MULTI-SENSORY SPEECH PROCESSING: INCORPORATING AUTOMATICALLY EXTRACTED HIDDEN DYNAMIC INFORMATION

Amarnag Subramanya

SSLI Lab,
University of Washington,
Seattle, WA - 98125.
asubram@ee.washington.edu

Li Deng, Zicheng Liu, and Zhengyou Zhang

Microsoft Research,
One Microsoft Way,
Redmond, WA - 98052.
{deng,zliu,zhang}@microsoft.com

Abstract

We describe a novel technique for multi-sensory speech processing for enhancing noisy speech and for improved noise-robust speech recognition. Both air- and bone-conductive microphones are used to capture speech data where the bone sensor contains virtually noise-free hidden dynamic information of clean speech in the form of formant trajectories. The distortion in the bone-sensor signal such as teeth-clacking and noise leakage can be effectively removed by making use of the automatically extracted formant information from the bone-sensor signal. This paper reports an improved technique for synthesizing speech waveforms based on the LPC cepstra computed analytically from the formant trajectories. When this new signal stream is fused with the other available speech data streams, we achieved improved performance for noisy speech recognition.

1. INTRODUCTION

Noise robustness is one of the major obstacles to mainstream adoption of speech processing systems. The presence of noise not only renders speech unintelligible but also results in poor performance of automatic speech recognition engines [8]. In [4, 5], a novel hardware solution was developed and described to combat against highly nonstationary acoustic noise such as background interfering speech. The device makes use of an inexpensive bone-conductive microphone in addition to the regular air-conductive microphone. The signal captured by the latter is corrupted by environmetal conditions, whereas the former is robust (to a large extent) to environmental noise. The bone sensor mostly captures the speech sounds uttered by the speaker but transmitted via the bone and tissues in the speaker's head. High frequency components ($> 4\ Khz$ at $16\ Khz$ sampling) are absent in the bone sensor signal.

The presence of two information streams requires the development of intelligent fusion techniques. The goal of this fusion process is to improve the overall intelligibility of speech, and recognition accuracy of ASR systems in noisy enviroments. In [4], we proposed an algorithm based on the SPLICE technique to learn the mapping between the two streams and the clean speech signal. One drawback of this approach is that it requires prior training of models of the signals in the two sensors. In [1], we proposed an algorithm that does not require proir training in order to estimate the clean speech signal. In [2], a fusion technique based on combining three streams, namely the air-conductive channel, the bone-conductive channel and a synthesized bone signal was developed. The work was based on the empirical observation that the underlying hidden dynamics of speech in the form of vocal tract resonances (VTRs) extracted from the bone-sensor signal and from the full-band clean speech signal are relatively invariant. Though this approach resulted in moderate gains in recognition accuracy with appropriate scaling of a confidence measure, a number of additional artifacts (not present in the original signal) were introduced into the synthesized bone signal.

One of the problems associated with the bone sensor signal in noisy environments is that a small amount of the noise leaks into the bone sensor. In [1], we proposed an algorithm to remove this leakage by estimating the transfer function between the two sensors during regions of non-speech activity. Another artifact that occur in the bone sensor are *teethclacks*. They are caused when the users' upper and lower jaws unconsiously come in contact with each other. They are characterized by high energy in the medium and high frequency bands. For detailed discussion of how teeth clacks effect the estimation of clean speech signal and and an algorithm for their removal, the reader is referred to [1].

In this paper we propose an alternative technique to remove both leakage noise and teethclacks. We track the hidden dynamics of the bone sensor and hence synthesize the clean bone sensor signal. This nonlinear procedure is capable of eliminating both leakage noise and teethclacks because the automatically tracked hidden dynamic variables with a fixed

0-7803-9332-5/05/$20.00 ©2005 IEEE

number are typically associated with high-energy regions of speech spectra, and the energy levels of the leakage noise and teethclacks are typically of lower energies. We also discuss an alternative technique to model the channel distortion from the close-talking channel to the synthezied bone channel.

