Narendra, N.P.; Airaksinen, Manu; Story, Brad; Alku, Paavo

Estimation of the glottal source from coded telephone speech using deep neural networks

Published in:
Speech Communication

DOI:
10.1016/j.specom.2018.12.002

Published: 01/01/2019

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Estimation of the glottal source from coded telephone speech using deep neural networks

NP Narendra, Manu Airaksinen, Brad Story, Paavo Alku

PII: S0167-6393(18)30144-4
DOI: https://doi.org/10.1016/j.specom.2018.12.002
Reference: SPECOM 2614

To appear in: *Speech Communication*

Received date: 18 April 2018
Revised date: 20 November 2018
Accepted date: 6 December 2018

Please cite this article as: NP Narendra, Manu Airaksinen, Brad Story, Paavo Alku, Estimation of the glottal source from coded telephone speech using deep neural networks, *Speech Communication* (2018), doi: https://doi.org/10.1016/j.specom.2018.12.002

Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 00 (2018) 1–16

Speech Communi-cation

Estimation of the glottal source from coded telephone speech using deep neural networks

N P Narendra [a], Manu Airaksinen [a], Brad Story [b], Paavo Alku [a]

[a] *Department of Signal Processing and Acoustics, Aalto University, Espoo 00076, Finland*

[b] *Speech Acoustics Laboratory, University of Arizona, Tucson, AZ 85718 USA*

Abstract

Estimation of glottal source information can be performed non-invasively from speech by using glottal inverse filtering (GIF) methods. However, the existing GIF methods are sensitive even to slight distortions in speech signals under different realistic scenarios, for example, in coded telephone speech. Therefore, there is a need for robust GIF methods which could accurately estimate glottal flows from coded telephone speech. To address the issue of robust GIF, this paper proposes a new deep neural net-based glottal inverse filtering (DNN-GIF) method for estimation of glottal source from coded telephone speech. The proposed DNN-GIF method utilizes both coded and clean versions of speech signal during training. DNN is used to map the speech features extracted from coded speech with the glottal flows estimated from the corresponding clean speech. The glottal flows are estimated from the clean speech by using quasi closed phase analysis (QCP). To generate coded telephone speech, adaptive multi-rate (AMR) codec is utilized which operates in two transmission bandwidths: narrow band (300 Hz - 3.4 kHz) and wide band (50 Hz - 7 kHz). The glottal source parameters were computed from the proposed and existing GIF methods by using vowels obtained from natural speech data as well as from artificial speech production models. The errors in glottal source parameters indicate that the proposed DNN-GIF method has considerably improved the glottal flow estimation under coded condition for both low- and high-pitched vowels. The proposed DNN-GIF method can be utilized to accurately [1] extract glottal source -based features from coded telephone speech which can be used to improve the performance of speech technology applications such as speaker recognition, emotion recognition and telemonitoring of neurodegerenerative diseases.

Keywords: Glottal source estimation, glottal inverse filtering, deep neural network, coded telephone speech.

1. Introduction

Glottal inverse filtering (GIF) is a method for estimating the glottal source (or voice source) from a recorded microphone speech signal [1][2]. GIF methods typically assume the source-filter model [3] for speech production. In this model, the source refers to the glottal volume velocity signal generated by quasi-periodic fluctuation of the vocal folds [4]. The filter refers to the combined effects of the time-varying vocal tract system and the lip radiation [3]. GIF methods estimate the glottal source signal by removing the effects of the vocal tract formants and lip radiation from the segment of recorded speech [1][5]. The advantages of the GIF method are that it can be performed solely based on speech signal, and it can be computed non-invasively and implemented in a completely automatic manner [6]. The glottal source estimation is a crucial problem in speech processing as it carries important information related to pitch

[1] In this article, the term "accurate/accuracy" is used only when referring to quantitative, objective measures.

Email address: narendra.prabhakera@aalto.fi, manu.airaksinen@aalto.fi,bstory@email.arizona.edu, paavo.alku@aalto.fi (N P Narendra [a], Manu Airaksinen [a], Brad Story [b], Paavo Alku [a])

and phonation type, which in turn can be related to various paralinguistic cues such as emotional state [7], individual speaker characteristics [8], and possible voice pathologies [9][10].

Several GIF methods have been developed in the past decades (see [1] for historical review of GIF methods). Some of the known GIF methods are closed phase covariance analysis (CP) [11], iterative adaptive inverse filtering (IAIF) [5], complex cepstral decomposition (CCD) [12], and quasi-closed phase analysis (QCP) [13]. These methods have been almost exclusively used in ideal conditions where speech recordings are carried out 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 a linear phase response) [14]. Even the best performing GIF methods are sensitive to recording conditions and conducting GIF in non-ideal circumstances can lead to significant distortions in glottal flow estimates [15][16]. The degradation of the glottal source due to non-ideal recording conditions was illustrated by Wong et al. [11] in different circumstances such as ambient noise, low-frequency bias due to breath burst on the microphone, phase distortions due to the recording equipment, and improper A/D conversion.

Even though speech is recorded under ideal conditions, there are different realistic scenarios, which results in distortion of the amplitude and phase properties of speech. Some examples of realistic scenarios under which distortions are introduced in the speech signal are addition of environmental noise, transmission errors, band-pass filtering, and coding (i.e. generation of quantization noise). The accuracy of GIF methods in such realistic scenarios cannot be assumed to be same as in ideal conditions. Even for the most powerful state-of-the-art methods such as CP, IAIF, and QCP, degradation in the accuracy of glottal flow estimation can be observed [6][17] (described in Section III). Furthermore, some of the GIF methods require additional parameter estimations (e.g. extraction of glottal closure instants (GCIs)) which may be affected due to noise [18] and other distortions [19] of the speech signal. Hence, there exists a need for a robust GIF method which can accurately estimate the glottal flow waveform in realistic conditions, for example, when the input signal is a 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. There are plenty of previous works on speech recognition [20] and speaker verification [21][22] from coded telephone speech where acoustical parameters, such as Mel frequency cepstral coefficients (MFCCs), and Linear frequency cepstral coefficients (LFCCs) have been used. But the 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 the glottal source waveform from coded telephone speech has potential applications in speaker recognition, emotion recognition, and in biomedical applications such as detection, classification, and telemonitoring of neurodegenerative diseases [23][24][25].

In the current study, the estimation of the glottal source is studied from coded telephone speech using two bandwidths, narrowband (300 Hz-3.4 kHz) and wideband (50 Hz-7 kHz), that have been standardized in speech transmission [26][27]. From the point of view of GIF, this scenario is severely non-ideal because the input signal is distorted both by quantization noise caused by low bit-rate coding and by bandpass filtering used in the speech transmission standards.

Deep neural networks (DNNs) are widely used powerful tools for finding non-linear relations between input and output features [28]. They have been used in a few recent studies as an alternative to conventional GIF methods to estimate the glottal source waveforms [17][29][30]. In these previous studies, the estimation of the glottal source using DNNs is mainly explored for statistical parametric speech synthesis to improve the quality of speech synthesis. In statistical parametric speech synthesis [29][30][31], the DNN is used to create a mapping between acoustic speech features and time-domain glottal flow waveforms, which are then used as excitation waveforms in the synthesizer. The DNN-based glottal flow excitation generation was shown to provide improvements in the quality of synthesized speech compared to state-of-the-art methods [32][33]. The speech data used in these synthesis-oriented studies is, importantly, mainly of high quality and recorded under ideal conditions.

