Speaker-dependent Mapping of Source and System Features for Enhancement of Throat Microphone Speech

Anand Joseph M., Sri Harish Reddy M. and B. Yegnanarayana

[1]International Institute of Information Technology Hyderabad, India

anandjm@research.iiit.ac.in, sriharsham@research.iiit.ac.in, yegna@iiit.ac.in

Abstract

A throat microphone (TM) produces speech which is perceptually poorer than that produced by a close speaking microphone (CSM) speech. Many attempts at improving the quality of TM speech have been made by mapping the features corresponding to the vocal tract system. These techniques are limited by the methods used to generate the excitation signal. In this paper a method to map the source (excitation) using multilayer feed-forward neural networks is proposed for voiced segments. This method anchors the analysis windows at the regions around the instants of glottal closure, so that the non-linear characteristics in these region of TM and CSM microphone is emphasized in the mapping process. The features obtained from these regions for both TM and CSM speech are used to train a MLFFNN to capture the non-linear relation between them. An improved technique for mapping the system features is also proposed. Speech synthesized using the proposed techniques was evaluated through subjective tests and was found to be significantly better than TM speech.

Index Terms: throat microphone, source mapping, system mapping, neural network, GCI

1. Introduction

Most methods for enhancement of throat microphone (TM) speech focus on mapping the spectral features corresponding to the vocal tract system of TM and close speaking microphone (CSM) speech [1] [2]. While these methods do improve the perceptual quality of TM speech, they are generally limited by the manner in which the excitation signals are derived from the speech signal. When residual templates derived from CSM speech [3] are used to generate an excitation signal, the quality was found to be poor.

Speech is generally considered to be the output of a vocal tract system excited by a time-varying excitation. The vocal tract system also varies with time, albeit slowly. This allows for the assumption that the vocal tract system is stationary over short durations, and so one can estimate the parameters of the system through techniques such as linear prediction analysis. Because the vocal tract system is slowly varying, it is also simpler to develop techniques to enhance TM speech by mapping the spectral features of TM and CSM speech. This mapping is performed by first computing the linear prediction (LP) coefficients using analysis frames of duration 20-30 ms, from simultaneously recorded TM and CSM speech. This is followed by training a multilayer feed-forward neural network (MLFFNN) with the input/output pairs of features corresponding to TM and CSM speech, respectively.

This paradigm for mapping the system has performed reasonably well, especially in voice conversion applications. However little progress has been made on mapping source-related features. Mapping of features corresponding to the excitation source is a more difficult problem than that of mapping the system features. This is primarily because the source (when compared to the vocal tract system) has very different characteristics. The source features are highly non-linear in nature and may vary rapidly both within a pitch cycle and also across pitch cycles. Any method for enhancement of the TM speech should take into consideration the characteristics of the source when extracting source features. When the vocal tract system is represented as an all-pole system characterized by a set of linear prediction coefficients, the error signal, i.e., the LP residual, primarily represents the characteristics of the source, especially around the instants of glottal closure. It also captures features such as the strength of excitation and the pitch period. In addition to these source characteristics, the LP residual also captures any deviations in the assumptions of linearity and time-invariance of the vocal tract system. Hence in addition to mapping the vocal tract system features, the source features should also be mapped.

This paper proposes a technique for mapping the residual of TM speech, so that the mapped residual when used to excite the mapped system features results in a speech signal with improved perceptual quality. During speech production, the significant excitation of the vocal tract system is at the glottal closure instant (GCI). The region around the GCI is a characteristic of the source. In this paper, the source features for mapping of voiced segments of speech are obtained from regions around the GCIs for both TM and CSM speech. A multilayer feed-forward neural network (MLFFNN) is used to capture the non-linear relation between the source features derived from TM and CSM speech, by training the network on a set of input/output source features corresponding to TM and CSM speech. Since the GCIs are used as anchor points, this method for source mapping is restricted to only voiced segments.

This paper also improves on the technique proposed earlier for mapping the system features [2], so that the temporal variations in the vocal tract system are also incorporated during mapping. The mapped source and system features so obtained are used to enhance the TM speech.

This paper is organized as follows. Section 2 describes the procedure for extracting the source features for mapping. Section 3 describes the procedure used to generate the mapped excitation. A subjective test is conducted to evaluate the performance of the proposed mapping technique. The methodology and results are described in Section 4. In Section 5 we briefly describe its application to voice conversion. Finally in Section 6 we summarize the work presented in this paper.

