Nonavinakere Prabhakera, Narendra; Airaksinen, Manu; Alku, Paavo

Glottal source estimation from coded telephone speech using a deep neural network

Published in:
Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH

DOI:
10.21437/Interspeech.2017-882

Published: 01/08/2017

Document Version
Publisher's PDF, also known as Version of record

INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Glottal source estimation from coded telephone speech using a deep neural network

N P Narendra, Manu Airaksinen, Paavo Alku

Aalto University, Department of Signal Processing and Acoustics, Espoo, Finland
{narendra.prabhakera, manu.airaksinen, paavo.alku}@aalto.fi

Abstract

In speech analysis, the information about the glottal source is obtained from speech by using glottal inverse filtering (GIF). The accuracy of state-of-the-art GIF methods is sufficiently high when the input speech signal is of high-quality (i.e., with little noise or reverberation). However, in realistic conditions, particularly when GIF is computed from coded telephone speech, the accuracy of GIF methods deteriorates severely. To robustly estimate the glottal source under coded condition, a deep neural network (DNN)-based method is proposed. The proposed method utilizes a DNN to map the speech features extracted from the coded speech to the glottal flow waveform estimated from the corresponding clean speech. To generate the coded telephone speech, adaptive multi-rate (AMR) codec is utilized which is a widely used speech compression method. The proposed glottal source estimation method is compared with two existing GIF methods, closed phase covariance analysis (CP) and iterative adaptive inverse filtering (IAIF). The results indicate that the proposed DNN-based method is capable of estimating glottal flow waveforms from coded telephone speech with a considerably better accuracy in comparison to CP and IAIF.

Index Terms: Glottal source estimation, glottal inverse filtering, deep neural network, telephone speech

1. Introduction

Glottal inverse filtering (GIF) is a method for estimating the voice source or glottal source from a recorded microphone speech signal. GIF methods assume the source-filter model [1] for speech production. In this model, the source refers to the voice source signal generated by quasi-periodic fluctuation of the vocal folds. The filter refers to the time-varying digital filter formed by a particular shape of vocal tract system. In GIF methods, the effects of the vocal tract formants are canceled by applying anti-resonances to the segment of recorded speech. The estimation of the glottal source signal is crucial as it carries important information related to type of phonation and pitch which can be related to various paralinguistic cues such as emotional state, individual speaker characteristics and possible voice pathologies.

Several GIF methods have been developed in the past decades and known examples are, for example, closed phase covariance analysis (CP) [2], iterative adaptive inverse filtering (IAIF) [3] and complex cepstral decomposition (CCD) [4]. These methods have been typically used in ideal conditions where speech recordings take place in an anechoic chamber and, most importantly, the input signal is recorded with very good audio quality (i.e., using a high-quality condenser microphone of a flat amplitude response and linear phase) [5]. The accuracy of GIF methods in more realistic scenarios, such as in noise or when the input signal is distorted by quantization and band-pass filtering, cannot be assumed to be same as in ideal conditions. In these realistic scenarios, the performance of GIF methods deteriorates severely. Even for the most powerful state-of-the-art methods such as CP, CCD, and quasi closed phase analysis (QCP) [6], degradation in the accuracy can be observed. These methods require additional parameter estimations (e.g. extraction of glottal closure instants (GCIs)) which may suffer from noise [7] and other distortions [8]. Hence, there exists a need for robust GIF methods which can accurately estimate glottal flow waveforms in more realistic conditions, for example, when the input signal is coded telephone speech. This is necessary as telecommunication networks have been extensively deployed in the recent years, and there is growing need for the speech processing tasks to be performed remotely after the transmission of speech. To the best of our knowledge, glottal source estimation has not been explored before from coded telephone speech due to known strict quality requirements of the GIF analysis. Robust estimation of glottal source from coded speech has potential applications in speaker recognition, emotion recognition and in biomedical applications such as detection and classification of neurodegenerative diseases.