Even though the DNN-based approach to estimate the glottal flow has been used in a few recent speech synthesis studies as described above, the DNN-based computation of the glottal flow has not been studied before from coded telephone speech. In addition, the accuracy of the glottal flow estimation has not been studied systematically in the previous few investigations on the DNN-based computation of the glottal flow because the emphasis has been on the perceptual speech synthesis quality. Therefore, a new method is proposed in the current study to estimate the glottal source from coded telephone speech. The proposed method utilizes DNNs to estimate the glottal flow waveforms using speech parameters extracted from coded telephone speech. During training, the speech parameters extracted from coded telephone speech are mapped with the corresponding reference glottal flow waveforms obtained from

clean speech[2] by using DNN. The reference glottal flow waveforms are estimated from clean speech by using the QCP method. Unlike the existing GIF methods, which attempt to accurately model the vocal tract filter, and estimate glottal flow by removing the contribution of the vocal tract from the speech signal, the proposed method relies on the non-linear mapping capability of DNNs to estimate the glottal flow. The proposed method requires only low level parameters extracted from coded speech to accurately estimate the glottal flow. Systematic accuracy evaluation of the glottal flow estimation is carried out on natural and synthetic speech data using well known objective measures, such as normalized amplitude quotient, amplitude difference between the first and second harmonic, as well as harmonic richness factor. During evaluation, the proposed method is compared with four existing GIF methods, QCP, CP, IAIF and CCD.

This paper is organized as follows. In Section 2, a description about the basic theory of speech production and a brief review of different glottal source estimation methods are provided. Section 3 provides the illustration of glottal source estimation from coded telephone speech using traditional GIF methods. The proposed DNN-based glottal source estimation method is described in detail in Section 4. In Section 5, experiments conducted with the proposed GIF method are explained and the evaluation results are reported in Section 6. In Section 7, the conclusions of the present study are presented.

2. Glottal source estimation

As the vocal folds oscillate due to interaction of the mechanical tissue properties with respiratory pressure, they modulate the glottis and generate a pulsatile air flow waveform referred to as glottal flow [3]. The glottal flow is filtered by the resonances of the vocal tract cavities and the resulting velocity flow at the lips is converted to the speech pressure waveform, which can be captured in a free field by a microphone. The (linear) source-filter model of speech production can be expressed mathematically in the z-domain as follows:

$$S(z) = G(z)V(z)L(z) = E(z)V(z) \tag{1}$$

where $S(z)$ is the speech signal, $G(z)$ is the glottal flow excitation, $V(z)$ is the vocal tract transfer function, and $L(z)$ is the transfer function of the lip radiation effect. The transfer function of the lip radiation effect can be modeled (for low frequencies) as a fixed differentiator [34]. Given the linearity of the system, the simplified lip radiation effect can be combined to the glottal flow thereby having a model (the right side of Eq. 1) in which the glottal flow derivative serves as the effective driving source (denoted as $E(z)$) to the vocal tract system to generate speech [3]. Existing methods followed for the estimation of glottal flow and their performance under coded condition are explained in remaining part of this section.

2.1. Existing GIF methods

GIF methods determine the glottal flow or its first time-derivative (i.e., the effective driving source) directly from the speech signal. These methods first estimate a parametric model of the vocal tract, and then remove the vocal tract contribution from the speech signal via inverse filtering to obtain the glottal flow. Note that formants are canceled in GIF using a minimum-phase FIR filter whose phase response is non-linear. This non-linear phase response, however, is canceled by the phase response of the vocal tract filter under the assumption that it can be faithfully modeled, both in magnitude and phase, by an all-pole filter. Therefore the phase response of the inverse filter does not distort the estimated glottal flow in a similar manner as external phase distortion that is present in speech signals due to, for example, using an AD-recorder of non-linear phase reponse. Some of the known GIF methods are described below.

2.1.1. Closed Phase Covariance Analysis

In the CP method [11], the glottal source signal is estimated by inverse filtering the speech signal using the vocal tract transfer function that is computed from samples that occur in the closed phase of the glottal cycle. The vocal tract filter is estimated using linear prediction (LP) analysis with the covariance criterion. Computation of the vocal tract from samples that are located in the closed phase is justified because during these samples the effects of the excitation

[2] In this study, clean speech refers to a high-quality speech signal which is recorded under ideal conditions.

are minimal and the signal can be regarded as a freely decaying oscillation caused by vocal tract resonances. Several modifications to the basic CP method have been proposed, which mainly address the difficulty of obtaining accurate closed phase positions [35] and the problem of the reliable vocal tract estimation from very short analysis windows during high-pitched voices [8][36].

In this study, the basic CP method proposed in [11] is adopted by modeling the vocal tract using covariance LP. Since the study includes speech data sampled both with 8 kHz and 16 kHz (see Section 3), the order of LP was p = 10 for sampling frequency (F_s) = 8 kHz and p = 18 for F_s = 16 kHz, which are suitable to model sufficiently all the formants in the given signal bandwidth [3]. The closed phase positions required for the CP method are obtained by estimating GCI locations and pitch information directly from the speech by using the SEDREAMS method [18]. The minimum duration of the closed phase was set to $p + 1$ samples. For high-pitched voices ($F_0 \geq 200$ Hz), two pitch periods were considered for the vocal tract filter estimation [8] in order to obtain more accurate results.

2.1.2. Iterative Adaptive Inverse Filtering

IAIF [5] is a LP-based method that employs an iterative analysis scheme to estimate the glottal source from a speech signal. Using the iterative analysis scheme, the tilting effect of the glottal source is removed from the speech spectrum. In the first iteration, a first-order IIR filter is used to model the contribution of the glottal source, the vocal tract transfer function (LP order = p) is obtained by using the estimated glottal source envelope, and then the speech frame is inverse filtered using this information. During the second iteration, the glottal source signal is estimated with similar steps as in the first iteration except that the LP order of q is used to obtain glottal source envelope. In this study, the IAIF algorithm was used with the orders p = 10, q = 6 for F_s = 8 kHz, and p = 18, q = 10 for F_s = 16 kHz. The LP analysis at each stage was performed by using autocorrelation criterion with the Hann window.

2.1.3. Complex Cepstrum Decomposition

The CCD method [12] is based on the mixed-phase model of speech [37]. In this model, speech is assumed to be composed of minimum-phase (causal) and maximum-phase (anti-causal) components. According to this model, the vocal tract impulse response and the glottal return phase are referred as minimum-phase signals, and the open phase of the glottal flow signal is referred as a maximum-phase signal. In order to compute the glottal source signal, these two components need to be separated. In the CCD method, the complex cepstrum (CC) of a two pitch-period long, GCI centered, and Blackman windowed speech frame is computed, and the positive and negative indices of the CC are separated (liftered). The negative/positive indices correspond to the anti-causal/causal components of the signal. The glottal source (without the return phase) is obtained by applying the inverse complex cepstrum operation to the liftered signal. In this study, the CCD method which is an implementation of [12] is obtained from the GLOAT toolbox [38].

2.1.4. Quasi closed phase analysis

The QCP method is a recently proposed GIF method [13] that is based on the principles of the closed phase analysis [11]. 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 [39], 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 [13] show that the accuracy of QCP is better than that of existing methods such as CP, IAIF, and CCD. However, QCP requires accurate locations of GCI for high performance.

2.2. Performance of existing GIF methods using coded telephone speech

Most of the existing GIF methods are evaluated by using high-quality speech which is recorded under ideal conditions [5][12][13]. Very few works have studied the performance of GIF methods in non-ideal conditions [6][17]. In [6], the robustness of GIF methods is studied by adding white Gaussian noise to the speech signal with various SNR levels. It is observed that the glottal source estimation methods are sensitive to slight distortions in the clean speech. Additional parameters (for example, locations of GCIs and duration of closed phase) required for the computation of glottal source signal are also sensitive to the distortions in the speech signal, which further degrade the performance of GIF methods [13]. On the whole, there is a need for a robust glottal source estimation method which can perform efficiently on speech signals collected in non-ideal/realistic scenarios, for example, coded telephone speech.

