IMPROVED SPEECH RECOGNITION IN NOISY ENVIRONMENTS BY USING A THROAT MICROPHONE FOR ACCURATE VOICING DETECTION

Tomas Dekens [1], Werner Verhelst [1], François Capman [2], Frédéric Beaugendre [3]

[1] Interdisciplinary Institute for Broadband Technology – IBBT, Vrije Universiteit Brussel,
dept. ETRO-DSSP, Pleinlaan 2, B-1050 Brussels, Belgium
tdekens@etro.vub.ac.be, wverhels@etro.vub.ac.be
[2] Thales Communications, Multimedia Processing, 146, Bd de Valmy, BP 82, 92704 Colombes, France
francois.capman@fr.thalesgroup.com
[3] Voice Insight, bat. EEBIC, avenue J. Wybran 40, 1070 Brussels, Belgium
frederic.beaugendre@gmail.com

ABSTRACT

In this paper we show that microphones that capture the bone-conducted voice can be used to improve Automatic Speech Recognition in noisy environments. These microphones exhibit good noise rejection properties and their signals are therefore good indications of speech activity, even in very noisy conditions. We conducted experiments where we used a throat microphone signal as a Voice Activity Detection (VAD) input signal and found that recognition accuracies in non-stationary noise improve significantly compared to when VAD is executed on a conventional microphone signal.

1. INTRODUCTION

Modern day speech recognizers achieve very high recognition rates. Even very large vocabulary continuous speech recognition is possible. One of the biggest problems that remain in the Automatic Speech Recognition (ASR) domain is noise robustness. Unwanted background signals that corrupt the desired speech signal cause a mismatch between this speech signal and the training data of the acoustic models of the recognizer. This leads to a degraded recognition performance. A second consequence of the added background signals is that it becomes more difficult to tell at what times exactly the user is speaking. This so called Voice Activity Detection is a very important aspect of ASR; it tells the recognizer when it has to listen to the input signal. If the recognizer listens while the user is silent, this can cause a high amount of insertion errors; not listening while the user is speaking will certainly lead to deletion errors. The VAD becomes particularly challenging when the background noises are non-stationary. In this paper we use special microphones to capture the body-conducted voice signal to help reduce the influence of background noise sources on the accuracy of a VAD.

In section 2 of this paper we give a brief description of the properties of non-conventional microphones and explain why they are of particular interest when it comes to Voice Activity Detection. We also explain the VAD algorithm that was used in our experiments. In section 3 of this paper we show the results of the Automatic Speech Recognition experiments that were conducted. Finally, in the last section we draw some conclusions.

2. VAD WITH BODY-CONDUCTING MICROPHONES

2.1 Body-conducting microphones

Body-conducting microphones are sensors that rely on the fact that the human voice signal is not only transmitted through the air. When a person speaks this will also lead to vibrations of the bones and tissues of this person, and the speech signal will propagate through these media as well. Because interfering signal sources are typically not in contact with a person, they will be present in the air surrounding this person but not in his or her internal structure. This means that if we would capture this so called body-conducted signal (or bone signal), we can capture an uncorrupted speech signal. A downside of this, however, is that the signal can be very limited in bandwidth. Since the signal propagates through human bone and tissue, its high frequencies will be attenuated and a low-pass speech signal will be retained. This property makes the body-conducted signal as such unsuitable for speech applications where a broadband signal is required. Today's speech recognizers are such applications. The acoustic models they use are those of "broadband" speech; using narrowband speech at the input side would cause discrepancies with the training data, leading to recognition errors. In [1] it was found that the body-conducted signal captured at the upper lip resembles the air signal the most and that using this signal as an ASR input signal is not impossible. A better solution to this problem would be to train a recognizer using a very large database consisting of body-conducted data. If their bandwidth is large enough to contain some formants, these signals will contain sufficient information for speech recognition purposes. However, no such database exists at this time. A different approach is to transform the bone signal into a regular air signal and use this signal as input. This transformed signal would also be suited for human communication and would be a solution to the problem of human communication in very noisy environments. Multiple attempts have been made to transform a body-conducted signal into a clean

 ISSN 2076-1465

air signal, either by using only the body-conducted signal e.g.[2][3], or by accompanying this signal with the noisy broadband signal picked up by a regular microphone [4]-[8]. In [9] the body-conducting microphone was used to help retrieve the clean speech recognition feature vector.