2. Source and system mapping using MLFFNN

The problem of source mapping/conversion techniques has been discussed in the literature in the context of voice conversion applications. The proposed methods can be broadly classified into source modeling [4][5] and residual prediction techniques [6]. The majority of source modeling techniques require estimation of the parameters of the Liljencrants-Fant (LF) model [5]. The parameters of the source speaker are then transformed into one matching that of the destination speaker. Since the LF model is a model of the glottal volume-velocity, its parameters should be similar for both CSM and TM speech, as they are of the same speaker. Residual prediction techniques deal with building of residual codebooks, and determining techniques for selection and smoothing of residuals. A similar approach was proposed in [3] for coding of TM speech, where its contribution to enhancement was limited. In this paper, the ability of an MLFFNN to capture the non-linear relations in the source and system features of TM and CSM speech is used for enhancement of TM speech.

When mapping the system features, a fixed frame size of 25 ms is used to extract the linear prediction coefficients (LPC). These LPCs are then converted to corresponding cepstral coefficents (LPCCs) and linearly weighted. The linearly weighted LPCCs are used as the system features. While this method is suitable for the vocal-tract system, it is not useful in the extraction of source features for mapping. In fact, any mapping of the source features extracted using arbitrarily positioned windows of duration 20-30 ms (or less) would perform poorly. This is because, for voiced speech, the excitation can vary significantly with time. Even more so, the significant part of the excitation is the region around the instants of glottal closure in the residual. This instant corresponds to the most significant excitation during a pitch cycle. The residual in the region around the instants of glottal closure is used to extract the source information. The analysis window is anchored around the instants of glottal closure. If the size of the window is too long compared to a pitch period, the source features in the region around the following GCI could influence those of the current GCI, especially in the case of high pitched speech.

2.1. Estimating GCIs from TM speech

Since the source features are to be extracted around the GCIs, it is necessary to have a reliable and accurate technique for estimating GCIs from the speech signal. The problem of low SNR does not arise when processing TM speech as the effect of the background noise on the TM is negligible. Hence a technique to reliably estimate the GCIs from clean speech is essential for source mapping. In this paper, the method based on the zero-frequency (ZF) filter is used for extracting the instants of glottal closure [7]. The steps involved in estimating the GCIs from the speech signal are outlined below:

(a) Difference the speech signal $s[n]$ (to remove any time-varying low frequency bias in the signal)

$$x[n] = s[n] - s[n-1] \tag{1}$$

(b) Pass the differenced speech signal $x[n]$ twice through an ideal resonator at zero frequency. That is

$$y_1[n] = -\sum_{k=1}^{2} a_k y_1[n-k] + x[n], \tag{2a}$$

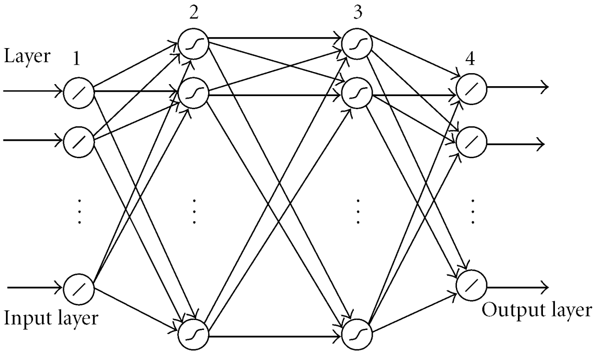

Figure 1: A 4 layer mapping neural network with 32L,80N,80N,32L, where L refers to a linear unit and N refers to a nonlinear unit.

and

$$y_2[n] = -\sum_{k=1}^{2} a_k y_2[n-k] + y_1[n], \tag{2b}$$

where $a_1 = -2$, and $a_2 = 1$.

(c) Remove the trend in $y_2[n]$ by subtracting the average over 10 ms at each sample. The resulting signal

$$y[n] = y_2[n] - \frac{1}{2N+1} \sum_{m=-N}^{N} y_2[n+m] \tag{3}$$

is called the zero-frequency filtered signal.