This paper is organized as follows: We outline, in Section 2, the algorithm for automatically tracking the dynamics of low-frequency VTRs from the bone-sensor signal and synthesis of bone sensor signal using the VTRs. The details of the fusion algorithm are given in section 3, followed by experimental setups in section 4 and results in section 5.

2. BONE CHANNEL SYNTHESIS

2.1. Extracting VTR's from Bone Sensor

In [2], it was shown that the hidden dynamics of speech extracted from the bone sensor and those from the clean speech (generally unobserved in practice) are very close to each other[1]. This provides the incentive for using the bone-sensor data to infer such invariant clean speech's properties. In order to track the hidden dynamics of speech and extract VTRs, we make use of the recently developed adaptive Kalman filtering algorithm, reported separately in [6].

To enable VTR estimation, a state-space formulation of the speech dynamic model is first constructed. The state equation of which is given by

$$\boldsymbol{x}(t+1) = \boldsymbol{\Phi}\boldsymbol{x}(t) + [\boldsymbol{I} - \boldsymbol{\Phi}]\boldsymbol{u} + \mathbf{w}(t), \qquad (1)$$

where $\boldsymbol{x}(t)$ is the hidden dynamic vector of the VTR sequence:

$$\boldsymbol{x} = (\mathbf{f},\ \mathbf{b})' = (f_1, f_2, ..., f_P, b_1, ..., b_3, b_P)', \qquad (2)$$

consisting of resonance frequencies and bandwidths corresponding to the lowest P poles in the all-pole speech model. $\boldsymbol{\Phi}$ is the system matrix, and $\boldsymbol{u}$ is the averaged VTR target vector, providing the constraint on the (phone-independent) mean VTR values.

The observation equation of the speech dynamic model is

$$\mathbf{o}(t) = \mathbf{C}[\boldsymbol{x}(t)] + \boldsymbol{\mu} + \mathbf{v}(t), \qquad (3)$$

where $\mathbf{o}(t)$ is the observation sequence from the bone sensor in the form of LPC cepstra. The nonlinear function $\mathbf{C}[\boldsymbol{x}(t)]$ has the following explicit form:

$$C(i) = \sum_{p=1}^{P} \frac{2}{i} e^{-\pi i \frac{b_p}{f_s}} \cos(2\pi i \frac{f_p}{f_s}), \quad i = 1, .., \mathrm{I} \qquad (4)$$

[1]The correlation coefficient between the two sensors for the first four formants was found to be 0.98.

where f_s is the sampling frequency, i is the order of the cepstrum up to the highest order of $I = 15$, and p is the pole order of the VTR up to the highest order of $P = 4$. To account for the modeling error due to the missing zeros and additional poles beyond P (i.e., source as well as filter modeling errors), we introduce the (trainable) residual vector $\boldsymbol{\mu}$ in addition to the use of the zero-mean noise $\mathbf{v}(t)$ in Eq. 3.

To construct the adaptive Kalman filtering algorithm for optimal estimation of the VTR sequence $\boldsymbol{x}(t)$ from the cepstral sequence $\mathbf{o}(t)$, we perform adaptive piecewise linearization on the nonlinear observation equation (3). In the mean time, the residual mean vector $\boldsymbol{\mu}$ and variances in $\mathbf{v}(t)$ are adaptively trained in an iterative manner as detailed in [6].

2.2. Synthesizing spectra and waveforms from the extracted hidden dynamics

In order to syntheize the bone channel we first use Eq. 4 to generate a linear cepstral sequence using the extracted VTR sequence, which is then used to compute the magnitude spectrum for the given frame. In order to generate the waveform we make use of the phase of the noisy bone signal with the magnitude spectral features generated above. The complex spectrum of the synthesized bone signal is generated using

$$\hat{B} = B\sqrt{M^{-1}e^{\left(C^{-1}(\hat{B}_m - B_m)\right)}} \qquad (5)$$

where, M and C are the mel and dct filters repectively, B and $\hat{B}$ are the complex spectra of the original bone and synthesized bone signals respectively, B_m and $\hat{B}_m$ are the mel-cepstrum of the bone and synthesized bone channels respectively. The synthesized bone channel waveform may then be obtained from $\hat{B}$ using the overlap and add technique.

