INTERSPEECH 2009 BRIGHTON

Speaker Dependent Mapping for Low Bit Rate Coding of Throat Microphone Speech

M. Anand Joseph[1], *B. Yegnanarayana*[1], *Sanjeev Gupta*[2] *and M. R. Kesheorey*[2]

[1]International Institute of Information Technology, Hyderabad, India
[2] Center for Artificial Intelligence and Robotics, Bangalore, India

anandjm@research.iiit.ac.in, yegna@iiit.ac.in,
sanjeev.guptapg08@research.iiit.ac.in, cair3@vsnl.net

Abstract

Throat microphones (TM) which are robust to background noise can be used in environments with high levels of background noise. Speech collected using TM is perceptually less natural. The objective of this paper is to map the spectral features (represented in the form of cepstral features) of TM and close speaking microphone (CSM) speech to improve the former's perceptual quality, and to represent it in an efficient manner for coding. The spectral mapping of TM and CSM speech is done using a multilayer feed-forward neural network, which is trained from features derived from TM and CSM speech. The sequence of estimated CSM spectral features is quantized and coded as a sequence of codebook indices using vector quantization. The sequence of codebook indices, the pitch contour and the energy contour derived from the TM signal are used to store/transmit the TM speech information efficiently. At the receiver, the all-pole system corresponding to the estimated CSM spectral vectors is excited by a synthetic residual to generate the speech signal.

Index Terms: linear prediction, neural network, spectral mapping, speech coding, throat microphone, vector quantization

1. Introduction

In environments with high levels of background noise, the intelligibility of speech signals collected through a close speaking microphone (CSM) is poor. In these noisy environments, transducers such as bone-conducting and throat microphones which pick up speech conducted as vibrations through the bones and tissues of the human body are more resilient to background noise, and are preferred for recording speech signals. A speech signal collected through a CSM is primarily due to the spectral shaping of the excitation by the vocal tract system consisting of the oral and nasal cavities. Unlike a CSM, a throat microphone (TM), due to its proximity to the larynx, may be less sensitive to finer articulatory modifications in the oral cavity. The damping effect of the nasal cavity may also not be reflected in the TM speech. In the case of unvoiced sounds like fricatives, the TM captures a signal which is due to the airflow before the constriction in the oral cavity. The effect of the back cavity on the TM speech is that most of the high frequency components of the fricatives are filtered out. In the case of voiced stops, the TM captures the voicing activity during the closure phase as well as the resonances of the oral cavity before the constriction. These factors cause the TM speech signal to sound less natural and muffled compared to the CSM speech.

In spite of this drawback, the resilience of a TM to background noise makes it a preferred choice in environments where CSM speech is not intelligible or buried in high levels of background noise. TMs are also preferred when operational constraints preempt the use of CSM. Typical usage scenarios where TM is preferred include helicopter or fixed wing aircraft, airplane cockpit, armoured vehicles, naval vessels and heavy machinery.

This paper proposes a technique for encoding TM speech so that the decoded TM speech is perceptually more natural than the original TM speech, and also intelligible. Speech is produced when a time-varying vocal tract (system) is excited by a time-varying excitation (source). If we assume that the systems producing the speech signals recorded through the TM and CSM differ in some aspects (due to the reasons mentioned previously), the goal is to find a (non-linear) mapping between the systems producing TM and CSM speech signals. This approach is proposed in [1], where the spectral features of the TM and CSM are mapped using a multilayer feed-forward neural network (MLFFNN). The current work is an improvement on the technique proposed in [2], where the spectral features of TM and CSM are both mapped and encoded using joint vector quantization. In this paper, a MLFFNN is used for mapping the spectral features of TM and CSM speech. The mapped features are then converted to line spectral pairs (LSPs) and vector quantized. The code-book indices obtained through vector quantization (VQ) are then transmitted along with energy and pitch contours. At the receiver, the codebook indices are used to obtain the corresponding spectral features from the codebook. A residual signal is generated using the pitch and energy contours estimated from the throat microphone signal. This residual signal is used along with the decoded spectral features for synthesizing the speech signal. The paper is organized as follows. Section 2 describes the procedure for mapping the spectral features of CSM and TM speech. In Section 3, we describe the procedure for coding the mapped spectral features. Section 4 describes the procedure for decoding and synthesizing the speech signal. In Section 5 we evaluate the use of the proposed method for mapping and encoding of the TM speech. In Section 6, we discuss the limitations of this approach and possible ways of addressing them.