(d) The positive zero crossings in the filtered signal correspond to the locations of the instants of glottal closure

2.2. Source feature extraction

For a given speaker, TM and CSM speech are recorded simultaneously at a sampling frequency of 8 kHz. Using overlapping frames of duration 20 ms, with an overlap of 15 ms, a 10^{th} order LP analysis is performed. This yields sets of LP coefficients which are used to inverse filter the speech signal to obtain the prediction error. The excitation signal (residual signal) is generated from the prediction errors of the analysis frames through overlap-save (OLS). This method is used to derive the residual signals from both TM and CSM speech. The next step is to identify the voiced and unvoiced regions in the speech signal. TM speech has significantly higher energy in the voiced segments than in the unvoiced segments, when compared with CSM speech. The energy of the ZF filtered signal is also significantly higher for voiced regions than for unvoiced regions [8]. Hence the voiced regions can be easily identified from the ZF filtered signal derived from the TM speech. In the voiced regions, the locations of the positive zero crossings in the ZF filtered TM speech signal are identified as the GCI locations.

2.3. Source mapping using MLFFNN

Using the GCIs in the TM residual signal as anchor points for the analysis frames, 4ms (i.e., 32 samples) of data for each frame is extracted from TM and CSM speech. This is repeated for successive GCI locations to form an input-output pair of 32-dimensional feature vectors. As in the case of mapping of system features [9], an MLFFNN is used to capture the implicit relation between the source features of CSM and TM speech. Once this relation is captured by the MLFFNN, given source features of a TM as input, it should be able to provide source features that have characteristics of CSM speech signals. Fig. 1 shows the structure of the MLFFNN used for mapping. The structure of the network in terms of the number of hidden layers and number of units in each hidden layer is not critical, except that there should be enough number of units in the hidden layers to achieve non-linear mapping. The choice of two hidden layers and 80 units in each layer has been determined empirically. Given a set of input-output pattern pairs $(\mathbf{x}_l, \mathbf{y}_l), l = 1, 2, \ldots L$, where $\mathbf{x}_l = (x_l[0], x_l[1], \ldots, x_l[32])$ and $\mathbf{y}_l = (y_l[0], y_l[1], \ldots, y_l[32])$ corresponding to the source features of TM and CSM for the frame anchored at the l^{th} GCI, the objective is to find a set of weights that capture the relationship between x_l and y_l. Once the relationship between the input-output pattern pairs has been captured, given some other x_l as input to the MLFFNN, the output $\hat{\mathbf{y}}_l$ will be an estimate of $\mathbf{y}_l$, for $l = 1, 2, \ldots L$. The error in the estimate is given by $||\mathbf{y}_l - \hat{\mathbf{y}}_l||^2$ for each l. This is achieved by iteratively determining a set of weights such that the total mean-squared error over all the input-output pairs used for training the MLFFNN is minimized. The average error E over all L input-output pattern pairs is given by

$$E = \frac{1}{L} \sum_{l=1}^{L} ||\, \mathbf{y}_l - \hat{\mathbf{y}}_l \,||^2 . \tag{4}$$

The estimated error for each presentation of $(\mathbf{x}_l, \mathbf{y}_l)$ is backpropagated from the output units to the hidden units, and is used to update the weights leading to the hidden units.

2.4. Alignment of the residual signals

The locations of GCIs computed from TM speech do not always align with those estimated from CSM speech. Furthermore, the distances between the location of the peaks in the Hilbert envelope of the residual and the GCIs estimated from TM speech differ from those of CSM speech. Hence there is a need to align the peaks of the Hilbert envelope in the residual obtained from CSM speech with that of the residual obtained from TM speech. Without such an alignment, the MLFFNN will be unable to effectively capture the relation between the non-linear regions of TM and CSM residuals using the extracted source features. This issue is addressed by repositioning the anchor points for extracting the source features using the procedure outlined below.

- Compute the GCI locations from the TM speech signal.
- Compute the Hilbert envelope [10] of TM and CSM residual signals.
- Using the GCIs as a reference for TM residual signals, find the locations of the maximum peak in the Hilbert envelope (in the region around the GCIs). For each GCI location there will be such a peak in the Hilbert envelope of the residual. A region of 3-4ms around the GCI is considered for this purpose. This peak will be the new anchor point to extract source features from TM residual.
- Repeat this for the CSM speech to obtain corresponding anchor points in the CSM residual.

