interspeech 2008 incorporating SST 08

Efficient Representation of Throat Microphone Speech

K. Sri Rama Murty[1], *Saurav Khurana*[2], *Yogendra Umesh Itankar*[2]
M. R. Kesheorey[3] *and B. Yegnanarayana*[2]

[1]Department of Computer Science and Engineering, IIT Madras, Chennai-600 036, India
[2]International Institute of Information Technology, Hyderabad -500 032, India
[3] Center for Artificial Intelligence & Robotics, Bangalore, India.

ksrmurty@gmail.com, sauravkhurana@students.iiit.ac.in, yogendra@students.iiit.ac.in,
cair3@vsnl.net, yegna@iiit.ac.in

Abstract

The objective of this work is to represent the information in the speech signal picked up by a throat microphone (TM) in an efficient manner in terms of number of bits required. Since the TM signal is unaffected by ambient noise, it is possible to extract the required information effectively under different environmental conditions. A spectral mapping technique is proposed from the TM speech to normal microphone (NM) speech to improve the perceptual quality. The mapping is done using vector quantization of pairwise spectral feature vectors derived from each frame of TM and the corresponding NM speech signals. Once the codebook is formed, the spectral features from a TM signal are represented as a sequence of codebook indices. 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. From the received sequence of codebook indices, the NM spectral vectors are retrieved due to pairwise vector quantization of the feature vectors. A synthetic residual signal is generated at the receiver from prestored residual templates by incorporating the pitch and the energy. The synthetic residual signal is used to excite the system corresponding to the NM spectral vectors to generate the speech signal.

Index Terms: Throat microphone, speech coding, spectral mapping, vector quantization.

1. Introduction

Communication in adverse conditions makes it difficult to process the speech due to high levels of ambient noise at the input of the microphone. The intelligibility of speech transmitted through low-bit rate coders severely degrades due to high levels of noise present in the acoustic environment. One effective way of overcoming noise is to collect the speech signal through bone conduction using a throat microphone. Though the quality of the throat microphone (TM) speech signal may be different from that of a normal microphone (NM) speech signal, the TM signal is not severely affected by ambient noise. Therefore the high quality of the TM signal can be exploited for transmitting the speech information at a low-bit rate. The objective of this work is to represent the information in the speech signal picked up by a TM in an efficient manner in terms of the number of bits required. Since the TM speech signal is unaffected by noise and degradation, it is possible to extract the required information effectively even under different environmental conditions.

While the signal collected through vibration pickup placed at the throat (near the glottis) is clean, it does not sound natural like a close-speaking microphone. The TM speech signal is typically a low bandwidth signal, whereas the NM signal is of wide bandwidth. Because of conduction through the bones and skin, the high frequency components are attenuated in the TM signal. As a result, the speech collected through a TM sounds slightly muffled and metallic. However, it is possible to map the features of TM speech to obtain the features of NM speech.

The spectral mapping methods from narrowband speech to wideband speech aims at improving the perceptual quality of the narrowband signals. There are several approaches for reconstruction of wideband spectrum from narrow band spectrum. Codebook mapping approaches rely on one-to-one mapping between the codebooks of narrowband and wideband spectral envelopes [1] [2]. Neural network approaches exploit the nonlinear properties of the network to estimate the missing frequency components [3]. In [4], a multilayer feedforward neural network was used to capture the functional relationship between the spectral vectors of TM speech and NM speech.

In this work, we propose a vector quantization method for coding and mapping the spectral features of the TM speech. Two codebooks are generated for TM speech and NM speech, which have one-to-one correspondence between their entries. The given throat microphone signal is converted into a sequence of codebook indices (symbols) by performing vector quantization using TM codebook. The symbols thus derived can be encoded in fewer bits compared to the number of bits required to represent the signal. During decoding, the spectral features of the NM speech are derived by using the symbols and the NM codebook. The NM residual signal is synthesized using the pitch and energy contours of the throat microphone signal. In Section 2, we explain the proposed vector quantization method for spectral mapping from TM speech to NM speech. Section 3 discusses a method to generate the NM residual from the pitch and energy contours of TM speech. In Section 4, we illustrate the application of the proposed representation of the TM speech. Finally, in Section 5, we summarize the contributions of this paper, and discuss some limitations of the proposed method which prompt for further studies.

