LEAKAGE MODEL AND TEETH CLACK REMOVAL FOR AIR- AND BONE-CONDUCTIVE INTEGRATED MICROPHONES

Zicheng Liu, Amar Subramanya, Zhengyou Zhang, Jasha Droppo, Alex Acero

Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA

ABSTRACT

Continuing our previous work [1, 2] on using air- and bone-conductive integrated microphones, and in particular on using the *direct filtering* approach [3] for speech enhancement in noisy environments, we present in this paper a refined version of the direct filtering algorithm. The new algorithm takes into account explicitly the leakage of background noise into the bone channel. We also present a new algorithm that detects and removes an artifact known as teeth clacks. Experiments show that the addition of the above algorithms improves system performance to a large extent even in highly non-stationary noisy environments.

1. INTRODUCTION

Speech Enhancement is one of the oldest disciplines of signal processing. Though many techniques have been proposed to enhance speech in the presence of stationary background noise, enhancement in the presence of non-stationary background noise is still an open problem. In our previous work [1, 2], we introduced air- and bone-conductive integrated microphones and showed that such devices can be used to reliably determine whether the speaker is talking or not, and furthermore, the two channels can be combined to remove overlapping noises. We use "WITTY", which stands for "Who Is Talking To You", as our acronym for the air- and bone-conductive microphones. A prototype of such devices is shown in Figure 1. It contains two sensors: a regular close-talk (air) microphone and a bone-conductive microphone. The close-talk microphone captures wideband high-quality speech but is noise sensitive. The bone sensor has the interesting property that it is insensitive to background noise but only captures the low frequency portion of the speech signals. Furthermore, the captured speech signals are distorted. Our aim is to combine the bone signals with the close-talk signals to remove environment noise. For a detailed description of the Air- and bone-conductive integrated microphones (WITTY microphones), the reader is referred to [2].

Our previous work [1, 2] used a channel mapping technique for speech enhancement. It works by training a piecewise linear mapping from the bone signal to the close-talk signal. One drawback of this approach is that it requires training for each speaker. In [3], we introduced a new technique called *Direct Filtering* which does not require any training. The basic idea is to directly design a filter which performs distortion correction on the bone signal and optimally combines the bone signal and the close-talk signal to remove the background noise.

One of the primary reasons for using a bone sensor to aid the close-talking channel is that it is insensitive to background noise. However, in closed environments with large amounts of background noise, a significant amount of the noise leaks into the bone channel, thus negating the effect of the bone sensor. In this paper, we present a modified version of the Direct Filtering algorithm that takes into account explicitly the leakage of the background noise into the bone channel. We show by means of *frequency weighted segmental SNR* that the new algorithm has an improved performance over the old one.

We have also observed an artifact called *teeth clack*. While talking, unconsciously, the upper and lower jaws come in contact with each other resulting in a 'click' in the bone sensor. If not tackled appropriately, this will distort the mapping function between the air and bone channels, resulting a negative effect on the enhancement. In this paper, we also present an algorithm to detect and remove automatically teeth clacks.

Fig. 1. Witty Prototype.

2. RELATED WORK

Graciarena et al. [4] combined the standard and throat microphones in the noisy environment. They trained a mapping from the concatenated features of both microphone signals in a noisy environment to the clean speech. Compared to their system, our algorithm does not need any training, is not environment dependent and produces an audible speech signal so that the output can be used for perception as well as speech recognition. Strand et. al. [5] designed an ear plug to capture the vibrations in the ear canal, and used the signals for speech recognition with MLLR adaptation. Heracleous et. al. [6] used a stethoscope device to capture

the bone vibrations of the head and use that for non-audible murmur recognition. Like [5], they only used the bone signals for speech recognition with MLLR adaptation.

Interestingly enough, on September 8, 2004, a commercial headset product called *Jawbone* was released [7], which, as we do, also integrates a bone sensor with an air microphone in order to reduce background noise. According to the information from their website, it seems that the bone signals are only used in their device for speech activity detection, which helps build a better noise model than if speech activity is only detected from the air channel. Possibly, a technique similar to traditional spectral subtraction is then employed to enhance the speech signals. In our work, the bone signals are used not only for speech activity detection but also directly for speech enhancement in an integral way, as to be shown in this paper.

3. LEAKAGE MODEL FOR DIRECT FILTERING

In the direct filtering model proposed in [3] it was assumed that the noise, including background speech, does not leak into the bone sensor. However, in closed environments, such as an office, at low SNRs, a significant amount of the background noise leaks into the bone channel. This leaked noise is then passed onto the enhanced output, thus defeating the purpose of using the bone sensor. Hence we propose a new formulation that explicitly models this leakage.