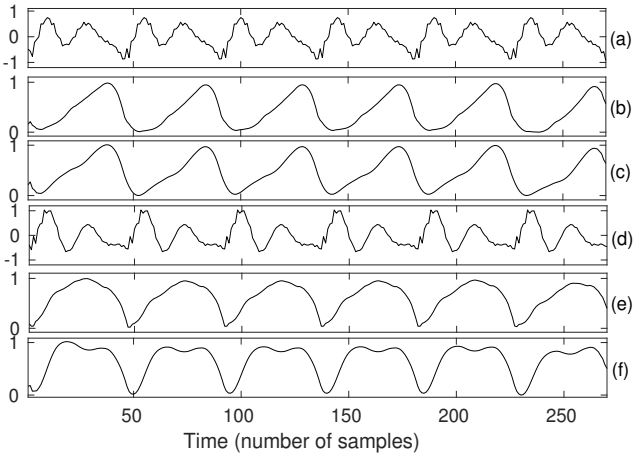

Figure 1. (a) Clean speech, glottal flow waveforms estimated from the clean speech by using (b) QCP, and (c) IAIF, (d) coded speech, and glottal flow waveforms estimated from the coded speech by using (e) QCP, and (f) IAIF.

In this work, an attempt is made to estimate the glottal flow waveform from coded telephone speech. Initially, a study on the effectiveness of existing GIF methods in estimation of glottal flow signal from coded telephone speech is performed. Here, two existing methods, QCP and IAIF, are considered for estimating the glottal flow waveforms from coded telephone speech, as well as from corresponding clean speech. Fig. 1 demonstrates how using simulated telephone transmission (i.e., band-pass filtering, low bit-rate coding, for more details see Section IV) as a source of speech degradation affects the glottal waveforms estimated by existing GIF methods. The figure shows the clean speech (Fig. 1(a)), the glottal flows computed from the clean speech by QCP (Fig. 1(b)) and IAIF (Fig. 1(c)), the coded speech signal (Fig. 1(d)), and the glottal flows computed from the coded signal by QCP (Fig. 1(e)), and IAIF (Fig. 1(f)). The coded speech signal shown in Fig. 1(d) is the coded version of the clean signal shown in Fig. 1(a). From figure, it can be observed that the shape of coded speech signal deviates from the shape of clean speech signal. Coding corrupts the speech signal by introducing different types of amplitude and phase distortions, and quantization noise. As a result of this, the shape of the flow waveform estimated from the coded speech deviates considerably from the corresponding, plausible-looking glottal waveform computed from the clean input. Even though previous studies have observed that the glottal waveforms estimated by traditional GIF methods are affected by, for example, the non-linear phase response of the recording equipment [11], this simple example illustrates clearly that severe distortions that are present in coded telephone speech make the estimation of glottal waveforms erroneous with traditional GIF methods.

3. Proposed method

In order to estimate the glottal source from coded telephone speech, a new data-driven method, called deep neural net-based glottal inverse filtering (DNN-GIF), is proposed. In this approach, a DNN is taken advantage of to map the acoustical features extracted from the input speech signal (coded telephone speech) to the time-domain glottal flow waveform. In the training phase, the DNN-GIF utilizes both the coded input speech signal and the corresponding clean signal recorded in ideal conditions. Prior to DNN training, reference glottal flow waveforms are estimated from the clean speech by using a GIF method. Simultaneously, acoustical speech parameters are extracted from the corresponding frames of the coded speech signal. After the DNN has been trained, it maps the input (a set of acoustical features extracted from the coded telephone speech) to the desired output (the time-domain waveform of the corresponding glottal source).

The block diagram of the proposed DNN-GIF method is shown in Fig. 2. First, a high quality multi-speaker speech database (described in detail in Section 3.1) is considered for the 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 reference glottal flow during the voiced segments of the signal. In the current study, QCP (regarded as one of the most accurate methods according to experiments in [13]), is used as a GIF method to compute the reference glottal flows. In order to compute the AME function in QCP, GCIs are detected using the SEDREAMS algorithm [18]. The glottal source obtained from QCP is pitch-synchronously decomposed into smaller segments. The glottal segments are computed from the derivative of the glottal flow which is decomposed as GCI-centered, two-pitch-period long segments. The glottal segments are then windowed with the square root Hann window and normalized in energy. The two-pitch period long glottal segments are interpolated to a constant length depending on the sampling frequency. Hann windowing of the glottal segments is essential to enable the use of overlap-add for generating the required glottal flow waveform from a fixed length glottal segment estimated by the DNN. The Hann windowing required for the overlap-add is carried out as square root Hann windowing in two parts: first time before training and second time after estimating the glottal segments by the DNN.

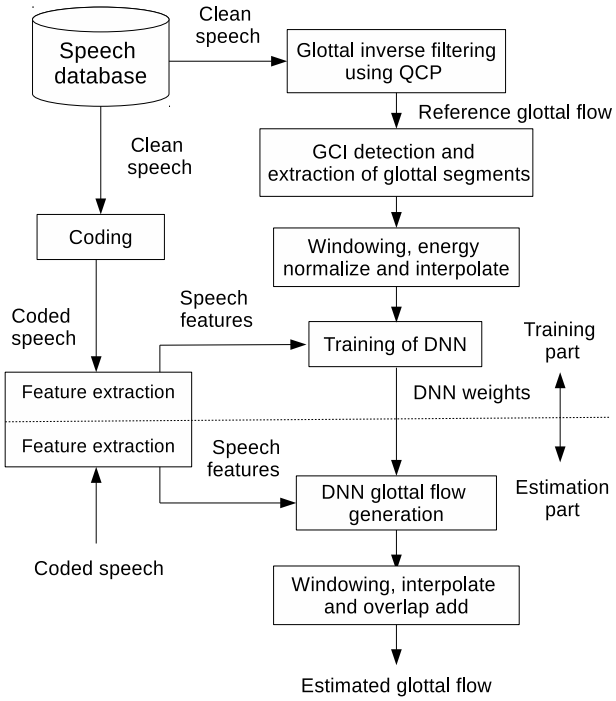

Figure 2. Block diagram of the proposed DNN-GIF method.

In order to simulate telephone speech, utterances of the database are coded using the adaptive multi-rate (AMR) codec [26], which is a widely used speech compression method standardized by the European Telecommunications Standards Institute (ETSI) [27]. The AMR codecs considered in this work employ two transmission bandwidths, narrowband (300 Hz - 3.4 kHz) and wideband (50 Hz - 7 kHz). According to transmission bandwidths, AMR can be categorized either as the AMR-narrowband (NB) codec or the AMR-wideband (WB) codec. The AMR-NB and AMR-WB codecs use sampling frequencies of 8 kHz and 16 kHz, respectively. In the proposed DNN-GIF method, the speech features are extracted separately from both NB- and WB-coded speech. In addition, the reference glottal flow signals are estimated with QCP separately from narrowband clean speech (sampled at 8 kHz) and from wideband clean speech (sampled at 16 kHz). Two separate DNNs are trained (one for narrowband speech, the other one for wideband speech) by mapping the speech features extracted from the AMR-coded speech to the reference glottal flow segments of the corresponding bandwidth. The input features considered can in principle be any acoustic representations of speech. The most commonly used speech features capture the spectral envelope hence involving information about the vocal tract resonances. The speech parameters considered in this study are line spectral frequencies (LSFs)

and the order of the LSF vector was set to 24. For both NB- and WB-coded speech, 24th order LSF was considered as the input speech features.