2. Spectral mapping through vector quantization

In this work, we use the linear prediction cepstral coefficients to represent the spectral information in the speech signal. In linear prediction (LP) analysis, each sample $s[n]$ is estimated as a linear weighted sum of past p samples [5]. The predicted

sample $\hat{s}[n]$ is given by

$$\hat{s}[n] = -\sum_{k=1}^{p} a_k s[n-k] \tag{1}$$

where p is the order of prediction, and $\{a_k\}$s are the linear prediction coefficients (LPCs). The LPCs are obtained by minimizing the mean squared prediction error over the analysis frame. The linear prediction cepstral coefficients (LPCCs) are derived from the LPCs through a recursive relation given by [5]

$$c_m = \begin{cases} a_m + \sum\limits_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k} & 1 \le m \le p \\ \sum\limits_{k=m-p}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k} & m > p. \end{cases} \tag{2}$$

Typically a cepstral representation with $q > p$ is used, where q is the number of LPCCs. The low-order LPCCs are sensitive to the spectral tilt and the high-order LPCCs are sensitive to noise and other forms of variability. Hence the cepstral coefficients are weighted by a bandpass lifter

$$w_m = \left[1 + \frac{q}{2} \sin\left(\frac{\pi m}{2}\right)\right], \qquad 1 \le m \le q, \tag{3}$$

to reduce the sensitivities. The weighting deemphasizes the LPCCs around $m = 1$ and $m = q$. The wLPCCs derived from the corresponding frames of TM speech signal and NM speech signal are used for spectral mapping.

2.1. Modeling

The wLPCCs extracted from a frame of the TM speech are appended with the wLPCCs extracted from the corresponding frame in the NM speech to form a joint-feature vector. The joint-feature vector represents the implicit relations between the spectral features of the TM speech and the NM speech. The joint-feature vectors of the entire speech data are pooled together and clustered to form a predecided number of clusters. For coding purposes, the number of clusters (M) is typically chosen in the form of 2^B, where B is the number bits required to transmit a quantized spectral vector. The choice of the number of clusters (M) depends on the amount of data available, and the tradeoff between quantization level and perceptual quality. The centroids of the clusters were initialized to arbitrarily chosen vectors from the training set. Clustering is done by minimizing the sum of squares of distances between the joint-feature vectors and the cluster centroids. The resulting cluster centroids are split to form a TM codebook and a NM codebook. The TM codebook is used at the encoding side, and the NM codebook is used at the decoding side. The block diagram of the proposed approach for speech coding and mapping is shown in Fig. 1.

The cluster centroids of the NM codebook are used to form the all-pole filter for speech synthesis. As the cluster centroid represents the mean of the wLPCCs of the frames that belong to a given cluster, it may not be physically generated by the speaker. Moreover, direct conversion of wLPCCs to LPCs is not guaranteed to result in a stable all-pole filter realization. In order to overcome this, we propose to store the LPCs of the nearest frame to the cluster centroid as the NM codeword instead of wLPCC vectors.

2.2. Encoding TM speech

During encoding, the LP analysis is performed on the TM speech signal that is to be transmitted. The LPCs are converted to LPCCs, and cepstral liftering is applied to obtain wLPCCs. Each of the wLPCC vectors is assigned a cluster index (0 to M-1) based on its similarity to the corresponding cluster centroid. The sequence of cluster indices is used for encoding and transmitting the spectral information of TM speech.

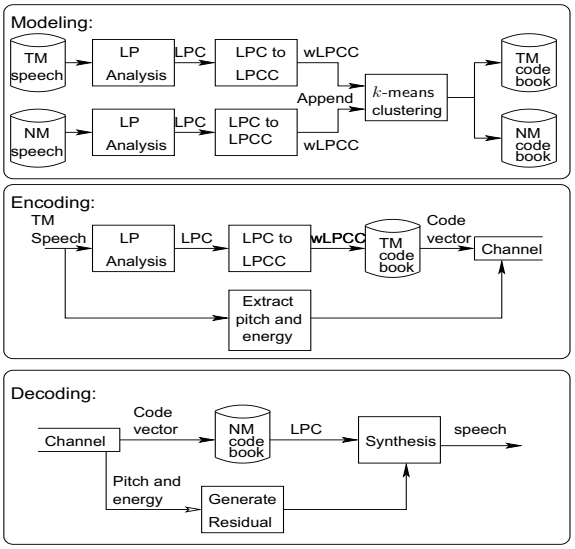

Figure 1: Block diagram of the proposed approach for pairwise vector quantization of spectral features derived from TM speech and NM speech