3. INFORMATION FUSION

The synthesized complex spectral sequence derived from the bone sensor, $\hat{B}(t,k)$ is combined with the two directly measured complex spectra, $Y(t,k)$, the close-talk signal and $B(t,k)$, the bone sensor, where (t,k), represents the k^{th} frequency bin at time t. The fusion rule to estimate the complex spectrum $X(t,k)$ of the clean speech signal is based on the following filtering model:

$$Y(t,k) = X(t,k) + \mathcal{N}(0,\sigma_1^2) \qquad (6)$$
$$B(t,k) = H(k)X(t,k) + \mathcal{N}(0,\sigma_2^2) \qquad (7)$$
$$\hat{B}(t,k) = G(k)X(t,k) + \mathcal{N}(0,\sigma_3^2), \qquad (8)$$

where $H(k)$ represents the bone microphone's channel distortion, and $G(k)$ represents the overall channel distortion

from clean signal to the synthesized speech in the bone channel, and $\sigma_1^2,\ \sigma_2^2,\ \sigma_3^2$ are the variances of the zero mean gaussian noise in the close-talk, bone and synthesized bone channels respectively. Under this model, an optimal (maximum likelihood) fusion rule can be shown to be

$$\hat{X}(t,k) = \frac{\sigma_2^2\sigma_3^2 Y(t,k) + \sigma_1^2\sigma_3^2 \bar{H}^*(k)B(t,k) + \sigma_1^2\sigma_2^2 \bar{G}^*(k)\hat{B}(t,k)}{\sigma_2^2\sigma_3^2 + \sigma_1^2\sigma_3^2|\bar{H}(k)|^2 + \sigma_1^2\sigma_2^2|\bar{G}(k)|^2} \tag{9}$$

where $\bar{H}$ is the estimated channel distortion function for the bone sensor [7]. There are two approaches that may be used to estimate $\bar{G}$, (a). to model it as the distortion between the close-talking channel and the extracted hidden dynamics and estimate $\bar{G}$ in a way similar to $\bar{H}$ or (b). to model the distortion between the bone channel and extracted hidden dynamics as say $\bar{G}_1$, and then set $\bar{G} = \bar{H}\bar{G}_1$ Henceforth we refer to these two approaches as Ω_1 and Ω_2 respectively. The channel distortion estimation technique assumes a linear relationsip between the two sensors. This assumption holds in a more stronger sense when modeling the distortion between the synthesized bone and bone channels, rather than the close-talking and the synthesized bone channels.

4. EXPERIMENTAL SETUPS

The test data was collected with two streams using the air- and bone-conductive microphone. One female speaker wears the headset and utters 42 sentences from the Wall Street Journal corpus in a cafeteria (ambient nosie level was 85 dbc) and in an office with a loud interfering speaker in the background. Henceforth we shall refer to these setups as test set ω_1 and ω_2 respectively. It should be noted here that the noisy utterances are not obtained by artificially corrupting the clean utterances, but are real-world noise corrupted.

In order to estimate the clean speech signal given the noisy utterance, we make use of the fusion rule discussed in section 3. The variances of the noise in the close-talking and the bone channel are estimated from the frames where speech activity is absent (for details related to the speech detector refer [4]). During our experiments, we found that the variance of the synthesized bone channel is zero for all practical purposes, which is a result of the noise removal by the synthesis algorithm. This (zero variance) is however, not desirable as it would cause the weights of both close-talking and bone channels to be zero. To avoid this condition we set the variance of the synthesized bone channel to be equal to the variance of the bone channel.

For all speech recognition experiments in this paper, we make use of Microsoft's internal large vocabulary HMM system, trained with a large amount of relatively clean speech data with a single stream. It should be noted here that the speech recognizer was not trained on any bone sensor data.

Setup	WER (ω_1)	WER (ω_2)
Y	45.00	38.15
$Y + B$	29.61	24.21
$Y + B + \hat{B}(\Omega_1)$	28.41	22.59
$Y + B + \hat{B}(\Omega_2)$	28.89	22.88

Table 1. Noisy speech recognition performance measured by percentage word error rate on a Wall Street Journal task: Y = close-talking channel, B = bone channel, and $\hat{B}$ = synthesized bone channel.