Deep neural networks (DNNs) are widely used powerful tools for finding nonlinear relations between input and output features [9]. DNNs have been studied in a few recent investigations as an alternative to conventional GIF methods to compute glottal source waveforms. DNNs were shown to provide significant improvement in glottal source estimation compared to state-of-the-art methods [10][11][12]. In statistical parametric speech synthesis [10][11][12], a DNN is used to create a mapping between acoustic speech features and time-domain glottal flow waveform. In the study by Airaksinen et al. [13], a DNN was utilized for estimating the glottal flow waveform by using acoustic features extracted from speech corrupted by additive noise. This study showed that the DNN was capable of generating noise robust estimates of the glottal flow signal. The advantage of using DNN is that it can accurately predict the glottal flow waveforms by using low level speech features. The input speech features which include spectrum and fundamental frequency can be easily and robustly extracted from speech.

This paper proposes a DNN-based glottal source estimation method that uses coded telephone speech as input. During training, both coded and clean, high-quality versions of speech are utilized. DNN is used to map the speech features extracted from coded speech with the corresponding glottal flow waveforms extracted with a GIF method from clean speech. The main contribution of the present work is the estimation of glottal source signal from coded telephone speech using a DNN-based approach, which has not been explored in previous studies. Background information about the inverse filtering method is provided in Section 2. The proposed method is described in detail in Section 3. The details about the speech database, experimental setup and results are provided in Section 4. The

http://dx.doi.org/10.21437/Interspeech.2017-882

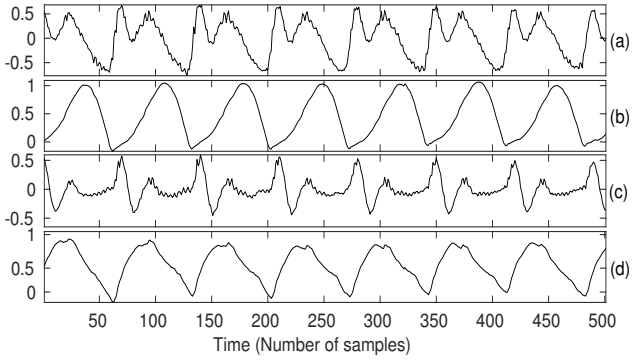

Figure 1: *(a) Clean speech, (b) glottal flow waveform estimated from clean speech, (c) coded speech, and (d) glottal flow waveform estimated from coded speech.*

summary of present work and discussion are given in Section 5.

2. Background

A new GIF method, Quasi-Closed Phase Analysis (QCP), was recently proposed in [6] and compared with several known reference techniques. Experiments in [6] indicate that QCP can be regarded as one of the most accurate GIF methods to estimate the glottal source from clean, high-quality speech signals. The QCP method is based on the principles of the closed phase analysis [2], which estimates the vocal tract transfer function during the closed phase of the glottal cycle. In contrast to the CP method, QCP does not compute the vocal tract response using the covariance method from few samples located in the closed phase. Instead, QCP creates a specific temporal weighting function, called the Attenuated Main Excitation (AME) function, using GCIs estimated from speech. The AME function is used to attenuate the contribution of the (quasi-) open phase in the computation of the Weighted Linear Prediction (WLP) coefficients, which results in good estimates of the vocal tract transfer function. Evaluation results in [6] show that the accuracy of QCP is better than that of existing methods such as CP, CCD and IAIF. However, QCP calls for conducting an additional parameter estimation (i.e., detection of GCIs) which makes the method sensitive to noise and other distortions (e.g. audio degradation caused by microphones of poor frequency responses).

Figure 1 demonstrates how using simulated telephone transmission (i.e., band-pass filtering, low bit-rate coding, for more details see Section 3) as a source of speech degradation affects the glottal waveform estimated by GIF. The figure shows the clean speech (Figure 1(a)), the glottal flow computed from the clean speech by QCP (Figure 1(b)), the coded speech signal (Figure 1(c)) and the glottal flow computed from the coded signal by QCP (Figure 1(d)). Coding corrupts the GIF input by generating different types of amplitude and phase distortion and quantization noise. As a result of this, the shape of the flow waveform estimated from the coded speech deviates greatly from the corresponding, plausible-looking glottal waveform computed from the clean input.