2. Mapping CSM and TM speech

The source-filter model is commonly used as the basis for speech production, wherein the excitation and the vocal tract are considered to be independent of each other. The vocal tract system can be modelled as an all-pole (AR) or a pole-zero (ARMA) model. For the purpose of mapping the TM and CSM speech, we assume that the vocal tract system is a linear time-invariant

system whose parameters can be obtained by linear prediction (LP) analysis. The coefficients for each frame of speech obtained using LP analysis are first converted to LP cepstral coefficients (LPCCs). Given the set of LP coefficients $\{a_n\}$, the LP spectrum for a frame of speech is given by

$$|H(k)|^2 = \left| \frac{1}{1 + \sum_{n=1}^{p} a_n e^{-j\frac{2\pi}{M}nk}} \right|^2, \quad k = 0, 1, \ldots, M-1, \tag{1}$$

where M is the number of spectral values. The inverse discrete Fourier transform (IDFT) of the logarithm of the LP spectrum gives the LPCCs $\{c_n\}$, which are obtained as follows:

$$S(k) = \log |H(k)|^2, \qquad k = 0, 1, \ldots, M-1, \tag{2}$$

$$c_n = \frac{1}{M} \sum_{k=0}^{M-1} S(k) e^{j\frac{2\pi}{M}kn}, \quad n = 0, 1, \ldots, M-1. \tag{3}$$

The first q cepstral coefficients are linearly weighted and are used to represent the features of the LP spectrum for each frame. Linear weighting of the LPCCs, by $n = 1, 2, ...q$, provides less emphasis on the lower order LPCCs which would otherwise dominate the error computation in the training of the multilayer feed-forward neural network (MLFFNN).

2.1. Feature mapping using MLFNN

Earlier studies [3, 4] have proposed techniques for bandwidth expansion of narrow-band signals. Some of them [5] use neural networks for bandwidth expansion. Here we use an MLFFNN to capture the implicit relation between the spectral features of CSM and TM speech. Once this relationship is captured by the MLFFNN, given spectral features of a TM as input, it should be able to provide spectral features that have characteristics of CSM speech signals. Fig. 1 shows the structure of the MLFFNN used for mapping. The structure of the MLFFNN in terms of no. of hidden layers and no. of units in each hidden layer is not critical, except that there should be enough no. of units in the hidden layers to achieve nonlinear mapping. The choice of two hidden layers and 30 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 $\mathbf{x}_l = (c_1^T, 2c_2^T, \ldots, qc_q^T)$ and $\mathbf{y}_l = (c_1^C, 2c_2^C, \ldots, qc_q^C)$ corresponding to the wLPCCs of the TM and CSM respectively for the l^{th} frame of the speech signal, the objective is to find a set of weights that capture the relationship between $\mathbf{x}_l$ and $\mathbf{y}_l$. Once the relationship between the input-output pattern pairs has been captured, then given some other $\mathbf{x}_l$ 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 total 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 back-propagated from the output units to the hidden units and used to update the weights of the hidden units.

3. Coding using vector quantization

The estimated CSM spectral features obtained using the MLFFNN is encoded using vector quantization (VQ). The number of clusters used for encoding is 2^B, where B is the number of bits required to code each frame. The choice of B depends on the trade-off between perceptual quality and bit-rate. The weighted cepstral features ($n\hat{c}_n$) estimated for each frame using an MLFFNN are deweighted to obtain the LPCCs, $\hat{c}_n$. From the cepstral coefficients, the power spectrum $\hat{P}(k)$ is estimated [6]. The LP coefficients are obtained by applying Levinson-Durbin algorithm on the autocorrelation coefficients computed from $\hat{P}(k)$ [6]. Since the LPCs are computed from an autocorrelation sequence, they correspond to a stable all-pole synthesis filter. As the LPCs are coefficients of a polynomial, VQ of LPCs may lead to unstable filters. These LPCs are therefore converted to line spectral pairs (LSPs) before quantization. LSPs are computed from antisymmetric ($P(z)$) and symmetric($Q(z)$) polynomials, which are related to the LP polynomial $A(z)$ as follows:

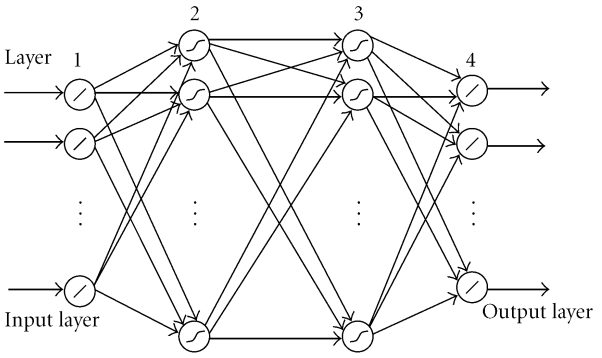

Figure 1: A 4 layer mapping neural network with 15L,30N,30N,15L, where L refers to a linear unit and N refers to a nonlinear unit.