The bone-sensor data is then used to track VTRs. During our experiments we found that temporal smoothing of the VTRs improved performance from both perception and recognition aspects. Thus we smooth the VTRs using a $1-2-1$ kernel across time in the magnitude spectral domain and then synthesize speech waveforms (as explained in section 2). This synthetic data stream, together with the two original data streams, are fused to estimate the clean speech waveform. This is then fed to the HMM system for recognition.

5. RESULTS

Figures 1(a) and (b) shows the spectrogram of the original close-talking sensor and bone sensor data repectively for a particular utterance in the Wall Street Journal corpus. As it can be seen some of the background noise leaks into the bone sensor. Also some spikes corresponding to teethclacks may be observed in the spectrogram of the bone channel. Note that this artifact is not present in the close-talking channel. Figure 1(c) shows the synthesized bone sensor data obtained using [2]. Figure 1(d) shows the synthesized bone sensor data obtained using the approach detailed in section 2. It may be observed that the the spectrum in figure 1(d) gets rid of the leakage noise and also removes artifacts such as teeth clacks.

The results of our recognition experiments are shown in table 1. There is a significant improvement in performance as a result of the fusion of the close-talking and bone sensors (compare rows Y and $Y + B$). The best performance is achieved by fusing all three channels (i.e. close-talking, bone and synthesized bone), resulting in 4% and 6.69% relative improvement over the two channel case in case of cafeteria and office background noise types respectively. It should be noted here that these results were achieved without modifying any scale parameters in the fusion rule of section 3 (compare with [2]).

6. CONCLUSIONS AND FUTURE WORK

The results show that the hidden dynamics of speech for both the close-talking and the bone microphones are relatively invariant (in the lower frequency regions). In this paper, we have shown one of the means in which this similarity may be exploited to improve system performance. Also the new technique for synthesis does not introduce any artifacts into the signal and is sucessful in getting rid of leakage noise. A comparison of the results of Ω_1 and Ω_2 illustrates that modeling the relationship between the close-talking and synthesized bone channels as a convolution (Ω_2) does not result in any improvement in recognition accuracy.

In our future research, we plan to collect a large amount of simulataneous air- and bone-conductive microphone data so that, the recongizer can be trained on both the close talking and bone sensor data. In this way the recognizer can learn the mapping between the streams.

7. REFERENCES

[1] Z. Liu, A. Subramanya, Z. Zhang, J Droppo, and A. Acero, "Leakage Model and Teeth Clack removal for Air-and-Bone conductive microphones", *Proc. of ICASSP, Philadelphia*, 2005.

[2] L. Deng, Z. Liu, Z. Zhang and A. Acero, "Non-Linear Information Fusion in multi-sensor processing – Extracting and Exploiting hidden Dynamics of Speech Captured by a Bone-Condutive Microphone", *Proc. of MMSP, Italy*, 2004.

[3] I. Bazzi, A. Acero, and L. Deng. "An expectation-maximization approach for formant tracking using a parameter-free non-linear predictor," *Proc. ICASSP*, 2003, pp. 464-467.

[4] Y. Zheng, Z. Liu, Z. Zhang, M. Sinclair, J. Droppo, L. Deng, A. Acero, and X Huang. "Air- and bone-conductive integrated microphones for robust speech detection and enhancement," *Proc. IEEE ASRU Workshop*, Dec. 2003, St. Thomas, US Virgin Islands.

[5] Z. Zhang, Z. Liu, M. Sinclair, A. Acero, L. Deng, J. Droppo. X. Huang, Y. Zheng. "Multisensory microphones for robust speech detection, enhancement, and recognition," *Proc. ICASSP*, Montreal, Canada, May 2004.

[6] L. Deng, L. Lee, H. Attias, and A. Acero. "A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal tract resonances," *Proc. ICASSP*, Montreal, Canada, May 2004.

