NONLINEAR INFORMATION FUSION IN MULTI-SENSOR PROCESSING — EXTRACTING AND EXPLOITING HIDDEN DYNAMICS OF SPEECH CAPTURED BY A BONE-CONDUCTIVE MICROPHONE

Li Deng, Zicheng Liu, Zhengyou Zhang, and Alex Acero

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

ABSTRACT

One well-known difficulty in creating effective human-machine interface via the speech input is the adverse effects of concurrent acoustic noise. To overcome this challenge, we have developed a joint hardware and software solution. A novel bone-conductive microphone is integrated with a regular air-conductive one in a single headset. These two simultaneous sensors capture distinct signal properties in the speech embedded in acoustic noise. The focus of this paper is exploration of the type of dynamic properties that are relatively invariant between the bone-conductive sensor's signal and the clean speech signal; the latter would not be available to the recognizer. Our approach is based on a nonlinear processing technique that estimates the unobserved (hidden) vocal tract resonances, as a representation of such invariant hidden dynamics, from the available bone-sensor signal. The information about these dynamic aspects of the clean speech is then fused with other noisy measurements to aim at improving the recognition system's robustness to acoustic distortion. The fusion technique is based on a combination of three sets of signals including the synthesized speech signal using the vocal tract resonance dynamics extracted nonlinearly from the bone-sensor signal.

1. INTRODUCTION

Noise robustness remains one major obstacle to mainstream adoption of speech recognition [1]. In [5], a novel hardware solution was developed to combat against highly nonstationary acoustic noise, including background interfering speech. A separate, very inexpensive sensor using bone conduction (i.e., bone-conductive microphone) was added to the same headset that also mounts a regular, air-conductive microphone. The regular microphone captures the acoustic signal comprising speech mixed in noise, while the bone sensor captures mostly the speech sounds uttered by the intended speaker but transmitted via the bone and tissues in the speaker's head. Therefore, the external noise is heavily reduced in the bone sensor but the signal, especially the high-frequency components, is highly distorted along the non-air signal transmission path.

The challenge is to intelligently fuse these complementary signals in order to derive the original, undistorted clean speech signal which is often unavailable to any direct measuring microphone sensor under noisy acoustic environments. The work reported in this paper is based on the empirical observation that the underlying hidden dynamics of speech in the form of vocal tract resonances (VTRs) are relatively invariant between those extracted from the bone-sensor signal (under any acoustic condition) and from the clean speech signal. That is, the bone conduction has not introduced the distortion in the low-frequency regions which is severe enough to affect the general dynamic properties of VTRs. On the other hand, under noisy conditions, VTR extraction is severely affected using the air-conductive microphone that captures acoustic noise as well as speech. Using this invariance, we are able to infer key properties of clean speech from the bone-conductive microphone measurement, while these properties are difficult to infer using the regular microphone alone since it captures acoustic noise also.

This paper is organized as follows. We outline, in Section 2, a novel algorithm for automatically tracking the dynamics of low-frequency VTRs from the bone-sensor signal that are relatively resistive to interfering noises. Exploitation of the extracted VTR dynamics is presented in Section 3, where synthesis of speech using the VTRs is described and its use as the new data stream in a novel three-stream information fusion is described. Then, in Section 4, experimental results are presented to demonstrate strong correlations between VTRs extracted from the bone sensor under noisy environments and those from clean speech. Positive speech recognition results are also presented using the new three-stream fusion technique that makes use of automatically extracted hidden VTR dynamics.

2. EXTRACTING HIDDEN DYNAMICS OF SPEECH FROM BONE SENSOR

We use the recently developed adaptive Kalman filtering algorithm reported in [2] to extract the VTRs from the bone-sensor signal. The underlying assumption is that the extracted VTRs from the bone sensor under noisy conditions should reflect aspects of realistic speech dynamic properties that are largely independent of the acoustic environment. (Empirical evidence supporting this assumption is provided in Section 4.1.) To enable VTR estimation, we construct a state-space formulation of the speech dynamic model. The state equation in this formulation is

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

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

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

consisting of resonance frequencies and bandwidths corresponding to the lowest P poles in the all-pole speech model. $\boldsymbol{\Phi}$ is the

We thank J. Droppo for suggestions of improving waveform synthesis mentioned in Section 3.1

system matrix, and $\boldsymbol{u}$ is the averaged VTR target vector, providing the constraint on the (phone-independent) mean values of the VTR.