Let $y(t)$ and $b(t)$ denote the close-talk and bone signals, respectively. Let $x(t)$ denote the clean speech which is to be estimated, $u(t)$ be the close-talking microphone sensor noise, $v(t)$ be the background noise and $w(t)$ the sensor noise associated with the bone microphone. The mathematical model for direct filtering with leakage may be represented as

$$y(t) = x(t) + v(t) + u(t) \tag{1}$$

$$b(t) = h(t) * x(t) + g(t) * v(t) + w(t) \tag{2}$$

where $h(t)$ is the speech mapping function from the clean signal in the close-talk channel to the signal in the bone channel, $g(t)$ is the noise leakage function that maps the background noise in the close-talking channel to the noise leaked in the bone channel, and $g(t) * v(t)$ models the amount of background noise that leaks into the bone sensor.

Equation (1) may be re-written in the complex-frequency domain as

$$Y_t(k) = X_t(k) + V_t(k) + U_t(k) \tag{3}$$

$$B_t(k) = H(k)X_t(k) + G_t(k)V_t(k) + W_t(k) \tag{4}$$

where k is the frequency band and $Y_t(k)$ is the k^{th} frequency component of $y_m[n] = y[n]w[m-n]$, a windowed version of $y[n]$ around time t. This notation applies to other quantities in the above equation. It is assumed here that the various frequency bands are independent, thus making the problem mathematically tractable. Also, it is assumed that $V_t(k) \sim N(0, \sigma_v^2(k))$, $W_t(k) \sim N(0, \sigma_w^2(k))$, $U_t(k) \sim N(0, \sigma_u^2(k))$ and are all independent of each other. It should be noted here that all the terms in equation (3) are in the complex frequency domain. In the following analysis we drop the argument (k) for simplicity.

We assume to be given an estimation of H_t, which is distributed as $N(H_0, \sigma_H^2)$, and an estimation of G_t, which is distributed as $N(G_0, \sigma_G^2)$. The clean speech signal X_t is given by minimizing the following probability:

$$\begin{aligned}
&p(X_t, V_t, H_t, G_t | Y_t, B_t, H_0, G_0, \sigma_u^2, \sigma_v^2, \sigma_w^2, \sigma_H^2, \sigma_G^2) \\
&\propto p(Y_t, B_t | X_t, V_t, H_t, G_t, \sigma_u^2, \sigma_w^2) \\
&\quad p(H|H_0, \sigma_H^2) p(G|G_0, \sigma_G^2) p(V_t) p(X_t) \\
&= p(Y_t | X_t, V_t, \sigma_u^2) p(B_t | X_t, V_t, H_t, G_t, \sigma_w^2) \\
&\quad p(H|H_0, \sigma_H^2) p(G|G_0, \sigma_G^2) p(V_t) p(X_t) \\
&\text{(if we ignore the prior speech model } p(X_t) \text{ by now)} \\
&\propto \frac{1}{(2\pi)^5 \sigma_u^2 \sigma_v^2 \sigma_w^2 \sigma_H^2 \sigma_G^2} \exp[-\frac{F_t}{2}]
\end{aligned}$$

where

$$\begin{aligned}
F_t = &\frac{|Y_t - X_t - V_t|^2}{\sigma_u^2} + \frac{|B_t - H_t X_t - G_t V_t|^2}{\sigma_w^2} + \frac{|V_t|^2}{\sigma_v^2} \\
&+ \frac{|H_t - H_0|^2}{\sigma_H^2} + \frac{|G_t - G_0|^2}{\sigma_G^2}
\end{aligned} \tag{5}$$

Setting $\partial F_t / \partial X_t = 0$, we get

$$V_t = \frac{\sigma_w^2 Y_t + \sigma_u^2 H_t^* B_t - (\sigma_w^2 + |H_t|^2 \sigma_u^2) X_t}{\sigma_w^2 + H_t^* G_t \sigma_u^2} \tag{6}$$

Setting $\partial F_t / \partial V_t = 0$, we get

$$\frac{G_t (B_t - H_t X_t - G_t V_t)^*}{\sigma_w^2} + \frac{(Y_t - X_t - V_t)^*}{\sigma_u^2} - \frac{V_t^*}{\sigma_v^2} = 0 \tag{7}$$

Substituting V_t above by (6), we get the solution for X_t as

$$X_t = \frac{(\sigma_w^2 + \sigma_u^2 H_t^* G_t) Y_t + [(\sigma_u^2 + \sigma_v^2) H_t^* - \sigma_v^2 G_t^*](B_t - G_t Y_t)}{\sigma_v^2 |H_t - G_t|^2 + \sigma_w^2 + \sigma_u^2 |H_t|^2} \tag{8}$$

The above equation is intuitive as $B_t - G_t Y_t$ removes the leakage in the bone sensor and the final output X_t is a weighted sum of the signal in the close talking microphone and the leakage removed bone sensor signal.