3. Proposed method

To accurately estimate the glottal flow waveform from coded speech, a data driven, DNN-based approach is proposed. This approach utilizes both clean and coded versions of speech in training. In the training phase, the glottal flow waveform is estimated from the clean speech by using a GIF method. Simultaneously, speech parameters are extracted from the corresponding frames of the coded speech signal. The speech parameters extracted from the coded speech and the reference glottal flow waveforms obtained from the clean speech are mapped using a DNN.

The block diagram of the proposed method is shown in Figure 2. First, a high quality multi-speaker speech database is considered for estimation of glottal flow waveforms. The amount of data should be sufficient for training the DNN. Next, glottal inverse filtering is applied to the clean speech to estimate the glottal flow during the voiced segments of the signal. In the current study, QCP is used for glottal inverse filtering. In order to compute the AME function, GCIs are detected using the SE-DREAMS method [14]. The glottal source is computed as the derivative of the flow, decomposed into two-pitch-period long segments, then windowed with the Hann window, normalized in energy, and zero-padded to constant length.

In order to simulate telephone speech, utterances from the database are coded using the adaptive multi-rate (AMR) codec [15] which is a widely used speech compression method standardized by the European Telecommunications Standards Institute (ETSI) [16]. AMR is a narrowband codec limiting the signal frequency range to 300-3400Hz. The speech features extracted from the AMR-coded telephone speech are then mapped via a DNN to the corresponding glottal flow segments estimated from the clean speech. The speech features considered can in principle be any acoustic representations of speech. The most commonly used speech features are the spectrum and fundamental frequency which both contain information about the voice source.

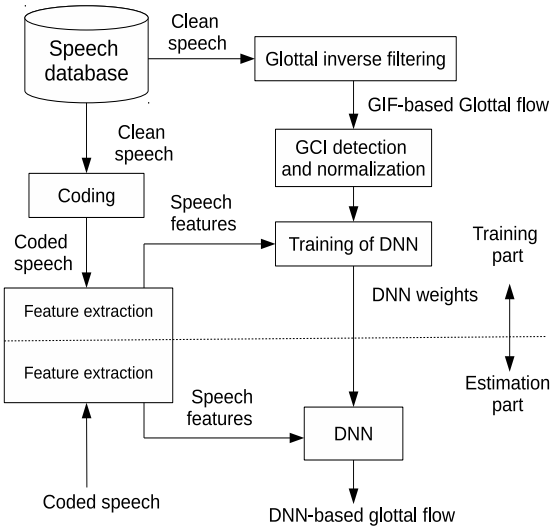

Figure 2: *Block diagram of the proposed DNN-based glottal source estimation.*

The DNN consists of an input layer, three hidden layers and an output layer. The size of the input and output layers depends on the dimension of the input feature vector and the length of the glottal flow waveform. As 8 kHz sampling rate is considered, the total length of the glottal flow segment is fixed to 300 samples, resulting in 300 neurons in the output layer. The number neurons in the three hidden layers are fixed to 250, 150 and 250. Sigmoid activation functions are used in the hidden layers and linear activation functions in the output layer. The

network weights are initialized by random Gaussian numbers with zero mean and standard deviation of 0.1. The network is trained using back-propagation. The DNN code is obtained from the GPU-based Theano software [17][18].

After the DNN training is converged, the DNN network can be used to estimate the glottal flow from the coded speech data. The same set of speech features which were used during training are extracted from the coded speech signal, and the extracted features are fed to the trained DNN, which finally outputs the glottal flow estimate.

4. Experiments

The experiments conducted in this study evaluate the effectiveness of the proposed DNN-based approach and the existing inverse filtering methods under coded conditions. The accuracy of the glottal flow estimates obtained under coded condition is compared with the corresponding glottal flow estimates obtained under clean condition.

