ESTIMATION OF ACOUSTIC MICROPHONE VOCAL TRACT PARAMETERS FROM THROAT MICROPHONE RECORDINGS*

Ülkü Çağrı Akargün and Engin Erzin

Multimedia, Vision and Graphics Laboratory
Koç University, Sarıyer, İstanbul, 34450, Turkey
uakargun,eerzin@ku.edu.tr, http://mvgl.ku.edu.tr

ABSTRACT

Recently, joint processing of throat and acoustic microphone recordings has been an attractive tool for robust speech processing. As the throat microphones record the acoustic sounds in the form of vibrations from skin attached sensors, they are more robust and highly correlated with the acoustic speech signal. We investigate the correlation of throat and acoustic microphone recordings. We propose a hidden Markov model (HMM) based structure to estimate acoustic speech features from throat speech features. The HMM based estimator will be used to estimate clean acoustic speech features from noisy throat and acoustic microphone recordings. Experimental results on acoustic speech feature estimation are provided.

1. INTRODUCTION

Since speech recognition is a natural interface for human-human communication, it becomes a natural source of interface for man-machine communication as well. However, speech recognition could not find a wide range of application areas, which is partly due to the robustness problems under varying environmental conditions. In the last two decades, many research articles address robustness issues in speech recognition under varying environmental conditions. Some of the mainstream robustness studies include speech enhancement, cepstral mean subtraction, model adaptation, etc, where these studies target to increase recognition performance under adverse conditions. Beside these uni-modal (speech only) approaches, recently multi-modal approaches try to benefit from robust modalities, such as use of lip movements with audio-visual speech recognition, to help speech recognition process. Multimodal approaches are beneficial when they include environment independent but speech correlated modalities. In this study, we as well propose a multimodal speech signal processing, which is targeting to investigate joint processing of throat- and acoustic-microphone (TA) recordings.

The multimodal approaches become increasingly attractive in the last decade, and among these efforts the joint processing of TA recordings gained momentum in the last couple of years. As the throat microphones record the acoustic sounds in the form of vibrations using throat and skin attached sensors, these recordings are more robust than acoustic microphone recordings to environmental conditions. However, they represent a lower bandwidth speech signal content compared to open-air acoustic recordings. Since the throat-microphone recordings are more robust and highly correlated with the acoustic speech signal, they become attractive candidates for robust speech recognition applications. In one of the early works using the voice vibrations recorded from throat in speech recognition applications, the throat and speech signals are linearly combined yielding a robust estimate of noisy speech signal [1]. The clean acoustic speech features are estimated using probabilistic optimum filter (POF) mapping in [2]. POF mapping [3] is a piecewise linear transformation applied to noisy feature space to estimate the clean feature space. A device that combines a close-talk and a bone-conductive microphone is proposed by Microsoft research group [4-5] for speech detection using a moving-window histogram. Clean speech is estimated from the bone sensor signals [4]. The other approach [5] aims to learn the mapping from the bone sensor to the close-talk microphone. In order to reconstruct the clean speech signals, the predicted speech from the bone sensor is fused with the noisy close talk speech. In another combined acoustic and throat microphone approach [6], the speech recorded from throat and acoustic channels is processed by parallel speech recognition systems and in the decision stage,

* This work has been supported by TUBITAK under project EEEAG-104E176.

combination of these two sources yields a recognition more robust to background noise.

In this study we investigate the correlation of TA recordings to build a mapping between throat and acoustic microphone vocal track parameters.