Setting $\partial F_t / \partial H_t = 0$, we get

$$H_t = \frac{\sigma_w^2 H_0 + \sigma_H^2 (B_t - G_t V_t) X_t^*}{\sigma_w^2 + \sigma_H^2 |X_t|^2} \tag{9}$$

We can obtain a similar equation for G_t by setting $\partial F_t / \partial G_t = 0$.

In order to estimate H_0 we adopt the same approach as in [3] by using the speech frames over the last N seconds (in our current implementation $N = 5$), resulting in

$$H_0 = \frac{\sum M_t \pm \sqrt{(\sum M_t)^2 + 4\sigma_v^2 \sigma_w^2 |\sum B_t^* Y_t|^2}}{2\sigma_v^2 \sum B_t^* Y_t} \tag{10}$$

where

$$M_t = (\sigma_v^2 |B_t|^2 - \sigma_w^2 |Y_t|^2) \tag{11}$$

A similar equation can be obtained for G_0, except the summation is over the non-speech frames.

3.1. Results

To test our techniques, we collected speech data from two speakers (one male and one female). Each read the first 42 sentences of the Wall Street Nov. 92 script, while another male speaker, sitting nearby, read newspaper. Note that the noise (background speech) is highly non-stationary.

To assess our algorithms, we use the *frequency weighted segmental SNR* measure [8]. Such a measure seeks to obtain a segmental SNR within a set of frequency bands normally spaced propotionally to the ear's critical bands. It is thus believed to produce an SNR measure more closely related to a listener's perceived notion of quality. Because we are dealing with real speech recordings, we do not know the signal and noise levels for each individual frame. Instead, for each sentence, we first perform speech detection as described in [1], and then compute the average energy from the speech frames as the signal energy and the average energy from the silent frames as the noise energy. Since the signal energy also contains the noise energy, the SNR is slightly over-estimated. The measure is given as follows:

$$\mathrm{SNR}_{\mathrm{fw}} = \frac{1}{M}\sum_{j=0}^{M-1}\left[\frac{1}{K}\sum_{k=0}^{K-1} 10\log_{10}\frac{E_{s,k,j}}{E_{n,k,j}}\right], \qquad (12)$$

where M is the number of utterances, K the number of frequency bands (we use the mel-frequency bands), $E_{s,k,j}$ the average short-term signal energy contained in the kth frequency band in the jth utterance, and $E_{n,k,j}$ the average short-term noise energy contained in the kth frequency band in the jth utterance.

Table 1. Comparative results of the enhancement algorithm with and without leakage modeling ($\mathrm{SNR}_{\mathrm{fw}}$ in dB)

	raw	without		with	
speaker	$\mathrm{SNR}_{\mathrm{fw}}$	$\mathrm{SNR}_{\mathrm{fw}}$	Δ_1	$\mathrm{SNR}_{\mathrm{fw}}$	Δ_2
speaker 1	23.9	32.3	8.4	35.7	3.4
speaker 2	15.2	25.8	10.6	28.5	2.7

The results are shown in Table 1. The second column shows the $\mathrm{SNR}_{\mathrm{fw}}$ of the raw data from the air microphone. The third and fifth columns show the $\mathrm{SNR}_{\mathrm{fw}}$ of the enhanced speech without and with leakage modeling, respectively. The fourth and sixth columns show the gains. The direct filtering algorithm without leakage modeling gains about 9.5 dB. By explicitly modeling the leakage, we gain another 3 dB.

4. TEETH CLACKS

While talking, unconsciously, the upper and lower jaws sometimes come in contact with each other resulting in a 'click' in the bone sensor which we refer to as *Teeth Clacks*. Teeth clacks are characterized by a high energy distribution in the medium and higher frequencies ($f \geq \frac{f_s}{4}$, f_s = 16 KHz and is the sampling frequency). Figure 2 shows the spectrogram of the close-talking (upper channel) and bone (lower channel) channels for a typical teeth clack. As it can seen, teeth clacks are characterized by a spike in the bone channel which is absent in the close talking channel. Failure to detect and remove them causes an annoying click in the enhanced signal. In the next section we analyze the effect of teeth clacks on direct filtering.