4.1. Speech data

In DNN training, a high-quality multi-speaker speech database was used. The speech database was recorded from 5 male and 6 female speakers. The ages of the speakers ranged from 18 to 48 years. All speakers were equipped with a headset microphone consisting of a unidirectional Sennheiser electret capsule. The microphone signal was passed through a microphone preamplifier and a mixer to iRiver iHP-140 digital audio recorder. The low-frequency phase distortion introduced by the digital recorder was corrected by convolving the recorded signals with the time-reversed version of the impulse response of the digital recorder.

Every utterance in the database is a sustained long vowel. Every speaker uttered 72 utterances, and thus the total number of utterances in the database is 792, comprising about 9.5 minutes of speech data. The utterances were recorded in an anechoic chamber with minimum noise and reverberation. It is worth emphasizing that in the context of glottal source analysis, considerably smaller volumes of speech data are needed in DNN training than, for example, in speech recognition and speech synthesis because of two reasons. The first reason is that as two pitch period long glottal flow waveforms are used as the output of DNN, a large number of glottal flow waveforms can be extracted by using a small volume of speech data (e.g. about 54000 glottal flow waveforms are obtained from 9.5 minutes of speech data). The second reason is that the glottal flow to be estimated by the DNN is an elementary waveform, which is formed at the level of glottis in the absence of vocal tract resonances. Therefore, a DNN can be trained more effectively to predict the waveform of the glottal source in comparison to, for example, the waveform of the speech pressure signal.

Test data consists of speech segments of the vowel [a] extracted from continuous speech. The data was recorded by asking 11 speakers (5 males, 6 females) to read three passages of Finnish text describing past weather conditions. The text was designed in order to have multiple long [a] vowels in contexts where the vowel is surrounded by either an unvoiced fricative or an unvoiced plosive. Hence, the vowel segments recorded were well-suited for GIF. From each speaker, 8 instances of the vowel [a] were extracted. A total of 40 utterances from male speakers and 48 utterances from female speakers were used as the test data. As the present work is the first and preliminary study on glottal source estimation under coded condition, only a small number of test utterances is considered. All speech utterances (both training and test set) were sampled at 8 kHz. Before feature extraction and glottal inverse filtering, the polarity of each of the utterance was checked and corrected if it was inverted to ensure correct estimation of glottal flow waveform.

4.2. Experimental setup

The speech data for the DNN training (both clean and coded) was analyzed using 30-ms frames at 15-ms intervals. Using the QCP method, the frames obtained from the clean speech were inverse filtered. From the clean speech data, a total of 54774 glottal flow waveforms were extracted. Corresponding to every glottal flow waveform, the speech parameters were simultaneously extracted from the coded speech frame. The dataset containing pairs of glottal flow waveforms and speech parameters were split into a training set consisting of 98% of the dataset, and a validation set containing the remaining 2%. The speech parameters considered in this study are fundamental frequency (F0) and line spectral frequencies (LSFs). LSFs represent the speech spectral information and the order of the LSF vector was set to 24.

For comparing the proposed DNN-based inverse filtering, two existing GIF methods, namely, CP [2] and IAIF [3] were used. In the CP inverse filtering method [2], the vocal tract filter is estimated by performing linear prediction (LP) analysis with the covariance criterion over the speech samples that are present in the closed phase of the glottal excitation. In the IAIF inverse filtering method [3], the contribution of the glottal flow is estimated with a low-order LP analysis. IAIF employs an iterative analysis scheme to remove the tilting effect of the glottal source from the speech spectrum. In QCP, CP and IAIF, inverse filtering was performed with the order of LP analysis set to 12.

During evaluation, the speech parameters were computed from the coded speech frames and fed into the DNN to obtain the glottal flow estimate. Using the CP and IAIF methods, the glottal flow estimates of the coded speech were obtained. In addition to this, the glottal flow estimates of the clean speech were obtained using the QCP [6] method which is denoted as the ground truth or reference estimate. The glottal flow estimates of the coded speech obtained from the DNN-based system, CP and IAIF were compared with the reference estimates.