2.5. Mapping of the system features using MLFFNN

The set of LPCs obtained in the generation of the residual signals are first converted to a set of corresponding LPCCs using the recursive relation given by [11].

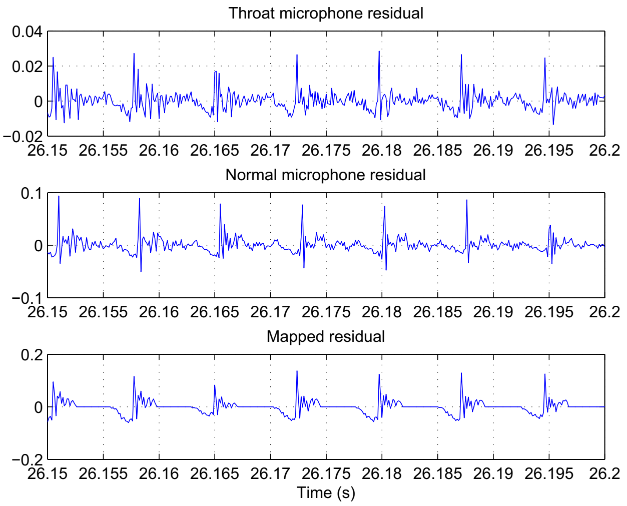

Figure 2: Source mapping of throat microphone residual. (a) Throat microphone residual, (b) close speaking microphone residual and (c) mapped residual.

$$c_n = \begin{cases} a_n + \sum_{k=1}^{n-1} \frac{k}{n} c_k a_{n-1} & 1 \leq n \leq p \\ \sum_{k=n-p}^{n-1} \frac{k}{n} c_k a_{n-k} & n > p \end{cases} \tag{5}$$

The coefficients are linearly weighted by $n = 1, 2, \ldots, q$. For each analysis frame, a set of 15 LPCs is obtained, to train the MLFFNN so that temporal information is also captured. For the i^{th} frame the LPCCs from the $(i-1)^{th}$, i^{th} and $(i+1)^{th}$ frames are concatenated to obtain a 45 dimensional feature vector. The 45 dimensional feature vectors so obtained from TM and CSM speech are used as the input and output pairs to train a 4-layer MLFFNN with the configuration 45L, 110N, 110N, 45L.

3. Generation of the excitation signal

Given the source features from the training data to train the MLFFNN, the goal is to capture the non-linear relation between the excitation source signal around the GCIs of TM and CSM speech. Once the network has been trained over sufficient number of input-output pairs of feature vectors, given the excitation source features of the TM as input, the output will be the estimated CSM source features. If the neural network has indeed learned the relation between the TM and CS excitation source features, the estimated CSM source features will have characteristics similar to that of the corresponding CSM speech.

For mapping the TM source features, at each GCI location, 4ms of TM residual data is provided as input to the trained MLFFNN, and its output is the estimated CSM residual for that input. Fig. 2 shows the LP residual segments corresponding to TM and CSM speech along with the estimated CSM residual segments. Since only 4ms around the GCI is used the rest of the samples in the segments have been set to zero. To generate the excitation signal, the segments around the TM residual which have been used for mapping are replaced with the estimated CSM segments. The excitation signal so obtained is used to excite the mapped LPCs.

Table 1: Average Itakura distances between CSM LP spectra and TM LP spectra, mapped LP spectra and mapped (with temporal information) spectra .

	TM	Mapped	Temporally Mapped
CSM	1.03	0.54	0.28

4. Evaluation

To evaluate the improvement in mapping of the system features obtained through incorporating temporal information, the Itakura distance [12] is used as an objective measure; the Itakura distance measures the similarity between two LP spectra. For 3 speakers (2 male, 1 female), 3 MLFFNNs are trained using 2-3 minutes of speech data per spealer. During testing 2 mts of speech data (consisting of 5 to 8 utterances per speaker), with no overlap with the training data, was used to evaluate the respective MLFFNNs ability to map the system features of the corresponding speaker. Table 1 shows the average Itakura distances for 2 minutes of test data between CSM and TM LPCs, CSM and mapped LPCs (without temporal information) and CSM and mapped LPCs (with temporal information). While the distance between CSM and TM spectra is largest, mapping using a MLFFNN has halved the distance. Use of temporal during training of the MLFFNN has reduced the distance even further.