4.1. Effect of Teeth Clacks on the estimation of H

The transfer function H is an optimal mapping (maximum likelihood sense) between the clean speech signal X and the bone signal B (energy in the low frequency and null in the high frequency). However, when teeth clacks occur, the estimated H is erroneous, which subsequently distorts the estimated clean speech signal.

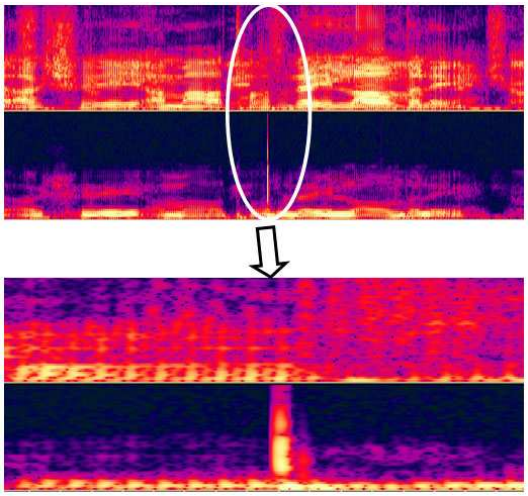

Fig. 2. Spectrogram of both close-talking and bone channel showing a teeth clack which is characterized by high energy in the medium and higher frequency bands in the bone sensor.

Refer to (10). In the absence of teeth clacks,

$$B_t(k) \approx 0 \text{ for all } k \geq \frac{f_s}{4} \qquad (13)$$

When teeth clacks occur the magnitude of the higher order spectral terms in the bone channel B are comparable to the magnitude of the terms in the close talking microphone channel Y. As a result, $\sigma_v^2|B_t|^2 - \sigma_w^2|Y_t|^2$ is increased. The increase in the numerator of equation 10 is much more than the increase in the denominator because $|B_t|^2 \geq B_t^*$, thus causing a spike in the higher frequency terms of H.

To understand the effect of this spurious spike in H on the estimation of clean speech, consider (8) with $\sigma_u^2 = 0$ (sensor noise in air channel is small), and ignore the effect of G_t since it is usually small (i.e., $G_t = 0$). This leads (8) to the following form

$$X_t \approx \frac{\sigma_w^2 Y_t + \sigma_v^2 H_t^* B_t}{\sigma_v^2|H_t|^2 + \sigma_w^2} \qquad (14)$$

As can be seen, X_t is essentially a weighted sum of the components of the close talking microphone signal and the bone signal. The weights are determined by H, σ_v, σ_w.

We need to consider two cases here.

4.1.1. *Processing a frame with* **no** *clacks.*

In this case the higher order terms of B_t are zero or close to zero. If H is as estimated above (with spikes in higher order terms), the terms that are affected by such a faulty computation are $\sigma_v^2 H_t^* B_t$ and $\sigma_v^2|H_t|^2$. In the high frequency bands where H_t spikes, $|B_t| \ll |H_t|$, which yields

$$\sigma_v^2 H_t^* B_t \ll \sigma_v^2|H_t|^2 \qquad (15)$$

As a result the denominator in (14) is much greater than the numerator, thus driving the estimated clean speech signal to **zero**.

4.1.2. Processing a frame **with** *clacks.*

Following a similar analysis as before, we need to compare the terms B_t and H_t. Since teeth clacks are present, the magnitudes of each of the terms are comparable. Thus clacks are simply passed through and in some cases may even be amplified if $|B_t| \gg |H_t|$.

4.2. Detection of Teeth Clacks

As explained in the previous section, teeth clacks are prominent in the bone channel and not present in the close talking microphone channel. This observation can be used to build a classifier to detect teeth clacks. In particular we make use of

$$\mathcal{J} = \sum_{k=1}^{K} \frac{1}{P(k)} |B_t(k) - H(k)Y_t(k)|^2 \tag{16}$$

$$\text{with } P(k) = \sigma_w(k)^2 + \sigma_v(k)^2 |H(k)|^2 + \sigma_H(k)^2 |Y_t(k)|^2 \tag{17}$$

as the discriminant function. In the above equation, K is the number of frequency bins/components and σ_H^2 is the variance of H. In essence the function H tries to match the the close talking signal with the bone signal. When a teeth clack occurs there is a mismatch between the two channels, resulting in a large value of $\mathcal{J}$; otherwise, the value of $\mathcal{J}$ should be small.