The accuracy of the glottal flow estimation was measured using two voice quality metrics. The first metric, normalized amplitude quotient (NAQ) [19], is a widely used temporal voice quality measure for the relative length of the glottal closing phase. The second metric, H1H2 [20], was computed as the magnitude difference between the first and the second harmonics. The H1H2 was used for quantifying the glottal source in the spectral domain.

4.3. Results

NAQ and H1H2 values are computed from the reference estimates (the glottal flow estimates of the clean speech) and from the glottal sources computed from coded speech using the proposed DNN-based system, CP and IAIF. NAQ and H1H2 values obtained from every glottal flow estimate are averaged for every utterance of the test data. From every utterance, errors are computed between the average NAQ and H1H2 values obtained from the coded signal estimate and the reference estimate. Average error values obtained for male and female speakers are shown in Table 1. From Table 1, it can be observed that the DNN-based system has the lowest errors compared to CP and IAIF. Further, the errors for the female speaker is observed to

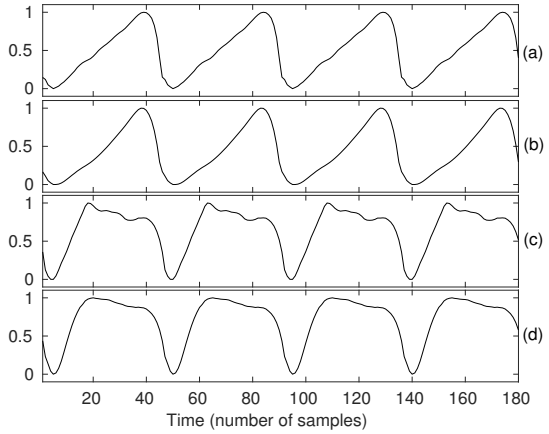

Figure 3: *(a) Glottal flow waveform obtained from clean speech by using QCP. Glottal flow waveforms estimated from coded speech by using (b) the proposed DNN-based system, (c) CP and (d) IAIF.*

be lower compared to the male speaker for all three methods. In the DNN-based system, the difference in NAQ and H1H2 errors for male and female speakers is very small, which indicates that the proposed method performs consistently for male and female speakers. In case of CP and IAIF, the difference in NAQ and H1H2 errors for male and female speakers is relatively high. The better performance of female speaker compared to male speaker for the coded speech is contrary to the common behavior of GIF under clean condition. This is due to band-pass nature of the coded speech, where the lower harmonic components are attenuated. The attenuation of the first harmonic components can significantly degrade the performance of GIF which mainly happens for male speakers, as they have lower pitch ranges compared to female speakers.

Table 1: *Average errors for NAQ and H1H2 for both male and female speakers. Glottal flows were estimated from coded speech by the DNN-based system, CP and IAIF. NAQ error is relative and HIH2 error is in dB.*

GIF methods	Male		Female	
	NAQ	H1H2	NAQ	H1H2
DNN-based system	**0.100**	**1.205**	**0.066**	**1.123**
CP	0.812	5.187	0.270	3.152
IAIF	0.933	5.496	0.437	2.993

Figure 3 illustrates the glottal flow waveforms estimated from clean and coded speech by using different GIF methods. The figure shows the glottal flow waveform estimated from clean speech by using the QCP method (Figure 3(a)), and the glottal flow waveforms estimated from the corresponding coded signal by using the proposed DNN-based system (Figure 3(b)), the CP method (Figure 3(c)) and the IAIF (Figure 3(d)) method. From figure, it can be observed that the glottal flow waveform estimated by the proposed DNN-based system is clearly closer to the glottal flow waveform estimated from clean speech compared to estimates given by CP and IAIF.

5. Conclusion

This paper proposes a DNN-based glottal source estimation method from coded telephone speech. DNN is used to map the speech features extracted from coded speech with the corresponding glottal flow waveforms extracted from clean speech. For extracting the glottal flow waveforms from clean speech, the QCP inverse filtering method is utilized. The proposed DNN-based glottal source estimation method was compared with two existing GIF methods, CP and IAIF. Evaluation results showed that the errors in two objective voice quality metrics, NAQ and H1H2, were considerably lower for the proposed DNN-based method compared to the two conventional GIF methods. The errors in NAQ and H1H2 values were similar for male and female speakers which indicates that the proposed method performs equally well for both male and female speakers.