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 the input speech parameter is a 24-dimension vector, the number of neurons in the input layer is 24. For 8 kHz sampling rate, the total length of the glottal flow segment is fixed to 300 samples, resulting in 300 neurons in the output layer. For sampling rate of 16 kHz, the total length of the glottal flow segment is fixed to 500 samples, resulting in 500 neurons in the output layer. The number of neurons in the three hidden layers are fixed to 250, 150 and 250. As the glottal flow segment predicted at the output will be having continuous values, a linear activation function is used for the output (regression) layer, and sigmoid activation functions are used in the hidden layers. The network weights are initialized by random Gaussian numbers with zero mean and standard deviation of 0.1. The network is trained by back-propagating derivatives of a cost function that measures the discrepancy between the target outputs and the actual outputs. In this work, mean squared error (MSE) is used as the cost function. The DNN code is obtained from the GPU-based Theano software [40][41].

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. The extracted features are then given as input to the trained DNN, which generates the glottal flow waveform. The generated flow waveform is the fixed length derivative of the glottal flow which is then windowed with square root Hann window and aligned to have GCI at the center of a fixed-length frame. The process of windowing and GCI alignment is done to enable proper overlap-add of fixed-length waveforms to generate a complete glottal flow waveform of desired length. In the context of modeling and generation of glottal flow waveform, windowing and GCI alignment is a standard procedure followed in the literature [30][42]. The resulting glottal flow is interpolated to a length equal to twice the length of the desired fundamental period. The desired fundamental period is the pitch period estimated from the coded speech signal, which is used to extract speech features. The pitch is estimated from coded speech by using an existing pitch tracking method proposed by Drugman et al. [43]. Two pitch-period glottal flow is pitch-synchronously overlap-added to generate the final glottal flow estimate.

3.1. Speech data in DNN training and validation

In order to train the DNN, 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 [14]. The microphone signal was passed through a microphone preamplifier and a mixer to iRiver iHP-140 portable digital audio recorder. Unfortunately, this device introduced low-frequency phase distortion to the recorded signal. To correct the undesirable effect of the digital recorder, we used a known compensation technique for phase distortion [44]: the impulse response of the device was first computed using the maximum-length sequences (MLS) measurement [45] and the recorded signal was then convolved with the time-reversed version of the impulse response.

Every utterance in the database is a sustained long vowel. Every speaker uttered sustained phonations of 8 Finnish vowels ([a], [e], [i], [o], [u], [y], [ae], and [oe]) using breathy, modal or pressed phonation types. 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 time-domain waveform of the glottal source in comparison to, for example, the waveform of the speech pressure signal.

The speech data (both clean and coded) was analyzed using 30-ms frames at 5-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 30-ms speech frame. The dataset containing pairs of glottal flow waveforms

and speech parameters were split into a training set consisting of 97.5% of the dataset, and a validation set containing the remaining 2.5%.

4. Experiments

The experiments conducted in this study evaluate the accuracy of the glottal source estimation from coded telephone speech using the proposed DNN-GIF approach and four traditional, known GIF methods. As traditional reference GIF methods, the evaluation included CP, IAIF, CCD, and QCP methods (description of each of these methods is provided in Section 2). The evaluation was performed on synthetic speech data, as well as on natural speech utterances. Synthetic speech data was obtained from two artificial speech production models, the Liljencrants-Fant (L-F) model and a physical model. For evaluating the proposed DNN-GIF method using natural speech data, two approaches were followed. In the first approach, 10-fold cross validation was performed by considering the same set of natural speech utterances, which are used for training the DNN. The second approach employed gender-specific evaluation. Here, the DNN trained using male (female) speech data is evaluated by using only male (female) test data. Male and female test data used in the evaluation were obtained from continuous speech. During the evaluation, a set of glottal source parameters were computed from each of the glottal flow waveforms estimated with the GIF methods under comparison. The errors in glottal source parameters were computed by comparing the glottal flow estimates obtained under coded condition with the corresponding glottal flow estimates obtained under clean condition.

4.1. Glottal source parameters and their error measures

The glottal source parameters considered in this work for the evaluation of the GIF methods are the normalized amplitude quotient (NAQ) [46], the level difference between the first and second harmonic (H1H2) [47], and the harmonic richness factor (HRF) [48]. These parameters describe different aspects of the estimated glottal flow. NAQ is a widely used temporal voice quality measure for finding the relative length of the glottal closing phase from the ratio of the peak flow and the negative peak amplitude of the glottal flow derivative, normalized with respect to the length of the fundamental period. H1H2 and HRF are used for quantifying the glottal source in the spectral domain. H1H2 is the absolute difference between the amplitudes of first and second harmonic frequencies of the glottal flow spectrum. HRF, which measures the decay of the glottal source spectrum, is computed as the ratio between the sum of the power spectral values at harmonics above the fundamental frequency and the power spectral value at the fundamental frequency. These parameters have been previously shown to be highly important in voice source analysis research [6][13]. Significance of NAQ, H1H2 and HRF have been exploited in the analysis of speaker characteristics, voice quality, emotion and singing voice analysis [7][14][48][49][50]. The procedure for computing the three selected parameters is straightforward without any ambiguity using APARAT toolbox [51]. This is in contrast to conventional time-domain parameters of the glottal flow, such as open quotient and closed quotient, whose estimation might be problematic due to, for example, gradual opening of the vocal folds.

NAQ, H1H2, and HRF values were computed from the reference estimates (the glottal flow estimates of the clean speech obtained using QCP) and from the glottal sources computed from coded speech using the proposed DNN-GIF system, QCP, CP, IAIF, and CCD. NAQ, H1H2, and HRF values obtained from every glottal flow estimate were averaged for every utterance of the test data. From every utterance, errors were computed between the average values obtained from the coded signal and the corresponding clean reference. The error for NAQ was obtained (on the linear scale) as the average absolute relative error, and the errors for H1H2, and HRF were obtained (on the dB scale) as the average absolute error:

$$E_{\mathrm{NAQ}} = \mathrm{E}\left[\frac{|\mathrm{NAQ}_{clean} - \mathrm{NAQ}_{coded}|}{\mathrm{NAQ}_{clean}}\right] \tag{2}$$

$$E_{\mathrm{H1H2}} = \mathrm{E}\Big[|\mathrm{H1H2}_{clean} - \mathrm{H1H2}_{coded}|\Big] \tag{3}$$

$$E_{\mathrm{HRF}} = \mathrm{E}\Big[|\mathrm{HRF}_{clean} - \mathrm{HRF}_{coded}|\Big] \tag{4}$$

where E[] denotes expectation.

4.2. Speech data

In this study, two sets of synthetic speech data and natural speech utterances were used as test data in evaluation. The first synthetic test set, the L-F model test set [52], was generated using the source-filter model, which uses the synthetic L-F model as an excitation to an all-pole vocal tract filter. The second synthetic test set, the physical model test set, was generated using a physical model of human speech production [53]. In order to mimic the large dynamics of natural speech, both of the synthetic test sets were generated using a wide range of F_0 values and phonation types. The synthetic test data used in our experiments do not contain jitter or shimmer. Glottal analysis of the synthetic test data with jitter and shimmer is a challenging task that warrants further investigation. However, various previous GIF studies (e.g., [6][13]) have also used similar kind of synthetic test data in evaluations. The natural speech utterances which are used as a test set in gender-specific evaluation are obtained by extracting the segments of vowels [a] from continuous speech.