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_p z^{-p}, \tag{5}$$

$$P(z) = A(z) - z^{-(p+1)} A(z^{-1}), \tag{6}$$

$$Q(z) = A(z) + z^{-(p+1)} A(z^{-1}). \tag{7}$$

The complex roots (θ_k) of the $P(z)$ and $Q(z)$ polynomials alternate in order on the unit circle in the z-plane. Any equivalent set of roots that alternate in this way will represent a stable LP filter. If θ_k is the set of complex roots, the LSPs (ω_k) expressed in radians are given by

$$\omega_k = \arctan\left(\frac{\Im(\theta_k)}{\Re(\theta_k)}\right), \tag{8}$$

where $\Re(.)$ and $\Im(.)$ denote the real and imaginary parts respectively. As the LSPs are roots of a polynomial, they are more tolerant to quantization errors and do not affect the spectra substantially, making them better suited for VQ based coding than the LPCs.

The LSPs obtained from the estimated CSM spectra are clustered into 2^B clusters using k-means algorithm. The cluster centers are initialized to vectors chosen arbitrarily from the training set. Clustering is performed by minimizing the sum of the squares of distances between the feature vectors and cluster centroids. The codebook thus formed is used at both the decoder as well as the encoder. The block diagrams for the encoding and decoding schemes are shown in Fig. 2.

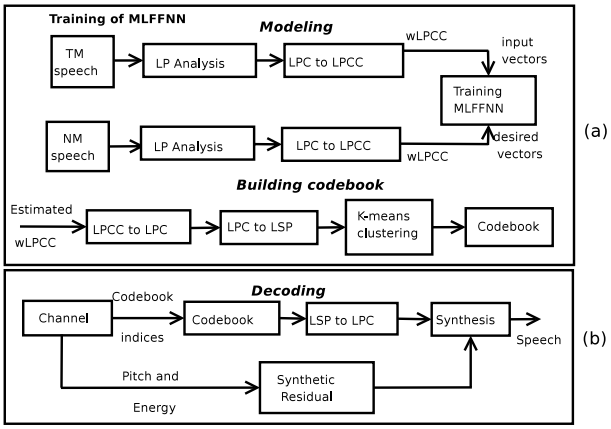

Figure 2: Block diagram of (a) mapping and encoding scheme of the spectral features derived from TM speech and (b) decoding and synthesis scheme.

3.1. Encoding source and system features

The cluster index is computed for the output of an MLFFNN for each frame of the TM speech. Along with the cluster indices, the energy and pitch contours also need to be transmitted. The energy contour is obtained from the TM residual for frames of 20 ms duration shifted by 10 ms. The pitch contour is obtained by autocorrelation analysis of the Hilbert envelope of the LP residual [7] of TM speech.

4. Decoding and synthesis

At the receiver, the encoded signal is decoded to obtain the sequence of cluster indices. The LSP values are obtained for each index using a codebook. The LSPs are converted back to LPCs, which are now an approximate representation of the spectra of CSM speech. Generation of the residual signal is similar to the approach in [2]. The energy and pitch contours are used to generate a residual signal. For each frame, a pre-stored residual signal for a pitch period is obtained from the CSM speech as a template for generating the synthetic residual. For every pitch period, the residual template is modified using pitch and energy contours of the TM speech signal. Increasing or decreasing the number of samples of the residual template is done by DFT interpolation [8]. The re-sampled residual signal is modulated with the energy contour to preserve the relative amplitudes of the sound units. As the duration, pitch and energy of the TM speech is used for synthesis, the true prosody of the TM speech is preserved in the synthesized speech signal. Fig. 2(b) shows the block diagram of the decoding and synthesis scheme described here.