VAD on the other hand is an application that does not necessarily require a broadband signal. VAD is very susceptible to noise, especially if this noise is non-stationary. The bone signal should typically contain a negligible amount of background signals, which should significantly simplify the VAD task as every signal activity would correspond to speaker activity. One problem however is that, since the bone signal is a low-pass signal, sounds that are characterized by high frequency energy, such as fricatives, will hardly be found at all in the bone signal. Detecting these sounds accurately using only the bone signal might be impossible, but the presence of speech could nevertheless be inferred from the bone signal activity as every syllable will include at least one voiced phoneme (the vowel nucleus). The multisensory speech enhancement techniques described in [4], [6]-[8] do implement VAD based on the body-conducted signal, but do not report on its direct impact on ASR performance. [10] on the other hand does report a relative improvement in error rate of 0%, 61.1% and 45.4% at 40dBA, 84dBA and 96dBA noise levels respectively, when the close talk microphone signal is replaced by a throat microphone signal for VAD. In [5] a reduction of 90% in the number of insertion errors is achieved when the signal is amplitude modulated by the speech confidence, which is derived from the bone signal's energy level.

2.2 Energy based VAD

We considered that analysing the energy of the bone-conducted signal should suffice for the purpose of Voice Activity Detection as this signal shows a high SNR such that high energy content becomes a direct indication of speaker activity. Thus, to detect the active speech parts in the body-conducted signal, we developed a simple energy based voice activity detector, that we named eVAD. The feature used in the eVAD algorithm is the smoothed energy, contained in the frequency region of interest. Let $Y(m,k)$ be the STFT of the input signal $y(t)$, with m the frame number and k the frequency index. The smoothed energy is then calculated as:

$$E(m) = mean\left\{ \frac{2}{Nfft} \sum_{k=k_1}^{k_2} \left|Y'(m+j,k)\right|^2 \right\}_{j=-N}^{j=+N} \qquad (1)$$

$$0 \le k_1, k_2 \le Nfft/2$$

where $Y'(m,k)$ is the same as $Y(m,k)$, except when k corresponds to DC or half the sampling frequency, then $Y'(m,k)$ is $Y(m,k)/\sqrt{2}$. This ensures that the energy at these frequency bins is only considered once. One can see that the parameter N in (1) controls the extent to which the feature is smoothed in time and by adjusting the range $[k_1, k_2]$ a given frequency band can be selected.

During an initialization phase the first frames of the input signal are used to calculate the noise energy using (1); the mean of these noise frame energies is an initial estimation of the smoothed noise energy E_{Noise}. Next, the smoothed energy is calculated for each signal input frame. This energy is then divided by E_{Noise} and the logarithm is taken:

$$Eratio(m) = 10\log_{10} \frac{E(m)}{E_{Noise}(m)} \qquad (2)$$

This ratio is then compared to a threshold. When the ratio is smaller than this threshold the frame is considered to contain only noise and the noise energy is updated:

$$E_{Noise}(m+1) = \alpha E_{Noise}(m) + (1-\alpha)E(m) \qquad (3)$$

with $0 \le \alpha \le 1$. If the ratio is larger than the threshold, speech is detected and the current noise energy estimate will remain unchanged:

$$E_{Noise}(m+1) = E_{Noise}(m) \qquad (4)$$

As stated in the previous section, unvoiced phonemes such as e.g. /s/ or /t/ might not be captured by a body-conducting microphone. This leads to a low instantaneous energy part in the signal while there is active speech. If these phonemes occur in the middle of a speech fragment, this does not pose a big problem in practice. First, the surrounding voiced phonemes ensure that the smoothed energy feature curve can not drop to too low values. Furthermore, one can select a minimum duration that a detected pause should have before it is classified as a pause. This can filter out pauses that would be detected due to a short unwanted drop in the energy curve.