2.3. Decoding and mapping

The received signal is decoded to obtain the sequence of cluster indices. The NM cluster centroids corresponding to the received sequence of cluster indices are used in synthesizing the speech signal. The sequence of cluster centroids gives an approximate representation of the spectral features of the NM speech signal. The all-pole filter corresponding to the LPCs is excited by the synthetic residual signal, modified using the pitch and energy contours of the TM speech signal. Generation of synthetic residual is discussed in the next section.

3. Generating synthetic residual

The residual signal is generated by modifying a prestored residual template of a pitch period of NM speech using the pitch and energy contours of the throat microphone signal. The energy contour is obtained by computing the energy of the throat microphone signal for every 20 ms interval shifted by 10 ms. The pitch contour is obtained by the autocorrelation analysis of the Hilbert envelope of the LP residual, which is derived from the throat microphone signal [6]. Fig. 2 shows the pitch contour and the energy contour derived from a throat microphone signal. Notice that the pitch frequency and energy contours are smooth and can be compressed to obtain low-bit rate coding. The frames with a pitch frequency value of 0 indicate a non-voiced (unvoiced or silence) frame.

At the receiver, the residual corresponding to a pitch period of the NM signal is modified according to the received pitch and energy contours to generate a synthetic residual signal. Increasing or decreasing the number of samples of the LP residual signal according to the desired pitch period is done by exploiting the interpolation property of the discrete Fourier transform [7].

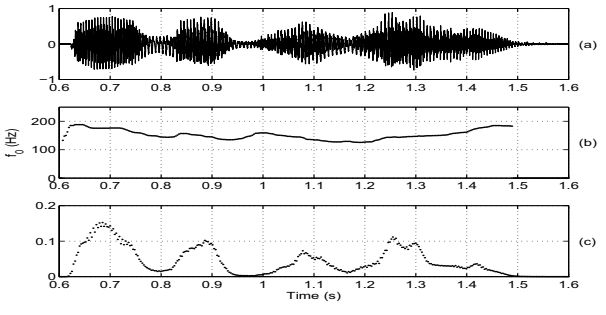

Figure 2: (a) TM speech signal, (b) Pitch contour derived from TM speech and (c) Energy envelope of TM speech

In this method, the residual samples can be resampled according to the pitch period of the frame, instead of deleting or inserting the samples arbitrarily. The process of resampling is illustrated in the Fig. 3. The resampling is done as follows: Let p be the number samples in the prestored template residual frame. Let q be the number of samples in the corresponding pitch period ($T_0 = \frac{1}{f_0}$) of the received frame, i.e., $q = T_0 * f_s$, where f_s is the sampling frequency. Resampling of the template residual $e[n]$ is performed by inserting $q-1$ zeros between successive samples of the $e[n]$. The resulting zero-padded residual signal $e_z[n]$ contains $p * q$ samples, and is given by

$$e_z[n] = \begin{cases} e\left[\frac{n}{q}\right], & n = 0,\ q,\ 2q,\ \ldots\ (p-1)q, \\ 0, & \text{Otherwise.} \end{cases} \quad (4)$$

A $p * q$ point discrete Fourier transform is performed on the zero-padded residual signal $e_z[n]$ to obtain its spectrum,

$$E_z[k] = \sum_{n=0}^{N} e_z[n] e^{-j\frac{2\pi}{N}kn}. \quad (5)$$

The resulting spectrum $E_z[k]$ is low-pass filtered up to the Nyquist frequency $\left(\frac{f_s}{2}\right)$ to preserve the spectral characteristics of the original residual signal, and thus avoiding the spectral folding due to upsampling. A $p*q$ point inverse discrete Fourier transform is performed on the low-pass filtered $E_z[k]$ to obtain the time-domain signal $e_{zi}[n]$ with interpolated samples. The desired number (q) of the residual samples are derived by selecting every p^{th} sample from the interpolated time-domain signal $e_{zi}[n]$. The modified residual frame $e_m[n]$ of length q samples is given by

$$e_m[n] = e_{zi}[n * p], \qquad n = 0,\ 1,\ 2, \ldots,\ (q-1) \quad (6)$$

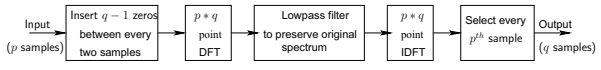

Figure 3: DFT interpolation technique for resampling the signal by a factor of $\frac{q}{p}$