4.2.1. The L-F model test set

The L-F model [52] is a parametric model for describing the glottal flow derivative waveform with a set of four L-F parameters. The L-F parameters consist of three time-domain values (t_e, t_p, t_a) and one amplitude domain value (E_e), and these parameters can be represented in an alternate dimensionless form (E_e, R_a, R_g and R_k). These parameters can be interpolated within their respective ranges as specified in [54] to get a wide range of possible excitations. A database of synthetic vowels was generated by using an analysis-by-synthesis optimization scheme. F_0 was varied from 80 Hz to 400 Hz with a 20 Hz-increment. The vocal tract was adjusted as in [55], by simulating three different vowels ([e], [o], [ae]) with an all-pole filter with 8 poles (resulting in 4 formants). The L-F test set contained a total of 31, 875 test vowels. The L-F parameters were given four linearly spaced values each, interlaced with the optimization set.

The test samples were divided into two categories based on their pitch: The 'Low' ($F_0 \leq 200$ Hz) and 'High' ($F_0 > 200$ Hz) categories. This was done to study whether the performance the glottal source estimation methods under comparison is dependent on F_0, which is known to be a factor that causes distortion to the estimation of the glottal flow [1][6]. It should be noted that the test sound synthesis is based on a simple linear system.

4.2.2. The physical model test set

In the physical modeling approach, a computational model of the human speech production system is used to generate test vowels instead of using the linear source-filter model as in the L-F model test set. The physical model adopted in the current study used the voice source component that consisted of a kinematic representation of the medial surface of the vocal folds [56][57] for which the surface bulging, adduction, length, and thickness are control parameters. As the vocal fold surfaces are set into vibration, the model generates a glottal area signal that is coupled to the acoustic pressures and air flows in the trachea and vocal tract through aerodynamic and acoustic considerations [58]. The resulting glottal flow was obtained by the interaction of the glottal area with the time-varying pressures present just inferior and superior to the glottis.

The vocal tract shape was specified by an area function representative of the vowels [a], [i], and [ae] of an adult male, adult female, and an approximately five-year-old child. Details about the parameter values used for the physical model are provided in [39]. Speech utterances of vowels were generated with eight values of F_0, ranging from 100 Hz to 450 Hz in 50 Hz-increments for the male, the female, and the child. Even though the full range of these F_0 is unlikely to be produced by either the male, female, or the child, conducting the experiment with the entire range was desirable for ease in comparison. Vowel duration was 0.4 seconds and F_0 was maintained constant during the utterance. Same as in L-F model test set, the physical model test set is also divided into 'Low' and 'High' categories based on their pitch.

4.2.3. Natural speech data set

In order to perform gender-specific evaluation (described in Section 5.3.2), the test data set was developed by extracting the segments of vowels [a] from continuous speech. There is no relation between this data set and the speech data used in training (described in Sec. 3.1). Initially, the speech 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 the continuous speech of every speaker, 8 instances of the vowel [a] were extracted. The average durations of vowels used in testing are 104.7 ms for male speakers and 130.4 ms for female speakers. A total of 40 utterances from male speakers, and 48 utterances from female speakers were used as the test data set.

Table 1. Errors for NAQ, H1H2 and HRF obtained for the L-F model test set. Best results printed in bold font.

	E_{NAQ}			E_{H1H2} (dB)			E_{HRF} (dB)		
L-F model (NB-coded)	All	Low	High	All	Low	High	All	Low	High
DNN-GIF	**0.104**	**0.094**	**0.113**	**0.898**	**0.883**	**0.913**	**0.947**	**0.914**	**0.980**
QCP	0.371	0.321	0.421	4.128	4.649	3.607	3.751	4.087	3.416
CCD	0.540	0.702	0.379	5.635	6.403	4.868	4.251	3.654	4.848
CP	0.525	0.470	0.580	5.023	5.212	4.834	4.917	4.957	4.877
IAIF	0.617	0.494	0.740	7.706	5.340	10.072	7.021	4.686	9.356
L-F model (WB-coded)	All	Low	High	All	Low	High	All	Low	High
DNN-GIF	**0.080**	**0.083**	**0.076**	**1.002**	**1.141**	**0.863**	**0.743**	**0.950**	**0.536**
QCP	0.216	0.158	0.275	3.658	4.911	2.406	3.496	4.609	2.384
CCD	0.561	0.687	0.435	5.153	5.774	4.532	5.097	5.446	4.748
CP	0.563	0.200	0.926	5.843	5.410	6.277	5.575	5.052	6.097
IAIF	0.725	0.263	1.186	7.720	5.748	9.692	7.006	5.010	9.003

Table 2. Errors for NAQ, H1H2 and HRF obtained for the physical model test set. Best results printed in bold font.

	E_{NAQ}			E_{H1H2} (dB)			E_{HRF} (dB)		
Physical model (NB-coded)	All	Low	High	All	Low	High	All	Low	High
DNN-GIF	**0.102**	**0.090**	**0.113**	**1.883**	**1.332**	**2.434**	**1.183**	**1.428**	**0.938**
QCP	0.460	0.238	0.683	4.556	3.452	5.661	4.161	3.310	5.012
CCD	0.591	0.704	0.478	5.070	4.590	5.550	3.213	2.543	3.883
CP	0.454	0.287	0.620	4.154	3.075	5.233	3.912	3.009	4.816
IAIF	0.591	0.369	0.813	6.076	3.991	8.161	5.565	3.707	7.424
Physical model (WB-coded)	All	Low	High	All	Low	High	All	Low	High
DNN-GIF	**0.103**	**0.085**	**0.121**	**1.629**	**1.902**	**1.356**	**1.101**	**1.246**	**0.956**
QCP	0.707	0.318	1.096	4.848	4.002	5.695	4.387	3.491	5.283
CCD	0.628	0.636	0.619	4.449	4.229	4.670	2.588	2.600	2.576
CP	0.775	0.248	1.302	4.656	3.550	5.761	4.518	3.242	5.795
IAIF	1.004	0.382	1.625	6.046	3.806	8.285	5.429	3.267	7.592

5. Results

5.1. Results on the L-F model test set

The average error values obtained for the L-F model test set are given in Table 1. It can be observed that the proposed DNN-GIF method has the lowest errors compared to existing methods in all categories. DNN-GIF performs better at 'low' frequencies than at 'high' frequencies for NB-coded speech, and vice versa for WB-coded speech. QCP has the second lowest errors after DNN-GIF at 'low' and 'high' frequencies, for both NB- and WB-coded speech (except for NAQ of NB-coded speech at 'high' frequency, and except for HRF of NB-coded speech at 'low' frequency). The CP method has the third lowest errors after DNN-GIF and QCP at 'low' frequencies, for both NB- and WB-coded speech (except for HRF of NB- and WB-coded speech at 'low' frequency). The CCD method shows

the third lowest errors after the DNN-GIF method at 'high' frequencies, for both NB- and WB-coded speech (except for H1H2 and NAQ of NB-coded speech at 'high' frequency). At 'low' frequencies, the CCD method shows highest errors among all the compared methods for both NB- and WB-coded speech (except for HRF of NB-coded speech at 'low' frequency). At 'high' frequencies, IAIF shows highest errors compared to all other methods for both NB- and WB-coded speech.

5.2. Results on the physical model test set

The average error values for the physical model test set are presented in Table 2. The proposed DNN-GIF method has the lowest errors compared to the traditional GIF algorithms in all categories. CP has the second lowest errors after DNN-GIF at 'low' frequencies, for both NB- and WB-coded speech (except for HRF of NB- and WB-coded speech and for NAQ of NB-coded speech at 'low' frequency). CCD has the second lowest errors after the proposed method at 'high' frequencies, for both NB- and WB-coded speech (except for H1H2 of NB-coded speech). At 'low' frequencies, the CCD method has highest errors compared to all other methods for both NB- and WB-coded speech (except for HRF of NB- and WB-coded speech at 'low' frequencies). At 'high' frequencies, IAIF has highest errors compared to all other methods for both NB- and WB-coded speech.