5. Experimental studies

Speech data was collected simultaneously for a speaker using both CSM and TM. The data was collected in a laboratory environment where the average SNR is about 30dB. For testing, the TM speech data in principle can be collected in an environment with higher background noise levels, as the TM data is not influenced by noise. The alignment of TM and CSM speech is crucial for the MLFFNN to capture the relationship between them. Both the TM and CSM speech signals were sampled at 8 kHz. For training the MLFFNN, 5 minutes of speech data was used. With a frame rate of 100 frames per second, and a

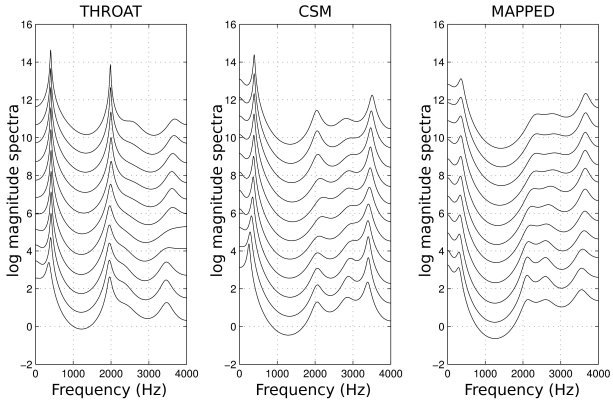

Figure 3: log magnitude LP Spectra of TM, CSM and the mapped log magnitude LP spectra for a sequence of frames.

duration of 20 ms for each frame, 10^{th} order LPCs are obtained for every frame. A 15 dimensional wLPCCs vector is derived from the LP spectrum of every frame. The input-output pattern pairs are presented in a batch mode to the neural network for training. The MLFFNN used for mapping the spectral features of the TM and CSM has 2 hidden layers each consisting of 30 non-linear units and input and output layers each consisting of 15 linear units. The MLFFNN is trained by presenting the wLPCCs of the TM as its input. The error between the output of the MLFFNN and the wLPCCs of the corresponding frame of CSM speech is used to update the weights of the hidden layer using backpropagation.

For testing, another set of TM and CSM speech data with a duration of 5 minutes is recorded. The wLPCCs for the TM from the test data set is provided as input to the trained MLFFNN. For each frame of TM wLPCCs, we obtain a corresponding frame of estimated CSM wLPCCs as output from the MLFFNN. As in the case of the training data, a frame rate of 100 frames per second and a frame duration of 20 ms is used. The output wLPCCs are converted to 10^{th} order LPCs and then to LSPs. The LSPs are vector quantized using *k*-means clustering. Euclidean distance is used as a distortion measure. The choice of the number of cluster centers depends on the amount of data and on the tolerable quantization error. In this paper, 1024 clusters are used for building the codebook. The codebook is made available to the receiver. During transmission, for every frame of TM speech, wLPCCs are computed, and using the trained MLFFNN, the estimated wLPCCs are obtained. The sequence of the estimated wLPCCs is represented as a sequence of codebook indices through VQ. The sequence of the codebook indices can be encoded at 1 kbps (1024 cluster centers, and 100 frames per second). Additionally we also require to transmit the pitch and energy contours derived from the TM speech signal. The pitch and energy contours can be encoded with fewer number of bits by exploiting their slow varying characteristics. On an average a bit-rate of 1.5 kbps can be realised using this method.

5.1. Results

Fig. 3 shows a sequence of log magnitude LP spectra for segments of TM, CSM and mapped speech. It can be visually observed that the MLFFNN is able to restore the high frequency components missing from the TM spectra, and that the spectra are continuous across the sequence of frames. The mapped spectra is also visually more similar to the CSM spectra than

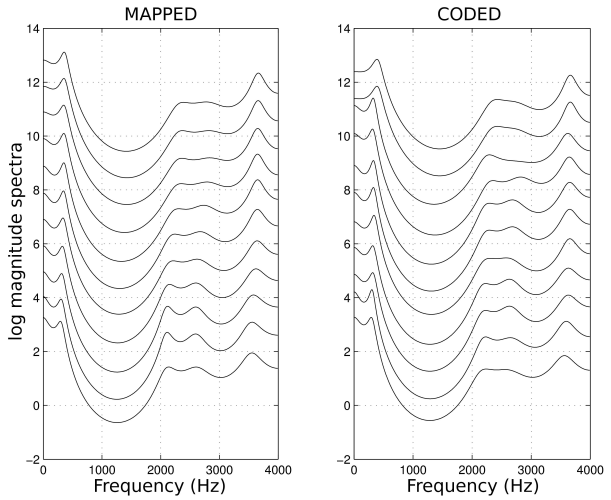

Figure 4: Mapped log magnitude LP spectra of CSM speech for a sequence of frames and corresponding coded log magnitude LP spectra.

the TM spectra. Even though they do not match perfectly, some features of the CSM spectra are brought out through mapping. At the receiver end, the codebook indices are used to retrieve the sequence of LSP vector centers. The LPCs derived from the sequence of LSP vectors represent the mapped spectra of TM speech signals. A synthesized residual is used to excite the all-pole model obtained through the LPCs. Fig. 4 shows the sequence of LP spectra of the mapped and coded frames. It can be observed that the quantization does not affect the spectral characteristics significantly.