Informal perceptual listening have indicated significant improvement over earlier methods. The presence of TM characteristics, if any, in the synthesized speech was found to be negligible at most. In general it has been observed that the quality of the speech using the mapped residual and mapped LPCs was much better than when only mapped LPCs were used. When a mapped residual was used as a source, the difference in quality of the synthesized speech using either CSM LPCs or mapped LPCs was observed to be negligible.

4.1. Limitations in the proposed approach

The proposed mapping technique does not explicitly consider the relative gains of TM and CSM residuals. This limits the quality of the enhanced speech to some extent. This method is also suitable only for voiced regions. Initial attempts at mapping unvoiced regions was found to produce little or no improvement in the quality of the enhanced TM speech. These two issues need to be addressed if the quality of the TM speech is to be enhanced further. The proposed method for source mapping as in the case of system mapping is still speaker-dependent. A speaker-independent mapping technique for both source and system is essential for wider applications

5. Extension to voice conversion applications

The proposed technique may also be applied for mapping the source features in voice conversion applications. Most voice conversion techniques only map the system features, and modify the average pitch of the source speaker to match that of the destination speaker. The region around the glottal closure is also a characteristic of a speaker. Hence by mapping the source using the proposed approach, the quality of the converted voice can be further improved. The main issue in voice conversion is the alignment of the residual signals corresponding to source and destination speakers.

6. Conclusions

For enhancement of TM speech it is not sufficient to merely map the spectral features, which characterize the vocal tract system. The source features, especially in the glottal closure region, significantly affect the perceptual quality of speech. Hence any attempt to enhance the TM speech should involve mapping of both the source and spectral features. This paper demonstrates that an MLFFNN can be used to capture the nonlinear mapping between the TM and CSM residual in the regions of glottal closure.

Although the gains and the unvoiced regions are not mapped in the TM residual, the quality of the TM speech was found too have improved significantly by mapping the source and system features. The proposed approach can also be extended to mapping of sources for voice conversion application.

7. References

[1] A. Uncini, Gobbi, and F. Piazza, “Frequency recovery of narroband speech using adaptive spline neural networks,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Ariz, USA, Mar. 1999, pp. 997–1000.

[2] A. Shahina and B. Yegnanarayana, “Mapping speech spectra from throat microphone to close-speaking microphone: A neural network approach,” *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 1–10, Jul. 2007.

[3] K. S. R. Murty, S. Khurana, Y. U. Itankar, M. R. Kesheorey, and B. Yegnanarayana, “Efficient representation of throat microphone speech,” in *Proc. Interspeech 2008*, Brisbane, Australia, Sep. 2008, pp. 2610–2613.

[4] D. G. Childers, “Glottal source modeling for voice conversion,” *Speech Commun.*, vol. 16, no. 2, pp. 127–138, 1995.

[5] A. del Pozo and S. Young, “The linear transformation of lf glottal waveforms for voice conversion,” in *Proc. Interspeech 2008*, Brisbane, Australia, Sep. 2008, pp. 1457–1460.

[6] D. Sndermann, A. Bonafonte, H. Ney, and H. Hge, “A study on residual prediction techniques for voice conversion,” in *IN INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING*, 2005, pp. 13–16.

[7] K. Sri Rama Murty and Yegnanarayana B., “Epoch extraction from speech signals,” *IEEE Trans. Audio Speech and Language Processing*, vol. 16, no. 8, Nov 2008.

[8] K. Sri Rama Murty, Yegnanarayana B., and Anand Joseph M, “Characterization of glottal activity from speech signals,” *IEEE Signal Processing Letters*, vol. 16, no. 6, June 2009.

[9] Anand Joseph M., B. Yegnanarayana, Sanjeev Gupta, and M. R. Kesheorey, “Speaker dependent mapping for low bit rate coding of throat microphone speech,” in *Proc. Interspeech 2009*, Brighton, UK, Sep. 2009, pp. 1087–1090.

[10] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Englewood Cliffs, New Jersey: Prentice Hall, 1975.

[11] J. Makhoul, “Linear prediction: A tutorial review,” *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[12] J. R. Deller, Jr., J. G. Proakis, and J. H. Hansen, *Discrete Time Processing of Speech Signals*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1993.