A close examination of $\mathcal{J}$ leads to the conclusion that it is the *Mahalanobis Distance* between $B_t(k)$ and $H(k)Y_t(k)$. Therefore, $\mathcal{J}$ follows a chi-squared distribution with K degrees of freedom.

The above distribution can be used in a significance (hypothesis) testing framework to automatically select the threshold for $\mathcal{J}$. If $\mathcal{H}_0$ is the hypothesis that a non-clack frame is classified as a non-clack frame, then we need to select the threshold α such that

$$P(\mathcal{J} < \epsilon | \mathcal{H}_0) = \alpha \tag{18}$$

In our algorithm we set $\alpha = 0.99$. The value of ϵ can be obtained from the χ^2 distribution table, and in our case $\epsilon = 365.4$.

4.3. Results

Figures 3 and 4 show the enhanced output for the same utterance without and with teeth clack removal, respectively. As it can be seen in Figure 3, when teeth clacks are not removed, a faulty transfer function H results in a number of nulls in the medium and high frequency bands. As a result the output is muffled. However, in Figure 4 the entire range of the spectrum of the output is retained.

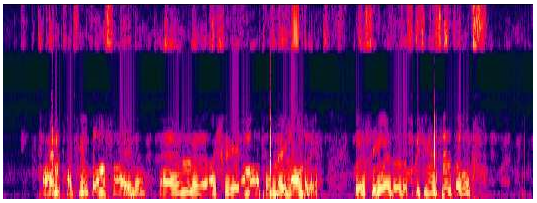

Fig. 3. Spectrogram of the enhanced signal **without** teeth clack detection and removal.

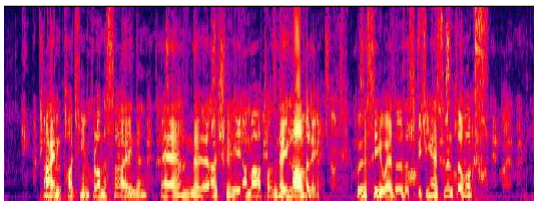

Fig. 4. Spectrogram of the enhanced signal **with** teeth clack detection and removal.

5. CONCLUSIONS AND FUTURE WORK

We have presented two new techniques: leakage model and teeth clack removal, to improve the direct filtering technique for the speech enhancement of the air- and bone-conductive integrated microphones. The leakage model is a mathematical framework to extend the direct filtering to the cases where the bone sensor has leakage. Wtih the teeth clack removal technique, the direct filtering algorithm can effectively handle the cases where there are a lot of teeth clacks. We have shown that the improved system is able to effectively enhance speech signals even in highly non-stationary noisy environment.

6. REFERENCES

[1] Y. Zheng, Z. Liu, Z. Zhang, M. Sinclair, J. Droppo, L. Deng, A. Acero, and X. D. Huang, "Air- and bone-conductive integrated microphones for robust speech detection and enhancement," in *ASRU 2003, St. Thomas, U. S. Virgin Islands*, 2003.

[2] Z. Zhang, Z. Liu, M. Sinclair, A. Acero, L. Deng, J. Droppo, X. D. Huang, and Y. Zheng, "Multi-sensory microphones for robust speech detection, enhancement, and recognition," in *ICASSP, Montreal, Canada*, 2004.

[3] Z. Liu, Z. Zhang, A. Acero, J. Droppo, and X. D. Huang, "Direct filtering for air- and bone-conductive microphones," in *IEEE International Workshop on Multimedia Signal Processing (MMSP), Siena, Italy*, 2004.

[4] M. Graciarena, H. Franco, K. Sonmez, and H. Bratt, "Combining standard and throat microphones for robust speech recognition," in *IEEE Signal Processing Letters*, 2003, vol. 10, pp. 72–74.

[5] O. M. Strand, T. Holter, A. Egeberg, and S. Stensby, "On the feasibility of asr in extreme noise using the parat earplug communication terminal," in *ASRU 2003, St. Thomas, U.S. Virgin Islands*, 2003.

[6] P. Heracleous, Y. Nakajima, A. Lee, H. Saruwatari, and K. Shikano, "Accurate hidden markov models for non-audible murmur (nam) recognition based on iterative supervised adaptation," in *ASRU, St. Thomas, U.S. Virgin Islands*, 2003.

[7] Jawbone, "http://jawbone.com/," Sept. 2004.

[8] J.M. Tribolet, P. Noll, and B.J. McDermott et al., "A study of complexity and quality of speech waveform coders," in *ICASSP, Tulsa, Okala*, 1978, pp. 586–590.