To the best of our knowledge, the current study is the first investigation in which glottal inverse filtering analysis of the voice source has been conduced from coded telephone speech. Possible future works are as follows. The proposed method should be evaluated with a larger test set by using also wideband speech that has been coded, for example, with the AMR-WB codec [21]. The proposed method can be utilized to extract glottal source-based features from telephone speech to be used in front ends of several speech technology applications (e.g. automatic speech recognition, speaker recognition, emotion recognition, identification of neurodegenerative diseases etc.). To understand whether these source-based features improve the efficiency of the underlying application is the main area of our future studies.

6. References

[1] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*, ser. Prentice-Hall signal processing series. Prentice-Hall, 1978.

[2] D. Wong, J. Markel, and A. G. Jr, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 350–355, 1979.

[3] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, no. 2-3, pp. 109–118, 1992.

[4] T. Drugman, B. Bozkurt, and T. Dutoit, "Complex cepstrum-based decomposition of speech for glottal source estimation," in *Proc. interspeech*, 2009, pp. 116–119.

[5] M. Airas, P. Alku, and M. Vainio, "Laryngeal voice quality changes in expression of prominence in continuous speech," in *Proc. International Workshop on Models and Analysis of Vocal Emissions in Biomedical Applications (MAVEBA)*, 2007, p. 135138.

[6] M. Airaksinen, T. Raitio, B. Story, and P. Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 596–607, 2014.

[7] T. Drugman, B. Bozkurt, and T. Dutoit, "A comparative study of glottal source estimation techniques," *Computer Speech and Language*, vol. 25, no. 1, pp. 20–34, 2012.

[8] J. N. Holmes, "Low-frequency phase distortion of speech recordings," *Journal of the Acoustical Society of America*, vol. 58, no. 3, pp. 747–749, 1975.

[9] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[10] T. Raitio, H. Lu, J. Kane, A. Suni, M. Vainio, S. King, and P. Alku, "Voice source modelling using deep neural networks for statistical parametric speech synthesis," in *Proc. European Signal Processing Conference (EUS1PCO)*, 2014.

[11] T. Raitio, A. Suni, L. Juve1a, M. Vainio, and P. Alku, "Deep neural network based trainable voice source model for synthesis of speech with varying vocal effort," in *Proc. Interspeech*, 2014, pp. 1969–1973.

[12] L. Juvela, B. Bollepalli, M. Airaksinen, and P. Alku, "High-pitched excitation generation for glottal vocoding in statistical parametric speech synthesis using a deep neural network," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5120–5124.

[13] M. Airaksinen, T. Raitio, and P. Alku, "Noise robust estimation of the voice source using a deep neural network," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5137–5141.

[14] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 994–1006, 2012.

[15] 3GPP TS 26.090, "Adaptive multi-rate (AMR) speech codec, transcoding functions," 3rd Generation Partnership Project, Tech. Rep., version 10.1.0, 2011.

[16] K. Järvinen, "Standardisation of the adaptive multi-rate codec," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2000.

[17] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: a cpu and gpu math expression compiler," in *Proc. The Python for Scientific Computing Conference (SciPy)*, 2010.

[18] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. J. Goodfellow, A. Bergeron, N. Bouchard, and Y. Bengio, "Theano: new features and speed improvements," in *Proc. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop*, 2012.

[19] P. Alku, T. Backstrom, and E. Vilkman, "Normalized amplitude quotient for parameterization of the glottal flow," *Journal of Acoustical Society of America*, vol. 112, no. 2, pp. 701–710, 2002.

[20] I. Titze and J. Sundberg, "Vocal intensity in speakers and singers," *Journal of Acoustical Society of America*, vol. 91, no. 5, pp. 2936–2946, 1992.

[21] 3GPP TS 26.171, "Adaptive multi-rate wideband (AMR-WB) speech codec, general description," 3rd Generation Partnership Project, Tech. Rep., version 10.0.0, 2011.