The observation equation of the speech dynamic model is

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

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:

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

where f_s is the sampling frequency, i is the order of the cepstrum up to the highest order of m, and p is the pole order of the VTR up to the highest order of P. 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 $\boldsymbol{v}(t)$ 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 [2].

3. EXPLOITING HIDDEN DYNAMICS: FUSION OF MULTIPLE SENSOR-DATA STREAMS

3.1. Synthesizing spectra and waveforms from the extracted hidden dynamics

As mentioned earlier, two sensor data streams, captured by the bone sensor and by the regular (air-conductive) microphone, respectively, have respective weaknesses in representing full properties of clean speech, which are not directly measurable, under noisy conditions. Given the extracted VTR dynamics that reflect one essential (but not complete) property of the non-measurable clean speech, we intend to create an additional data stream for the sensor fusion.

The specific technique we have developed is described here. First, we use Eq. 4 to generate linear cepstral sequence using the extracted VTR sequence. When the spectral distortion in the original bone sensor is weak, we add to these synthesized cepstra the residual mean vector $\boldsymbol{\mu}$ in Eq. 3. This compensates for, at least partially, the approximation errors of Eq. 3 to true cepstra due to well known limitations of the all-pole model of speech with finite orders. However, if the spectral distortion is severe (e.g., strong teeth clacking or noise leakage through the bone sensor), we remove the above compensation step since it would otherwise add back such spectral distortion to the synthesized cepstra. Second, we perform inverse discrete cosine transform (IDCT) on the synthesized cepstra above to generate the log-spectral sequence. From the log-spectral spectrograms, we have observed that different ways of adding the compensation vector $\boldsymbol{\mu}$ (e.g., varying the number of iterations in training) gave different tradeoffs between modeling inaccuracy and spectral distortion in the original bone-sensor signal. Finally, we exponentiate the above synthesized log-spectra, take square root, and then add the phase information derived from the bone-sensor signal. This gives a synthesized complex spectral sequence, and then a synthesized speech waveform, after applying the overlap-and-add technique.

3.2. Fusion of three input data streams

The above synthesized complex spectral sequence[1] derived from the bone sensor and denoted by $Y_3(t,k)$ is combined with the two directly measured complex spectra, $Y_1(t,k)$ for the close-talk air-conductive microphone and $Y_2(t,k)$ for the bone sensor. The FFT's frequency index is k for each window, and the windowed time (i.e., frame) sequence index is t. The fusion rule to estimate the complex spectrum $X(t,k)$ of clean speech is based on the following highly simplified linear filtering model:

$$Y_1(t,k) = X(t,k) + \mathcal{N}[0,\Sigma_1] \tag{5}$$

$$Y_2(t,k) = H(k)X(t,k) + \mathcal{N}[0,\Sigma_2] \tag{6}$$

$$Y_3(t,k) = G(k)X(t,k) + \mathcal{N}[0,\Sigma_3], \tag{7}$$

where $H(k)$ represents the bone microphone's channel distortion, and $G(k)$ represents the overall channel distortion (from clean speech to bone-sensor distorted speech and then to the synthesized speech), and $\mathcal{N}[0,\Sigma_1]$ denotes the normally-distributed random vector representing the additive interfering noise's spectrum. Under this model, an optimal (maximum likelihood) fusion rule can be shown to be

$$\hat{X}(t,k) = \frac{\Sigma_2\Sigma_3 Y_1(t,k) + \Sigma_1\Sigma_3 \bar{H}^*(k) Y_2(t,k) + \Sigma_1\Sigma_2 \bar{G}^*(k) Y_3(t,k)}{\Sigma_2\Sigma_3 + \Sigma_1\Sigma_3 |\bar{H}(k)|^2 + \Sigma_1\Sigma_2 |\bar{G}(k)|^2}, \tag{8}$$

where $\bar{H}$ is the estimated channel distortion function for the bone sensor [3]. $\bar{G}$ is estimated in a similar way, but estimation uses the synthesized speech waveform based on the extracted hidden dynamics instead of on the bone-sensor data directly.