On the whole, CCD has high errors at low frequency and IAIF has high errors at high frequency for both the L-F test set and the physical model test set. The poor performance of the CCD method at low frequency might be caused by the fact that the obtained version of the CCD method estimates only the opening and closing phases of a glottal period until the GCI, discarding the return phase completely. The return phase is relatively constant at high frequency, but at low frequency the return phase tends to increase linearly with the pitch period [59], leading to deterioration of glottal flow estimates which results in high objective error measures. In IAIF, at high frequency, the LP-based spectral envelope tends to become biased towards the harmonic peaks, leading to erroneous estimation of the vocal tract filter which in turn results in erroneous glottal flow estimates and high objective errors. QCP performs better with the L-F model test set compared to the physical model test set. The poor performance of QCP with the physical model test set might be caused due to erroneous detection of GCIs from coded speech.

5.3. Results on natural speech data

In order to evaluate the proposed DNN-GIF method using natural speech data, two approaches were followed. The first approach employed 10-fold cross validation using vowels produced by both male and female talkers. In the second approach, gender-specific evaluation was performed. In this approach, DNNs trained separately using the speech data of male and female speakers are evaluated individually using the test data of male and female speakers.

5.3.1. 10-fold cross validation

In 10-fold cross validation, the same set of natural speech utterances which were used for training DNNs was utilized (details about speech data are provided in Section 3.1). In 10-fold cross validation, speech utterances of all subjects are randomly shuffled. Here, the speech utterances are not identified by their subject identity. The speech utterances are divided into 10 smaller sets. Among 10 sets, a single set is used as the validation data for testing, and the remaining 9 sets are used for training the DNN. In the 9 sets used for training, speech utterances are coded separately using the NB and WB codecs. From every coded speech utterance, features are extracted and from the corresponding clean speech utterance, glottal flow waveforms are estimated. DNNs are trained to map the features extracted from coded speech with the glottal flow waveforms estimated from clean speech. Separate DNNs are developed using NB- and WB-coded speech. The validation data set used for testing is also coded separately with the NB and WB codecs. Utilizing the developed DNN-GIF, average error measures are computed using the validation data set. The cross-validation process is then repeated 10 times (the folds), with each of the 10 sets used exactly once as the validation data. The objective error measures obtained from all the 10 folds are averaged to produce a single set of error measures. The average error measures are computed for both NB- and WB-coded speech. In addition to computation of objective measures for DNN-GIF, the objective measures are obtained from four existing GIF methods. Since all the existing GIF are signal processing-based methods, there is no need for training stage, and all 10 sets of natural speech utterances are directly used to compute objective measures. The average error measures obtained from 10-fold cross validation are shown in Table 3.

Error measures computed from DNN-GIF are lower compared to other existing GIF methods for both NB- and WB-coded speech. Error measures obtained from DNN-GIF are lower compared to those of the L-F model test set,

physical model test set and gender-specific evaluation (described in Section 5.3.2). The main reason for improved performance is that the same kind of natural speech utterances (i.e. sustained vowels) are used in both training and testing. From Table 3, it can also be noticed that the error measures of DNN-GIF are slightly lower for NB-coded speech compared to WB-coded speech (except HRF for NB- and WB-coded speech).

Table 3. Average errors for NAQ, H1H2 and HRF obtained from 10-fold cross validation.

NB-coded	E_{NAQ}	E_{H1H2} (dB)	E_{HRF} (dB)
DNN-GIF	**0.082**	**0.720**	**0.725**
QCP	0.9422	3.8349	3.6757
CCD	0.5520	6.6378	5.9311
CP	1.0439	4.3286	4.6247
IAIF	1.3539	4.6324	4.5427
WB-coded	E_{NAQ}	E_{H1H2} (dB)	E_{HRF} (dB)
DNN-GIF	**0.093**	**0.760**	**0.703**
QCP	2.4302	3.7635	3.7082
CCD	3.2258	7.8831	5.3302
CP	1.9352	3.9740	3.8956
IAIF	2.4488	4.7082	4.3297

5.3.2. Gender-specific evaluation

The gender-specific evaluation is performed to understand the potential influence of the gender on training and testing of DNN-GIF. For gender-specific evaluation, the natural speech data used in DNN training is divided into two sets containing male and female utterances (details about speech data are provided in Section 3.1). Among natural speech data containing 792 utterances, 360 utterances are from male speakers and 432 utterances are from female speakers. DNNs are trained separately using utterances from male and female speakers for both NB- and WB-coded speech. For evaluation, test data containing natural speech utterances of male and female speakers are used (description about the test data is provided in Section 4.2.3). The DNN trained with speech utterances of male (female) voice is evaluated using the test utterances obtained from male (female) voice. Table 4 shows the average errors obtained from the gender-specific evaluation. Results of Table 4 show that the error measures are slightly lower if the same gender speech data is used during training and testing, and the error measures are slightly higher if there is a gender mismatch between training and testing for both NB- and WB-coded speech. The error measures are slightly higher for WB-coded speech compared NB-coded speech. From these results, it can be concluded that slight gender-specific influences are present in the trained DNN. Even though natural speech utterances are used in training and testing, slightly higher error measures can be observed in the gender-specific evaluation compared to the 10-fold cross validation. The main reason for this is that the male and female speech data used for testing in the gender-specific evaluation is different from speech data used for training. The vowel speech samples that are used for testing in the gender-specific evaluation are extracted from continuous speech utterances. In the 10-fold cross validation, sustained vowels are used for both training and testing. Also, the 10-fold cross validation results are obtained by considering both male and female speech data together in testing.

Fig. 3 illustrates the glottal flow waveforms estimated from clean and coded speech by using different GIF methods. Clean speech is obtained from natural speech data and the AMR-NB codec is used to obtain the corresponding coded telephone speech version. The figure shows the glottal flow waveform estimated from clean speech by using the QCP method (Fig. 3 (a)), and the glottal flow waveforms estimated from the corresponding coded speech signal by using the proposed DNN-GIF method (Fig. 3(b)), the QCP method (Fig. 3(c)), the CCD method (Fig. 3(d)), the CP method (Fig. 3(e)), and the IAIF method (Fig. 3(f)). From the figure, it can be observed that the glottal flow waveform estimated by the proposed DNN-GIF method is clearly closer to the glottal flow waveform estimated from clean speech compared to estimates given by the four traditional GIF methods.

Table 4. Average errors for NAQ, H1H2 and HRF obtained for gender-specific evaluation.

Natural speech data	E_{NAQ}	E_{H1H2} (dB)	E_{HRF} (dB)
DNN-GIF (NB-coded) (train: male, test: male)	0.113	0.792	0.708
DNN-GIF (NB-coded) (train: male, test: female)	0.177	0.738	0.737
DNN-GIF (NB-coded) (train: female, test: female)	0.084	0.968	0.755
DNN-GIF (NB-coded) (train: female, test: male)	0.111	1.019	1.080
DNN-GIF (WB-coded) (train: male, test: male)	0.129	0.816	0.803
DNN-GIF (WB-coded) (train: male, test: female)	0.174	0.782	0.693
DNN-GIF (WB-coded) (train:female, test: female)	0.110	1.009	0.967
DNN-GIF (WB-coded) (train:female, test: male)	0.155	1.375	1.099

6. Conclusion

A new method, deep neural net-based glottal inverse filtering (DNN-GIF), is proposed to estimate the glottal source signal from coded telephone speech. The proposed method utilizes DNN to establish a mapping between acoustic speech features extracted from coded telephone speech and the glottal flow waveform obtained from clean speech. For extracting the glottal flow waveform from clean speech to be used as the reference signal in the DNN training, the QCP inverse filtering method is utilized. Using the trained DNN, the glottal flow waveforms are estimated from telephone speech that has been coded with two standardized speech codecs, AMR-NB and AMR-WB. The proposed method is compared with four existing GIF methods by computing objective measures from the glottal flow estimates of sustained vowels obtained from synthetic speech data which is generated by using different models of speech production. The results indicate that the errors in objective measures were considerably lower for the proposed method compared to other four GIF methods for both low-pitched and high-pitched vowels.