[7] Z. Liu, Z. Zhang, A. Acero, J. Droppo, and X. Huang. "Direct filtering for air- and bone-conductive microphones." *Proc. MMSP*, Siena, Italy, Sept. 2004.

[8] L. Deng and X. Huang. "Challenges in adopting speech recognition," *Communications of the ACM*, Vol. 47, No. 1, January 2004, pp. 69-75.

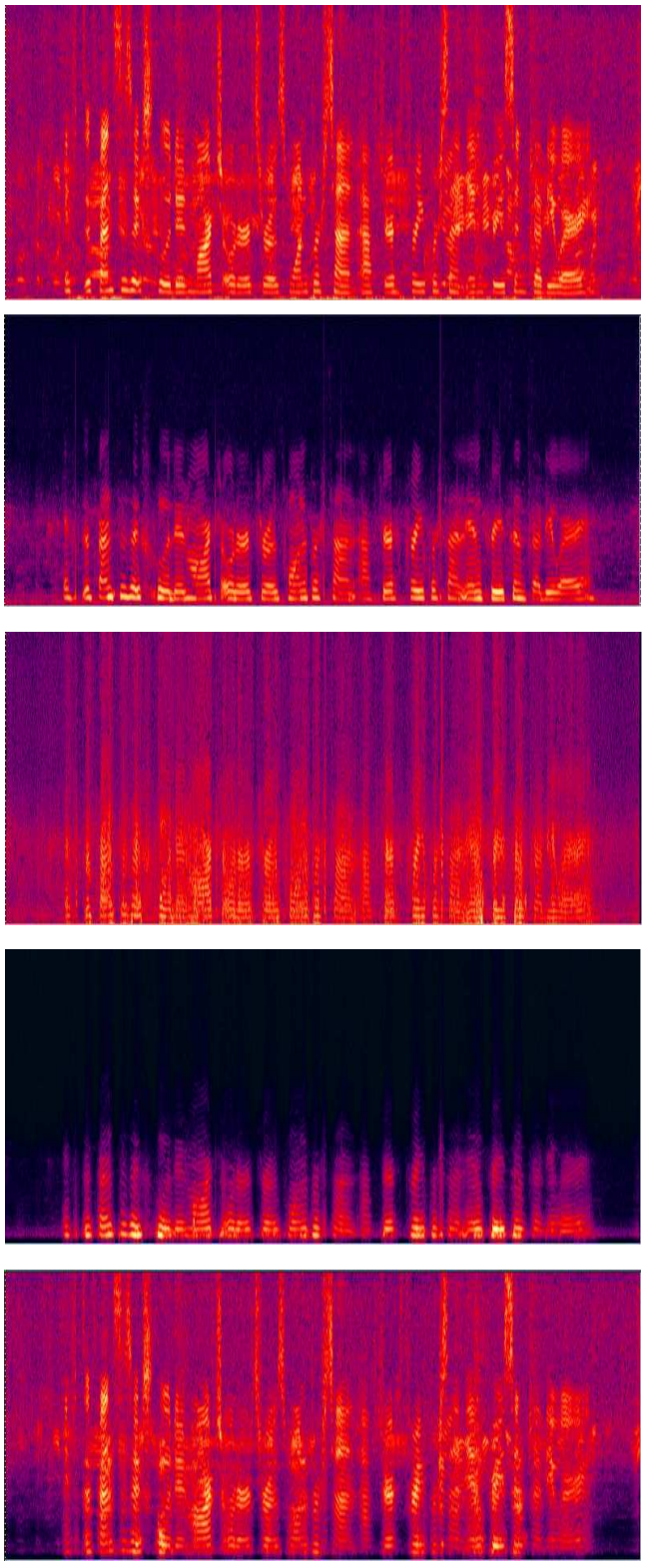

Fig. 1. Illustrations of five spectrograms (sequentially from top to bottom): a) the original close-talking sensor data; b) the original bone sensor signal; c) the synthesized bone sensor signal using [2]; d) the synthesized bone sensor signal from the new proposed algorithm and e) estimate of the clean signal ($Y + B + \hat{B}$).