Also, future work should be aware of the dependence of the performance on speaking styles and try to avoid producing far negative performance outliers for any combination of speakers *and* speaking styles.

7. References

[1] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative Embeddings for Segmentation and Separation," in *2016 IEEE International Conference on Acoustics,*

Speech and Signal Processing (ICASSP), Shanghai, China, Mar. 2016, pp. 31–35.

[2] D. Yu, M. Kolbaek, Z.-H. Tan, and J. Jensen, “Permutation Invariant Training of Deep Models for Speaker-Independent Multi-Talker Speech Separation,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, Mar. 2017, pp. 241–245.

[3] M. Kolbæk, D. Yu, Z. Tan, and J. Jensen, “Multitalker Speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[4] Z. Chen, Y. Luo, and N. Mesgarani, “Deep Attractor Network for Single-Microphone Speaker Separation,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, Mar. 2017, pp. 246–250.

[5] L. Drude, T. v. Neumann, and R. Haeb-Umbach, “Deep Attractor Networks for Speaker Re-Identification and Blind Source Separation,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 11–15.

[6] Y. Luo and N. Mesgarani, “TaSNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 696–700.

[7] Z.-Q. Wang, J. L. Roux, D. Wang, and J. Hershey, “End-to-End Speech Separation with Unfolded Iterative Phase Reconstruction,” in *Proc. Interspeech 2018*, Hyderabad, India, Sep. 2018, pp. 2708–2712.

[8] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, “Single-Channel Multi-Speaker Separation Using Deep Clustering,” in *Proc. Interspeech 2016*, San Francisco, USA, Sep. 2016, pp. 545–549.

[9] Z. Wang, J. L. Roux, and J. R. Hershey, “Alternative Objective Functions for Deep Clustering,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, Apr. 2018, pp. 686–690.

[10] Y. Luo, Z. Chen, J. R. Hershey, J. L. Roux, and N. Mesgarani, “Deep Clustering and Conventional Networks for Music Separation: Stronger Together,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, Mar. 2017, pp. 61–65.

[11] O. Yilmaz and S. Rickard, “Blind Separation of Speech Mixtures via Time-Frequency Masking,” *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[12] E. Vincent, R. Gribonval, and C. Fevotte, “Performance Measurement in Blind Audio Source Separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[13] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*. John Wiley & Sons, Aug. 2006.

[14] P. Boersma, “Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-To-Noise Ratio of a Sample Sound,” *Proceedings of the Institute of Phonetic Sciences*, vol. 17, no. 1193, pp. 97–110, 1993.

[15] P. Boersma and V. van Heuven, “Speak and unSpeak with PRAAT,” *Glot International*, vol. 5, no. 9, pp. 341–347, Nov. 2001.

[16] A. C. Lammert and S. S. Narayanan, “On Short-Time Estimation of Vocal Tract Length from Formant Frequencies,” *PLOS ONE*, vol. 10, no. 7, Jul. 2015.

[17] A. Gray and D. Wong, “The Burg Algorithm for LPC Speech Analysis/Synthesis,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 6, pp. 609–615, Dec. 1980.