To gain insight into the fusion rule (8), we rewrite it as the following weighted sum of three signal components:

$$\hat{X}(t,k) = W_1 Y_1(t,k) + W_2[\bar{H}^{-1} Y_2(t,k)] + W_3[\bar{G}^{-1} Y_3(t,k)], \tag{9}$$

where

$$W_1 = \frac{\Sigma_1^{-1}}{\Sigma_1^{-1} + \Sigma_2^{-1}|\bar{H}(k)|^2 + \Sigma_3^{-1}|\bar{G}(k)|^2} \tag{10}$$

$$W_2 = \frac{\Sigma_2^{-1}|\bar{H}(k)|^2}{\Sigma_1^{-1} + \Sigma_2^{-1}|\bar{H}(k)|^2 + \Sigma_3^{-1}|\bar{G}(k)|^2} \tag{11}$$

$$W_3 = \frac{\Sigma_3^{-1}|\bar{G}(k)|^2}{\Sigma_1^{-1} + \Sigma_2^{-1}|\bar{H}(k)|^2 + \Sigma_3^{-1}|\bar{G}(k)|^2}, \tag{12}$$

and where the three weights sum to one: $W_1 + W_2 + W_3 = 1$. Note the estimate expressed in (9) contains the inverse-filtered bone-sensor signal $[\bar{H}^{-1} Y_2(t,k)]$ and inverse-filtered synthetic signal $[\bar{G}^{-1} Y_3(t,k)]$.

4. EXPERIMENTS AND RESULTS

4.1. Results of Hidden Dynamics Tracking

In this section, we first show the results of tracking VTR dynamics, demonstrating that such inherent properties of speech in relatively invariant between the clean and noisy acoustic environments. This hence provides evidence supporting the rationale behind this work.

In Fig. 1, we show comparisons between the VTR frequencies extracted from the bone sensor with noisy speech input and those extracted from clean speech for a typical utterance in our speech

[1] In speech recognition experiments, due to system implementation constraints, we use the synthesized speech waveform. The front-end of the speech recognizer then takes a sequence of FFT's to obtain the complex spectra.

database. The noisy speech used as the input is created artificially by adding background interfering speech into the clean speech, with SNR=5dB. Close correlation between the two sets of VTR frequencies is observed, especially for the second resonance frequency (labeled as f_2 in Fig. 1), and the correlation coefficient is computed to be as high as 0.98 in this typical example. The correlation is found to be much lower between the VTR frequencies of clean speech and those extracted from the regular, air-conductive microphone. This is because the VTR extraction is subject to errors due to the interfering noise, especially for low-energy portions of the speech signal where the noise spectral components dominate the observed spectra with mixed speech and noise.

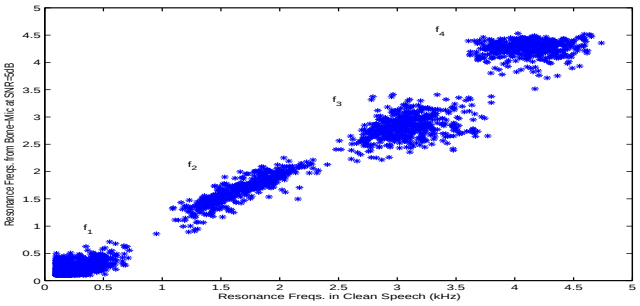

Fig. 1. Comparison between 1) VTR frequencies extracted from the bone sensor with noisy speech input (y axis); and 2) VTR frequencies extracted from clean speech (x axis).

More detailed illustrations are provided in Figs. 2 and 3 for the extracted VTR frequencies superimposed on the spectrograms of the clean speech signal captured by the air- and bone-conductive microphones, respectively. The same illustrations for noisy speech are shown in Figs. 4 and 5, respectively, for the two types of microphones. The bone sensor captures much less noise than the air-conductive microphone, while cutting off most of the high-frequency energies. However, the tracked f_1 to f_4 VTRs shown as the four separate lines in Figs. 2-5 are quite similar in values for clean speech in the air-conductive microphone (Fig. 2) and for the bone sensors (Figs. 3, and 5). In particular, the tracked VTRs from the signal captured by the bone sensor for noisy speech (Fig. 5) are much less affected than those by the air-conductive microphone (Fig. 4).