2. ACOUSTIC-THROAT CORRELATION MODEL

Throat microphone speech recordings are correlated with acoustic microphone signal. We focused on the analysis of this correlation over the source filter model. Let us represent the acoustic and throat microphone recordings by y(t) and b(t), and the corresponding source filter model as,

$$Y(w) = E(w)H(w) + N(w) = E_1(w)H_1(w) \quad (1)$$

$$B(w) = E(w)H(w)M(w) = E_2(w)H_2(w) \quad (2)$$

Here, e(t) is the excitation signal, h(t) is the vocal track filter impulse response, n(t) is the possible additive environmental noise and m(t) is the filter impulse response that forms the vocal tract in throat microphone recordings. We can obtain source Ei(w) and vocal tract filter Hi(w) using source-filter analysis of TA recordings. Vocal tract filter, H2(w), that corresponds to throat microphone recordings is a low-pass version of the actual vocal tract, H(w), in this model. The throat microphone recordings are more robust to environmental noise than acoustic speech signals, hence they result a good representation of the low-pass nature of the actual vocal tract. We can obtain an estimate of the clean acoustic vocal track parameters from the noisy vocal track filter $H_1(w)$ and the throat microphone vocal track filter $H_2(w)$. Let us define this estimator as:

$$\hat{H}(w,t_n) = \Phi(H_1(w,t_n - \Delta \le t \le t_n + \Delta), H_2(w,t_n - \Delta \le t \le t_n + \Delta)) \quad (3)$$

Here, the function Φ(.,.) estimates clean vocal tract filter from the noisy vocal tract filter in the Δ neighborhood of the time instant *tn*. The synchronous clean TA recordings are processed with source filter analysis and line spectral pairs (LSPs), representing the vocal tract filter, are extracted synchronously for acoustic and throat microphones. The LSP parameters are calculated using a 16th order linear prediction filter over a window of size 20 msec for every 10 msec frame. Let us define the LSP vector for i-th channel and k-th frame as, $L_k^i = [l_{k,1}^i, l_{k,2}^i, ..., l_{k,16}^i]^T$. The L_k^1 represents the acoustic microphone filter models $H_1(w)$ and $H(w)$, which are equal to each other for the clean environment recordings. Likewise, L_k^2 vector represents the throat microphone filter model $H_2(w)$ for k-th window. Joint synchronous filter parameters can be represented as $L_k = [L_k^{1T}, L_k^{2T}]^T$ for k-th window.

We consider two different approaches to estimate acoustic microphone filter models from throat microphone filter models. The structure of the general estimator in (3) is modified to estimate the acoustic filter from the throat filter as follows:

$$\hat{L}_k^1 = \Phi(L_{k-\Delta}^2, L_{k-\Delta+1}^2, ...L_k^2, ...L_{k+\Delta-1}^2, L_{k+\Delta}^2) \quad (4)$$

This estimator is realized using two different approaches: the commonly used vector quantization approach and the Hidden Markov Model (HMM) based approach. The estimators and their performances are presented in the following sub-sections.

2.1. Vector Quantization Based Estimator

A vector quantizer can be designed to jointly quantize the synchronous TA vocal track filter parameters. This quantizer can be used to estimate acoustic filter model from the throat filter model. Let the joint vector quantizer L_{QB} is designed over the multi-stream filter parameter vectors L_k with 2^B levels using the Linde-Buzo-Gray (LBG) training. Hence each element of L_{QB} quantizer is a 32 dimensional joint acoustic and throat microphone filter parameters. L_{QB} vector quantizer can be split into two conjugate vector quantizers, L_{QB}^1 and L_{QB}^2, which will represent acoustic and throat channels. We can quantize any throat filter parameter vector L_k^2 for the k-th frame:

$$\bar{L}_k^2 = \underset{L_{QB}^2(i) \quad \forall i}{\arg\min} \| L_k^2 - L_{QB}^2(i) \|$$

$$I_k^2 = \underset{0 \le i < 2^B}{\arg\min} \| L_k^2 - L_{QB}^2(i) \| \quad (5)$$

Here, $\bar{L}_k^2$ is the quantized throat filter parameter vector and I_k^2 is the index of the quantized vector. We can estimate the acoustic filter parameter vector

using the quantized throat filter parameter vector $\bar{L}_k^2$ and the conjugate vector quantizer L_{QB}^1 as:

$$\hat{L}_k^1 = L_{QB}^1(I_k^2) \quad (6)$$

The estimation error between the estimate $\hat{L}_k^1$ and the original L_k^1 parameter vectors can be computed as the logarithmic spectral distortion between $\hat{H}_1(w)$ and $H_1(w)$ as follows:

$$d(H_1(w),\hat{H}_1(w)) = \frac{1}{2\pi}\int_{-\pi}^{\pi}[10\log\frac{1}{|H_1(w)|^2} - 10\log\frac{1}{|\hat{H}_1(w)|^2}]^2 dw \quad (7)$$