In the voiced frames (indicated by a nonzero pitch frequency value), the resampled residual signal is used as the excitation. In the unvoiced frames (indicated by a zero pitch frequency value), white noise is used as excitation. The resampled residual signal is modulated with the energy contour to preserve the relative amplitudes of the sound units. Fig. 4 shows a segment of TM residual and the corresponding reconstructed NM residual. Since the true prosody (duration, pitch period and energy) of the TM signal is incorporated in the synthesized residual, it does not introduce significant perceptual distortions.

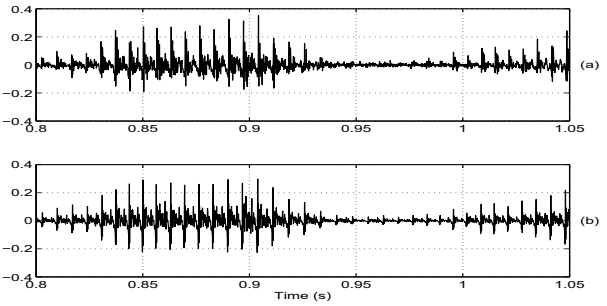

Figure 4: Illustration of the synthetic residual generated by proposed method. (a) Residual signal of TM speech and (b) Synthesized residual from pitch and energy contours of TM speech.

4. Experimental studies

The training phase involves recording speech from a speaker using the TM and the NM simultaneously. Simultaneous recording is essential for understanding the differences between components of speech in both the signals, and for capturing the implicit relations between the spectra of the two signals. For training, 5 minutes of speech data (read from a text, and containing both speech as well as nonspeech regions) is used. The speech signals from a TM and a NM are sampled at a rate of 8 kHz. A 10^{th} order LP analysis is used on (Hamming) windowed speech frames, each of 20 ms duration, at a rate of 100 frames per second. A 15-dimensional wLPCC vector is extracted from the 10 LPCs of each frame of the speech signal. The choice of LP analysis and the number of LPCCs is not critical. The 15-dimensional wLPCCs extracted from a frame of the TM speech signal are appended with the 15-dimensional wLPCCs extracted from the corresponding frame of the NM speech signal to form a 30-dimensional joint-feature vector. k-means clustering is performed on the joint-feature vectors extracted from the entire 5 min of speech data, with Euclidean distance as the distortion measure. The number of clusters was chosen to be 512. It was observed that there was no significant difference in the perceptual quality of the synthesized speech with $M = 512$ and $M = 1024$. The first 15 coefficients of each of the cluster centroids are stored as the TM codebook at the transmitter. The LPCs of the NM speech of the nearest frame to the cluster centroid is stored as the NM codeword.

During testing a new utterance (not used during modeling) is recorded from the TM, and 10th order LP analysis is performed on frames of 20 ms duration at a rate of 100 frames per second. The wLPCCs extracted from the test utterance are quantized into sequence of codebook indices. The pitch and energy contours are also obtained for every frame. The sequence of codebook indices, the pitch contour and the energy contour can be used to transmit the information in TM speech signal at a low-bit rate. The sequence of codebook indices can be encoded at 900 bits/s (512 clusters, 100 frames per second). Since the pitch and the energy contours are smooth, they also can be coded in smaller number of bits. On an average, we expect to

represent the information in the TM speech signal at 1000 bits/s using this method.

At the receiver, the codebook indices are used to retrieve the LPCs of the NM speech signal. The derived sequence of LPCs provide mapping from actual TM speech signal to the reconstructed NM speech signal. The LP spectra for a sequence of frames of the TM speech and NM speech, and the corresponding reconstructed spectra are shown in Fig. 5. The high frequency content missing in the TM spectra is incorporated in the reconstructed spectra. It is also seen from Fig. 5 that the codebook mapping seems to provide an estimate of the NM spectra which are smooth across consecutive frames. The residual signal generated from the pitch and energy contours (as described in Section 3) is used to excite the time-varying all-pole filter, corresponding to the retrieved LPCs, for synthesizing the speech signal.