In the current study, detailed investigation on the glottal inverse filtering analysis of the voice source has been conducted from coded telephone speech. The proposed method requires precise GCI locations and glottal flow during training. As both GCIs and glottal flow are estimated from clean speech, the trained DNN is fairly accurate and results in better performance even though the input coded speech used in training and testing is corrupted due to amplitude and phase distortions. As a result, it has been shown that using low level parameters extracted from coded speech, and using an effective DNN-based mapping between coded telephone speech and clean speech, a reliable estimation of the distortion-free glottal flow is achieved for both low-pitched and high-pitched voices. Even though the proposed method is evaluated in the current study only on sustained vowels, the proposed method can be extended for applications involving continuous speech. Using appropriate phonetic segmentation and voicing decision techniques [60], the proposed method can be readily applied on voiced sections of continuous speech.

The basic principle of the proposed method, the use of DNNs in the estimation of glottal source waveforms, is similar to two previous investigations [17][29]. There are, however, clear differences between the current study and these two previous investigations. In [17], a noise mismatch scenario is studied in which a DNN trained using clean speech is tested with speech corrupted by additive white Gaussian noise. In the current study, however, DNN-GIF uses realistic coded telephone speech in both training and testing phases, an issue that is essential in applications such as telemonitoring of voice pathologies where speech parameterization needs to be computed remotely after the transmission of speech. In addition, [17] includes only two reference GIF methods (QCP and IAIF) while in the current study the proposed DNN-GIF method is compared with four reference methods (QCP, IAIF, CP, and CCD).

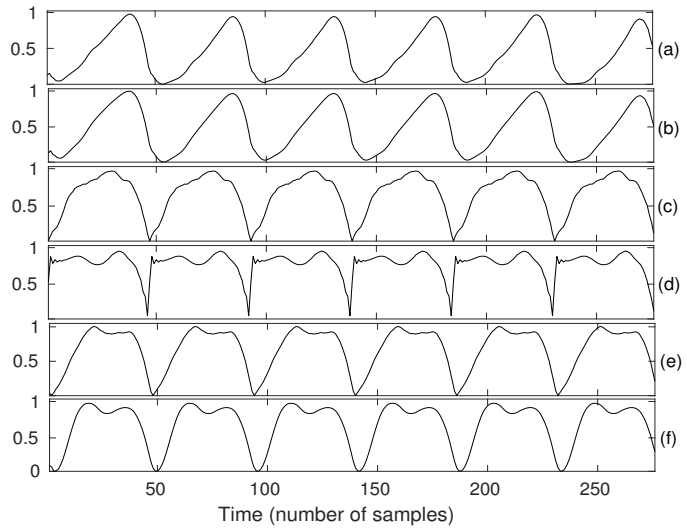

Figure 3. (a) Glottal flow waveform obtained from clean speech by using QCP. Glottal flow waveforms estimated from coded speech by using (b) the proposed DNN-GIF method, (c) QCP, (d) CCD, (e) CP and (f) IAIF.

In [29], the DNN-based glottal flow estimation is applied in statistical parametric speech synthesis and the framework allows only parameters generated by the synthesizer's acoustic models as input to DNN and, hence, the developed DNN is confined with specific type of input features. Finally, the DNNs in [17] and [29] are trained using either speech data of a single talker or speech data of multiple speakers of the same gender (mostly male) while in the current study, the DNN is trained using both male and female speakers.

Possible future works are as follows. Apart from the AMR-NB and AMR-WB codecs, the proposed method can be evaluated using recent codecs, for example, Enhanced Voice Services (EVS) codec [61]. 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.

7. Acknowledgment

This research has been funded by the Academy of Finland (project no. 284671 and 312490).

References

[1] P. Alku, Glottal inverse filtering analysis of human voice production a review of estimation and parameterization methods of the glottal excitation and their applications, Sadhana (2011) 623–650.
[2] T. Drugman, P. Alku, A. Alwan, B. Yegnanarayana, Glottal source processing: From analysis to applications, Computer Speech and Language 28 (5) (2014) 1117–1138.
[3] L. Rabiner, R. Schafer, Digital Processing of Speech Signals, Prentice-Hall signal processing series, Prentice-Hall, 1978.
[4] G. Fant, Glottal source and excitation analysis, STL-QPSR 20 (1) (1979) 85–107.
[5] P. Alku, Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering, Speech Communication 11 (2-3) (1992) 109–118.
[6] T. Drugman, B. Bozkurt, T. Dutoit, A comparative study of glottal source estimation techniques, Computer Speech and Language 25 (1) (2012) 20–34.
[7] M. Airas, P. Alku, Emotions in vowel segments of continuous speech: analysis of the glottal flow using the normalised amplitude quotient, Phonetica 63 (1) (2006) 26–46.
[8] M. Plumpe, T. Quatieri, D. Reynolds, Modeling of the glottal flow derivative waveform with application to speaker identification, IEEE Transactions on Speech and Audio Processing 7 (5) (1999) 569–586.
[9] Y.-R. Chien, M. Borský, J. Gudnason, Objective severity assessment from disordered voice using estimated glottal airflow, in: Proc. interspeech, 2017, pp. 304–308.