2.2. Hidden Markov Model Based Estimator

An unsupervised classifier based on Hidden Markov model (HMM) is used to jointly segment temporal acoustic and throat microphone filter parameters. Joint temporal filter parameter patterns are used to form a correlation between acoustic and throat filter parameters. This temporal correlation is used to estimate acoustic filter parameters from throat filter parameters.

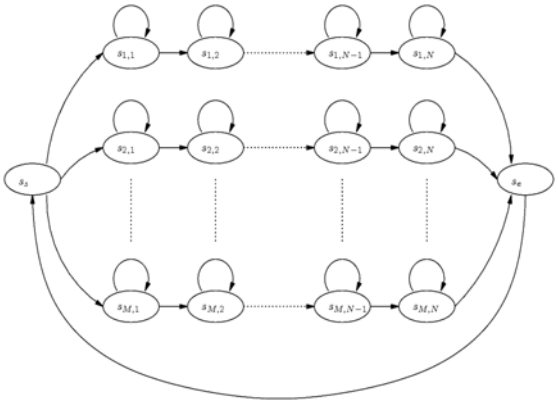

Figure 1: Hidden Markov model based unsupervised classifier

The multi-stream joint filter parameter vectors, L_k, are used to train a parallel branch HMM structure. The parallel branch HMM structure, as shown in Fig. 1, is used to perform unsupervised temporal clustering. In the parallel branch HMM structure, each branch corresponds to a temporal pattern. After the training process, the multi-stream HMM is split into acoustic only and throat only models. In the estimation process, the throat filter parameter sequence is temporally segmented using throat HMM, and in the resulting Viterbi state sequence the acoustic filter parameters are estimated as the mean vectors of the corresponding state probability distributions. In our simulations, each branch is selected as a 4-state left-to-right HMM, and the estimation performance is tested for varying number of branches and Gaussian mixtures.

3. EXPERIMENTAL RESULTS

We build a synchronous acoustic and throat microphone database, which consist of 400 sentence recordings from a single subject under clean conditions. In our experimental studies we split this database into two equal parts to perform training and testing of the vector quantizer and HMM based estimators.

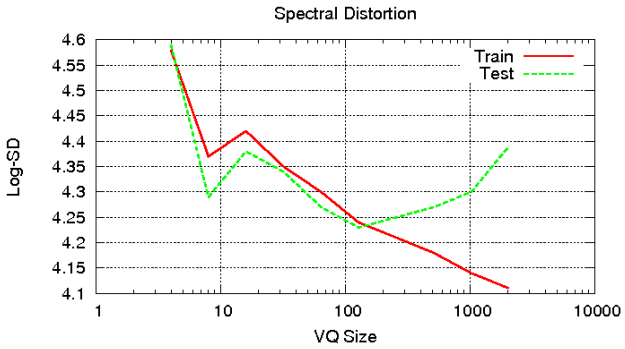

Figure 2: The average log-spectral distortion performance of the vector quantizer based estimator on training and test sets

The performance of the vector quantizer based estimator is analyzed for varying codebook sizes. The average log-spectral distortion values within train and test sets are given in Fig 2. As expected, spectral distortion tends to decrease as the dimension of the vector quantizer increases for the training set. As for the test data, first the spectral distortion tends to decrease and then increase. It is observed that the best estimation is obtained with a 128 codebook size.