The performance of the proposed mapping technique is evaluated using the Itakura distance measure as the objective criterion. The Itakura distance measures the similarity between two LP spectra [8]. The Itakura distance between the LPCs of two frames ($\mathbf{a}$ and $\mathbf{b}$) is given by

$$d_{ab} = \frac{\mathbf{b}^{\mathrm{T}}\mathbf{R_a}\mathbf{b}}{\mathbf{a}^{\mathrm{T}}\mathbf{R_a}\mathbf{a}} \tag{7a}$$

$$d_{ba} = \frac{\mathbf{a}^{\mathrm{T}}\mathbf{R_b}\mathbf{a}}{\mathbf{b}^{\mathrm{T}}\mathbf{R_b}\mathbf{b}} \tag{7b}$$

where d_{ab} and d_{ba} are the asymmetric distances from $\mathbf{a}$ to $\mathbf{b}$ and vice versa, respectively, and $\mathbf{R}_a$ and $\mathbf{R}_b$ is the Toeplitz matrices formed from the autocorrelation sequences of the speech frame corresponding to $\mathbf{a}$ and $\mathbf{b}$. The symmetric Itakura distance between two vectors is given by $d = \frac{1}{2}(d_{ab} + d_{ba})$. The Itakura distance between the TM spectra and NM spectra, and the NM spectra and reconstructed spectra are computed for each frame. Fig. 6 shows the Itakura distance plot for a segment of an utterance. It can be observed that the distance between the NM spectra and reconstructed spectra is smaller compared to the Itakura distance between the NM spectra and TM spectra. This shows that the reconstructed spectra are very close to the NM spectra. Thus, the proposed method of spectral mapping is able to capture the implicit relations among the TM spectra and NM spectra. Informal listening to the reconstructed speech (speech synthesized using LPCs derived from NM codebook and synthetic residual signal) showed that it is of perceptible quality.

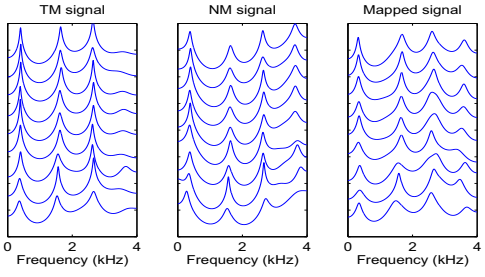

Figure 5: LP spectra of the TM speech, NM speech and the estimated LP spectra for a sequence of speech frames.

5. Summary and conclusions

In this paper, we proposed a framework for representing the information in the TM speech signal using a small number of bits.

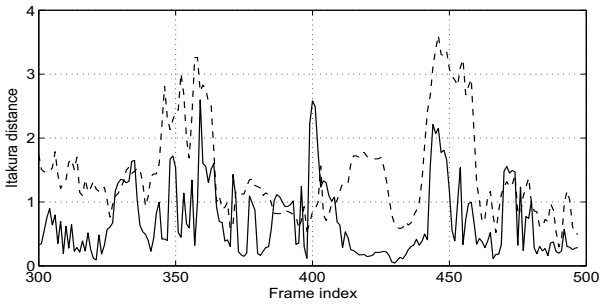

Figure 6: Itakura distance between the NM and TM spectra (dashed lines) and NM and estimated spectra (solid lines) for speech utterances.

The high SNR property of the TM speech signal is exploited for this efficient representation. The proposed vector quantization method also provides a speaker-dependent mapping of TM spectral features to NM spectral features to improve the perceptual quality. The information in the TM speech signal is represented as a sequence of codebook indices (corresponding to spectral information), and pitch and energy contours. Using the proposed method, we expect to represent the TM speech signal at 1000 bits/s, which can be reconstructed with reasonable perceptual quality. Our future efforts focus on generating a residual signal using more number of templates to improve the perceptual quality. Another objective is to perform spectral mapping in a speaker-independent manner.

6. References

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

[2] 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.

[3] 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.

[4] 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, pp. 1–10, June 2007.

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

[6] 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.

[7] 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.

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