[10] T. Drugman, T. Dubuisson, T. Dutoit, On the mutual information between source and filter contributions for voice pathology detection, in: Proc. Interspeech, 2009, pp. 1463–1466.
[11] D. Wong, J. Markel, A. G. Jr, Least squares glottal inverse filtering from the acoustic speech waveform, IEEE Transactions on Audio, Speech, and Language Processing 27 (4) (1979) 350–355.
[12] T. Drugman, B. Bozkurt, T. Dutoit, Complex cepstrum-based decomposition of speech for glottal source estimation, in: Proc. interspeech, 2009, pp. 116–119.
[13] M. Airaksinen, T. Raitio, B. Story, P. Alku, Quasi closed phase glottal inverse filtering analysis with weighted linear prediction, IEEE/ACM Transactions on Audio, Speech, and Language Processing 22 (3) (2014) 596–607.
[14] M. Airas, P. Alku, 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, pp. 135–138.
[15] A. O. Cinnéide, D. Dorran, M. Gainza, E. Coyle, Exploiting glottal formant parameters for glottal inverse filtering and parameterization, in: Proc. Interspeech, 2010.
[16] D. Childers, J. Naik, J. Larar, A. Krishnamurthy, G. R. Moore, Vocal Fold Physiology, Biomechanics, Acoustics and Phonatory Control, The Denver Center for The Performing Arts, Denver, 1983, Ch. Electroglottography, speech, and ultra-high speed cinematography, pp. 202–220.
[17] M. Airaksinen, T. Raitio, 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.
[18] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, T. Dutoit, Detection of glottal closure instants from speech signals: A quantitative review, IEEE Transactions on Audio, Speech, and Language Processing 20 (3) (2012) 994–1006.
[19] J. N. Holmes, Low-frequency phase distortion of speech recordings, Journal of the Acoustical Society of America 58 (3) (1975) 747–749.
[20] I. Medennikov, A. Prudnikov, A. Zatvornitskiy, Improving English conversational telephone speech recognition, in: proc. Interspeech, 2016.
[21] H. Lei, E. Lopez-Gonzalo, Mel, linear, and antimel frequency cepstral coefficients in broad phonetic regions for telephone speaker recognition, in: Proc. Interspeech, 2009, pp. 2323–2326.
[22] L. F. Gallardo, M. Wagner, S. Möller, I-vector speaker verification based on phonetic information under transmission channel effects, in: Interspeech, 2014, pp. 696–700.
[23] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, L. O. Ramig, Suitability of dysphonia measurements for telemonitoring of parkinsons disease, IEEE Transactions on Biomedical Engineering 56 (4) (2009) 1015–1022.
[24] P. Klumpp, T. Janu, T. Arias-Vergara, J. C. V. Correa, J. R. Orozco-Arroyave, E. Nth, Apkinson - a mobile monitoring solution for parkinsons disease, in: Proc. Interspeech, 2017, pp. 1839–1843.
[25] E. Vaiciukynas, A. Verikas, A. Gelzinis, M. Bacauskiene, Detecting parkinsons disease from sustained phonation and speech signals, Plus One 12 (10) (2017) 1–16.
[26] 3GPP TS 26.090, Adaptive multi-rate (AMR) speech codec, transcoding functions, Tech. rep., 3rd Generation Partnership Project (version 10.1.0, 2011).
[27] K. Järvinen, Standardisation of the adaptive multi-rate codec, in: Proc. European Signal Processing Conference (EUSIPCO), 2000.
[28] Z.-H. Ling, S. yin Kang, H. Zen, A. Senior, M. Schuster, X.-J. Qian, H. Meng, L. Deng, Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends, IEEE Signal Processing Magazine 32 (3) (2015) 35–52.
[29] T. Raitio, H. Lu, J. Kane, A. Suni, M. Vainio, S. King, P. Alku, Voice source modelling using deep neural networks for statistical parametric speech synthesis, in: Proc. European Signal Processing Conference (EUS1PCO), 2014.
[30] L. Juvela, B. Bollepalli, M. Airaksinen, 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.
[31] T. Raitio, A. Suni, L. Juve1a, M. Vainio, 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.
[32] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, P. Alku, HMM-based speech synthesis utilizing glottal inverse filtering, IEEE Transactions on Audio, Speech, and Language Processing 19 (1) (2011) 153–165.
[33] T. Raitio, A. Suni, M. Vainio, P. Alku, Comparing glottal flow-excited statistical parametric speech synthesis methods, in: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 7830–7834.
[34] J. L. Flanagan, Speech Analysis, Synthesis and Perception, 2nd Edition, New York: Springer-Verlag, 1972.
[35] D. E. Veeneman, S. BeMent, Automatic glottal inverse filtering from speech and electroglottographic signals, IEEE Transactions on Acoustics, Speech, and Signal Processing 33 (2) (1985) 369–377.
[36] P. Alku, C. Magi, T. Bäckström, B. Story, Glottal inverse filtering with the closed-phase covariance analysis utilizing mathematical constraints in modelling of the vocal tract, Journal of the Acoustical Society of America 125 (5) (2009) 3289–3305.
[37] B. Bozkurt, T. Dutoit, Mixed-phase speech modeling and formant estimation using differential phase spectrums, in: Proc. ISCA Workshop on Voice Quality (VOQUAL), 2003, pp. 21–24.
[38] T. Drugman, GLOttal Analysis Toolbox (GLOAT) (2012).
URL `http://tcts.fpms.ac.be/drugman/Toolbox/`
[39] P. Alku, J. Pohjalainen, M. Vainio, A.-M. Laukkanen, B. H. Story, Formant frequency estimation of high-pitched vowels using weighted linear prediction, Journal of the Acoustical Society of America 134 (2) (2013) 1295–1313.
[40] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, Y. Bengio, Theano: a CPU and GPU math expression compiler, in: Proc. The Python for Scientific Computing Conference (SciPy), 2010.
[41] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. J. Goodfellow, A. Bergeron, N. Bouchard, Y. Bengio, Theano: new features and speed improvements, in: Proc. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.
[42] N. P. Narendra, K. S. Rao, Time-domain deterministic plus noise model based hybrid source modeling for statistical parametric speech synthesis, Speech Communication 77 (2016) 65–83.
[43] T. Drugman, A. Alwan, Joint robust voicing detection and pitch estimation based on residual harmonics, in: Proc. Interspeech, 2011, pp. 1973–1976.

[44] M. Airas, P. Alku, Comparison of multiple voice source parameters in different phonation types, in: Proc. Interspeech, 2007, pp. 1410–1413.
[45] D. D. Rife, J. Vanderkooy, Transfer-function measurement with maximum-length sequences, Journal of the Audio Engineering Society 37 (1989) 419–444.
[46] P. Alku, T. Backstrom, E. Vilkman, Normalized amplitude quotient for parameterization of the glottal flow, Journal of the Acoustical Society of America 112 (2) (2002) 701–710.
[47] G. Fant, The LF-model revisited. transformations and frequency domain analysis, STL-QPSR 36 (2-3) (1995) 119–156.
[48] D. G. Childers, C. K. Lee, Vocal quality factors: Analysis, synthesis, and perception, Journal of the Acoustical Society of America 90 (5) (1991) 2394–2410.
[49] N. Campbell, P. Mokhtari, Voice quality: the 4th prosodic dimension, in: Proc. International Congress of Phonetic Sciences, 2003, pp. 2417–2420.
[50] I. Arroabarren, A. Carlosena, Effect of the glottal source and the vocal tract on the partials amplitude of vibrato in male voices, Journal of Acoustical Society of America 119 (4) (2006) 2483–2497.
[51] M. Airas, H. Pulakka, T. Bäckström, P. Alku, A toolkit for voice inverse filtering and parametrisation, in: Proc. Interspeech, 2005, pp. 2145–2148.
[52] G. Fant, J. Liljencrants, Q. Lin, A four-parameter model of glottal flow, STL-QPSR 26 (4) (1985) 1–13.
[53] B. Story, Physiologically-based speech simulation using an enhanced wave-reflection model of the vocal tract, Ph.D. thesis, University of Iowa, Iowa City, IA, USA (1995).
[54] C. Gobl, The voice source in speech communication - production and perception experiments involving inverse filtering and synthesis, Ph.D. thesis, KTH, Speech Transmission and Music Acoustics, Stockholm, Sweden (2003).
[55] B. Gold, L. Rabiner, Analysis of digital and analog formant synthesizers, IEEE Transactions on Audio Electroacoustics 16 (1) (1968) 81–94.
[56] I. R. Titze, Parameterization of the glottal area, glottal flow, and vocal fold contact area, Journal of the Acoustical Society of America 75 (2) (1984) 570–580.
[57] I. R. Titze, The Myoelastic Aerodynamic Theory of Phonation, National Center for Voice and Speech, 2006.
[58] I. R. Titze, Regulating glottal airflow in phonation: Application of the maximum power transfer theorem to a low dimensional phonation model, Journal of the Acoustical Society of America 111 (1) (2002) 367–376.
[59] M. Tooher, J. G. McKenna, Variation of glottal LF parameters across F0, vowels, and phonetic environment, in: Voice Quality: Functions, Analysis and Synthesis (VOQUAL), 2003, pp. 41–46.
[60] D. Talkin, REAPER: Robust Epoch And Pitch EstimatoR (2015).
URL `https://github.com/google/REAPER`
[61] 3GPP TS 26.445, EVS Codec Detailed Algorithmic Description; 3GPP Technical Specification (Release 12), 2014.