A similar performance analysis is performed for the HMM based estimator. The average log-spectral distortion performances for varying number of classes with single and two Gaussian mixtures are given in Fig. 3 and 4, respectively. The best performances are obtained with 37-class HMM structure with single Gaussian mixture and 33-class

HMM structure with two Gaussian mixtures. Note that, there is equivalence between vector quantizer and HMM based estimators in term of total number of clusters. The HMM based estimator has 4 states in a single branch and 37/33 total branches, on the other hand vector quantizer based estimator has 128 codebooks. However, HMM based estimator captures the temporal changes, and with this property, it is observed that its log-spectral distortion is lower than the log-spectral distortion of the vector quantizer based estimator.

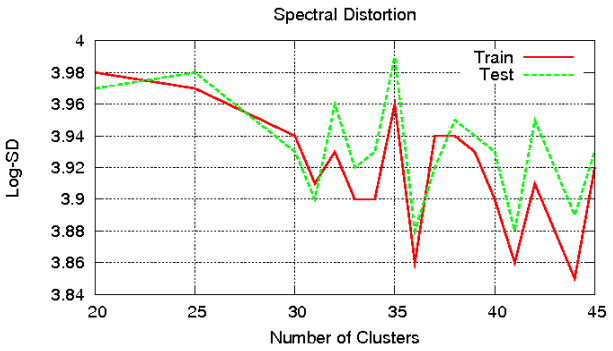

Figure 3: The average log-spectral distortion performance of the HMM based estimator with single Gaussian mixture on training and test sets over varying number of classes

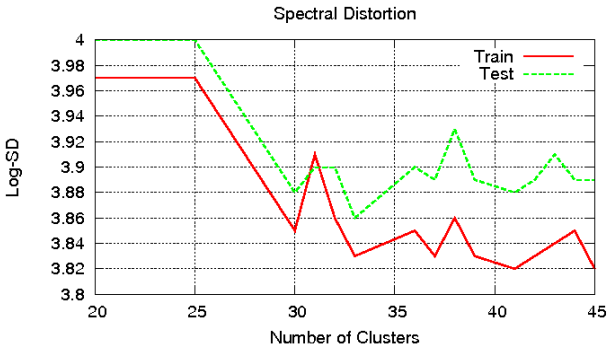

Figure 4: The average log-spectral distortion performance of the HMM based estimator with two Gaussian mixtures on training and test sets over varying number of classes

4. CONCLUSION

In this paper, we focused on the analysis of the correlation between the throat and acoustic microphone recordings. Vector quantization and HMM based estimators are examined under clean environment recordings to estimate acoustic filter parameters from throat filter parameters. We observed that the average log-spectral distortion values for the HMM based estimator is better compared to the vector quantization based estimator. As future work, we will study the correlation analysis of the throat and acoustic microphone recordings under noisy environments.

5. REFERENCES

[1] S. Roucos, V. Viswanathan, C. Henry, R. Schwartz, "Word recognition using multisensor speech input in high ambient noise", Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '86. , Volume: 11 , Pages:737 – 740, Apr 1986.

[2] M. Graciarena, H. Franco, K. Sonmez, H Bratt, "Combining Standard and Throat Microphones for Robust Speech Recognition", IEEE Signal Processing Letters, Vol. 10, No. 3, pp. 72-74, March 2003.

[3] L.Neumeyer and M.Weintraub, "Probabilistic optimum filtering for robust speech recognition," Proc. ICASSP Adelaide, Australia, pp. I417-I420, 1994.

[4] Y. Zheng, Z. Liu, Z. Zhang, M. Sinclair, J. Droppo, L. Deng, A. Acero, X. Huang, "Air- and Bone-Conductive Integrated Microphones for Robust Speech Detection and Enhancement", Proc. of ASRU 2003, U.S. Virgin Islands, Dec. 2003.

[5] Z. Zhang, Z. Liu, M. Sinclair, A. Acero, L. Deng, J. Droppo, X. D. Huang, Y. Zheng, "Multi-sensory microphones for robust speech detection, enhancement, and recognition", ICASSP04, Montreal, May17-21, 2004.

[6] S. Dupont, C. Ris, "Combined use of close-talk and throat microphones for improved speech recognition under non-stationary background noise," Proc. of Robust 2004 (Workshop (ITRW) on Robustness Issues in Conversational Interaction), Norwich, Aug. 2004.