Another advantage of the nonlinear process of VTR extraction as discussed in this paper is to eliminate possible noise leakage to the bone sensor in the low-frequency region. Typically, the magnitude of such leaked noise is relatively small. When the spectral peaks due to the leaked noise are competing with those in the clean speech, the VTR extraction technique described in Section 2 is often effective in discarding the interfering spectral peaks. This is because not only the observation equation of the dynamic speech model ((3) in Section 2) tends to fit the spectral peaks with large magnitudes, but also the model incorporates powerful prior knowledge about the VTR dynamics in clean speech based on state equation (1). This constrains the possible range for each VTR component among a small number of total VTR components (four in the current implementation). Any spectral components due to interfering noise are likely to create mismatch to the prior dynamic patterns of clean speech and are thus likely to be rejected.

4.2. Preliminary Results on Noise-Robust Speech Recognition

We have conducted preliminary experiments on noise-robust speech recognition, where the largely invariant hidden dynamics of speech extracted from the bone-sensor signal are exploited. The maximum-likelihood fusion rule in (8) is used in the experiments. In implementing (8), the variances, Σ_1, Σ_2, and Σ_3, are estimated from three separate data streams, $Y_1(t,k), Y_2(t,k)$, and $Y_3(t,k)$, respectively, using the utterance-initial, speech-free frames. The estimation is carried out for each test utterance separately.

Because the fusion model for the synthesized spectral sequence $Y_3(t,k)$ in (7) is very crude, the estimate of variance, $\bar{\Sigma}_3$, using speech-free portions of $Y_3(t,k)$ sequence could be very inaccurate. One technique to compensate for such inaccuracy is to empirically scale the estimated variance.[2] Speech recognition results with this technique will be reported in this section.

In our experiments, we use a Microsoft's internal large vocabulary HMM system, trained with a large amount of relatively clean speech data with a single data stream (i.e., with no bone-sensor data). The test data are collected with two streams using simultaneous air- and bone-conductive microphones. One female speaker wears a headset mounted with both types of microphones and utters 42 sentences from the Wall Street Journal corpus in an office with a loud interfering speaker in the background. The bone-sensor data collected are used to track VTRs and then to synthesize speech waveforms. 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.

The speech recognition accuracy, listed as a function of the variance scaling factor, is shown in Table 1. The baseline accuracy is 72.21% for two-stream fusion, using $Y_1(t,k)$ and $Y_2(t,k)$ (and 55.00% for one-stream input using noisy air-conductive microphone speech $Y_1(t,k)$ only). Adding the new stream $Y_3(t,k)$ produces virtually no improvement if the unscaled variance is used in the fusion. However, when the variance scaling factor is increased to a value between five and six, a sizable accuracy improvement is obtained. When the variance scaling factor is further increased, the accuracy drops back to the baseline performance. Indeed, when $\Sigma_3 \to \infty$, Eq. 8 is reduced to

$$\hat{X}(t,k) \to \frac{\Sigma_2 + \Sigma_1 \bar{H}^*(k) Y_2(t,k)}{\Sigma_2 + \Sigma_1 |\bar{H}(k)|^2}, \qquad (13)$$

which is the two-stream fusion rule.

scale factor	0.1	0.2	0.5	1	2
Rec.Acc(%)	51.65	65.46	70.40	72.37	72.70

scale factor	3	5	6	10	1000
Rec.Acc(%)	73.45	74.11	74.10	72.81	72.21

Table 1. Speech recognition accuracy (%) as a function of the variance scaling factor, which adjusts the relative contributions of the new data stream and the original data streams in the information fusion. The baseline accuracy is 72.21% for two-stream fusion, and 55.00% for the one-stream noisy-speech input.

[2] Similar variance scaling was found useful in other noise-robust speech recognition experiments [4], where inaccuracy of the variance estimate was due to ignorance of the phase relationship between clean speech and interfering noise.

5. SUMMARY AND FUTURE WORK

In this paper we present a novel technique that recovers hidden VTR dynamics of speech under high-noise conditions. It uses the speech signal captured by a bone sensor that distorts high-frequency energies but retains most of noise-reduced low-frequency energies where major VTRs of speech lie. We have discovered that the extracted VTR frequencies from the bone sensor under noisy conditions have exceedingly high correlations with those extracted from clean speech. A three-stream fusion technique is further developed that capitalizes on the synthetic spectra or waveforms based on the extracted VTRs. While the fusion technique itself is linear, the new stream is derived from an original stream in a highly nonlinear fashion. This nonlinear VTR extraction process creates complementary signal properties in the new stream and is responsible for the speech recognition performance gain reported in our experiments.