To objectively measure the performance of the proposed approach to mapping and low bit-rate coding, we compute the Itakura distances [9] between the spectra of the CSM and the TM speech, the spectra of the CSM and the mapped speech and the spectra of the CSM and the coded speech. Fig. 5 (a) shows the Itakura distance computed for 100 frames (1 s) of the test data set between the CSM spectra and the mapped spectra. Fig. 5 (b) shows the Itakura distance computed between the CSM spectra and the coded spectra. The Itakura distance between the spectra of CSM and TM is shown in dashed lines in these figures. It can be observed that the Itakura distance for both the mapped and coded spectra is significantly less than for the TM spectra. Table. 1 shows the average Itakura distance for 5 minutes of the test data for TM, mapped and coded spectra. It can be seen that the distance between the spectra of the CSM and the TM is the largest. In the case of mapped spectra this distance is halved indicating that the MLFFNN indeed captures the implicit relation between the spectra to some extent. In the case of the coded spectra, there is an insignificant increase in the distance due to quantization.

Table 1: Average Itakura distance between the CSM spectra and the TM spectra, mapped spectra and coded spectra.

	TM	Mapped	Coded
CSM	1.19	0.58	0.61

6. Summary and conclusions

In this work, a system is proposed which maps and encodes TM speech signal at low bit-rates. This method uses an MLFFNN, trained with pairs of spectral features derived from frames of

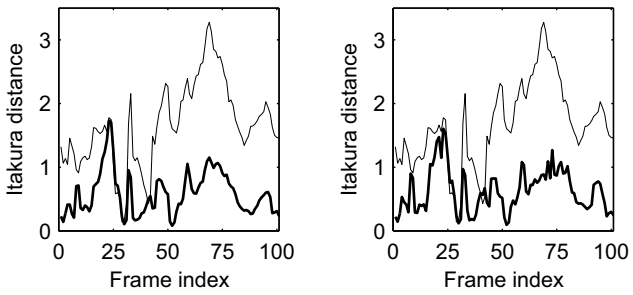

Figure 5: Itakura distance computed between (a) CSM spectra and mapped spectra and (b) CSM spectra and coded spectra are shown for 100 frames. The Itakura distance between the CSM and TM spectra is plotted using thinner lines in both figures.

simultaneously recorded TM and CSM speech, for estimating CSM-like spectral features from TM speech. This mapping is speaker-dependent as the MLFFNN is trained with spectral features for a particular speaker. The mapped spectral features are quantized using VQ to achieve low bit-rate coding. Using the proposed method, bit-rates of about 1.5 kbps can be achieved. The speech signal reconstructed using spectral features obtained at such low bit-rates was observed to be of perceptible quality.

The focus of the current work has been on mapping of spectral features (system) of the TM and CSM speech signals. The primary limitation of the proposed approach is in the synthesis of the residual signal (source). Since the residual signal is generated using a single residual template, the synthesized speech, while intelligible, suffers from the distortion due to discontinuities at the concatenation points. These discontinuities affect the perceptual quality of the synthesized speech. A future objective is to focus on alternative techniques of generating a residual signal that improves the perceptual quality of the synthesized speech signal.

7. References

[1] A Shahina and B Yegnanarayana, "Mapping speech spectra from throat microphone to close-speaking microphone: A neural netowrk approach," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 2, June 2007.

[2] K. Sri Rama Murty, Saurav Khurana, Yogendra Umesh Itankar, M R Kesheorey, and B. Yegnanarayana, "Efficient representation of throat microphone speech," in *INTERSPEECH*, Brisbane, Australia, Sept. 2008.

[3] B Geiser, P Jax, and P Vary, "Artificial bandwidth extension of speech supported by watermark-transmitted side information," in *INTERSPEECH*, Lisbon, Portugal, Sept. 2005, pp. 1497–1500.

[4] J A Fuemmeler, R C Hardie, and W R Gardner, "Techniques for the regeneration of wideband speech from narrowband speech," *EURASIP Journal on Applied Signal Processing*, vol. 2001, no. 4, pp. 266–274, 2001.

[5] 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.

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

[7] S R M Prasanna and B Yegnanarayana, "Extraction of pitch in adverse conditions," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Canada, May 2004, pp. 109–112.

[8] K S Rao and B Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 3, pp. 972–980, May 2006.

[9] J R Deller and JG Proakis, *Discrete-Time Processing of Speech Signals*, Mcmillan, New York, USA, 1993.