While the largely invariant hidden dynamic properties of speech are discovered, extracted, and successfully exploited in this work, significant challenges remain for future research. VTR extraction from the bone sensor as a novel nonlinear processing technique can remove non-dominant noise components, but the synthesized spectra of speech also contain significant distortions compared with clean speech. To reduce such distortions while retaining the phonetically rich information contained in the extracted VTRs, we need to develop better synthesis techniques (e.g., formant vocoding). This will require careful source modeling of speech, rather than using the highly empirical compensation vector $\boldsymbol{\mu}$ as used in the current work. In addition, we have found significant phase distortions in the synthetic waveforms. Our current speech recognizer implementation requires the input in the form of waveforms. In the new implementation that directly accepts cepstral-based features, the effect of such phase distortions can be eliminated. Finally, the fusion technique experimented in this work is at the waveform level, without any modification of any element in the speech recognizer. When new fusion techniques are developed at the feature or decision level that requires expansion of the recognizer's feature set or modification of the decision rule, greater recognition performance improvement is expected.

6. REFERENCES

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

[2] 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.

[3] 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.

[4] L. Deng, J. Droppo, and A. Acero. "Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features," *IEEE Trans. Speech and Audio Processing*, Vol. 12, No. 3, May 2004, pp. 218-233.

[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.

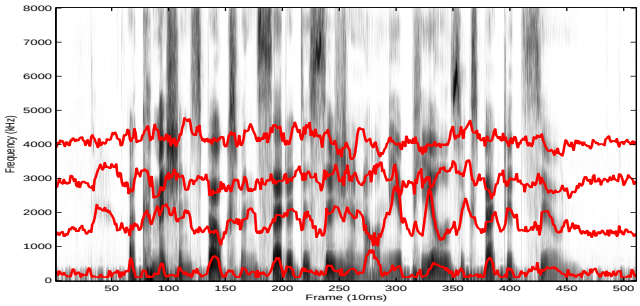

Fig. 2. Extracted VTRs (f_1 to f_4) superimposed on the spectrogram for the "clean" (female) speech recorded by the regular, air-conductive microphone in a quiet offi ce environment. Utterance: *A commission spokesman said a decision on the appeal is expected soon.*

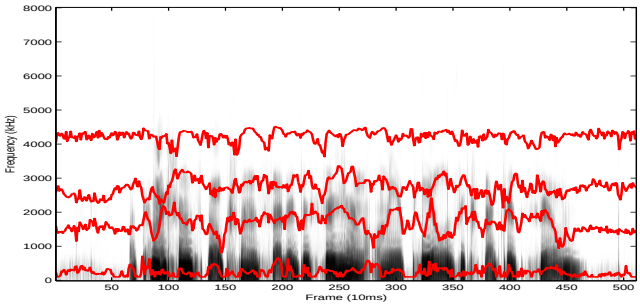

Fig. 3. Extracted VTRs from clean speech captured by the bone sensor.

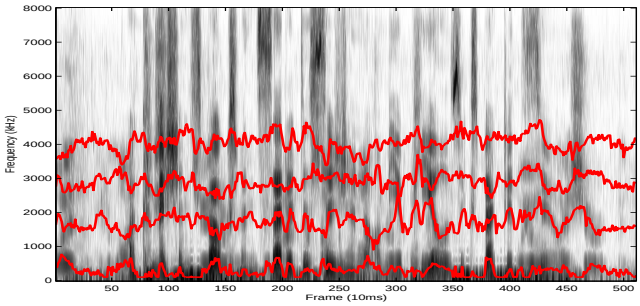

Fig. 4. Extracted VTRs (f_1 to f_4) superimposed on the spectrogram for noisy speech, where noisy speech is created by artificially adding interfering speech into the clean speech with SNR=5dB.

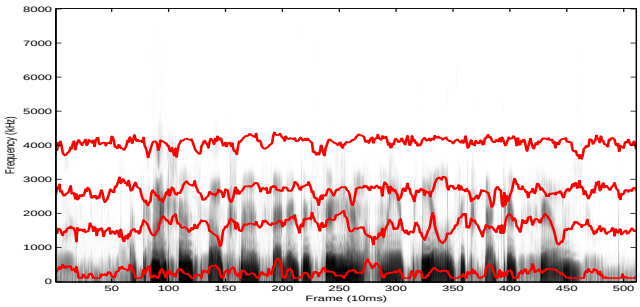

Fig. 5. Extracted VTRs from the noisy speech captured by the bone sensor.