On the other hand, very often a speech fragment starts or ends with problematic unvoiced phonemes and this can not be dealt with by selecting a minimum pause length. That is why detected speech portions are extended in time (front and back) by a given small amount, i.e. the regions before and after a detected speech fragment will be classified as speech regions as well.

Signals captured by a body-conducting microphone can also contain undesired sounds such as teeth clacks or swallowing sounds. These sounds cause a rise in the energy curve, but generally have a short duration, especially compared to speech. To cope with these short bursts of high energy a minimum speech length can be selected. If a speech region is detected that is shorter than this minimum length, it will be classified as noise.

Besides the relative energy ratio (2), we found it useful to also use an absolute power measure. This will make the VAD deaf to signals whose power is below a threshold value. So if

$$\frac{2}{Nfft\ Nwin} \sum_{k=k_1}^{k_2} \left|Y'(m,k)\right|^2 < \delta \qquad (5)$$

frame m will be classified as a noise fame, where $Nwin$ is the length of the window used.

3. ASR EXPERIMENTS

Our goal is to determine the influence of VAD with body-conducting microphones on automatic speech recognition performances. During previous work we recorded a multilingual database of noisy speech using multiple microphones [11]. It was noticed that the throat microphone signals in this database exhibited good noise rejection properties and there-

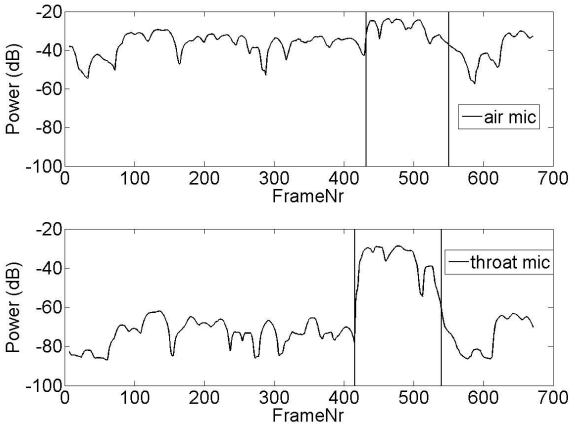

Figure 1- Signal power, top: air signal, bottom: throat signal

fore we selected this throat microphone (a Clearercom Stryker PC [12]) for our experiments. The air signal we used in our experiments was produced by a Bluetooth close-talk microphone, designed by Voice Insight [13].

The signals of the throat microphone contain some low frequency noise but no high frequency voice energy. For this reason, the energy in the region [250 5000] Hz was used to calculate the energy ratios used in the eVAD.

Figure 1 depicts an example of the power (in the selected frequency region) of the throat microphone signal and that of an air microphone. The background noise source was a second speaker. The vertical lines indicate the speech start and stop positions. It can be clearly seen that the throat signal exhibits a much higher SNR than the air signal, making it well suited for energy based VAD indeed.

The multilingual database contains noise corrupted utterances for two speech recognition tasks and some numbers. For the experiments in this paper, we chose to let the speech recognizer recognize isolated numbers, since this is a more generic task from which we expect the most clear results as the other two tasks come with a very restrictive grammar.

We picked four noise types for these experiments: diesel and jet engine noise, since these are stationary noise types; a siren, as this is a noise type that is stationary on a short time basis, but globally alternates periodically between two sounds; babble noise, since this is a speech-like noise and more or less stationary. The last noise type that was selected was an interfering speaker, which is the most challenging background sound that exists for a VAD or speech recognizer, since it is non-stationary and its characteristics are identical to those of the desired signal.

The loudness levels of the noise at the ear of the speaker during the recordings of the database were set to 80dBA, except for the levels of babble noise and of the interfering speaker, which were chosen in a way that they sounded realistic. The number of utterances (utterance being a number in the range [0-9999]) per noise type was equal to 25 for each speaker.

The ASR engine that was used was the Nuance Vocon 3200 [14]. In one set-up we just fed the noisy signals to the

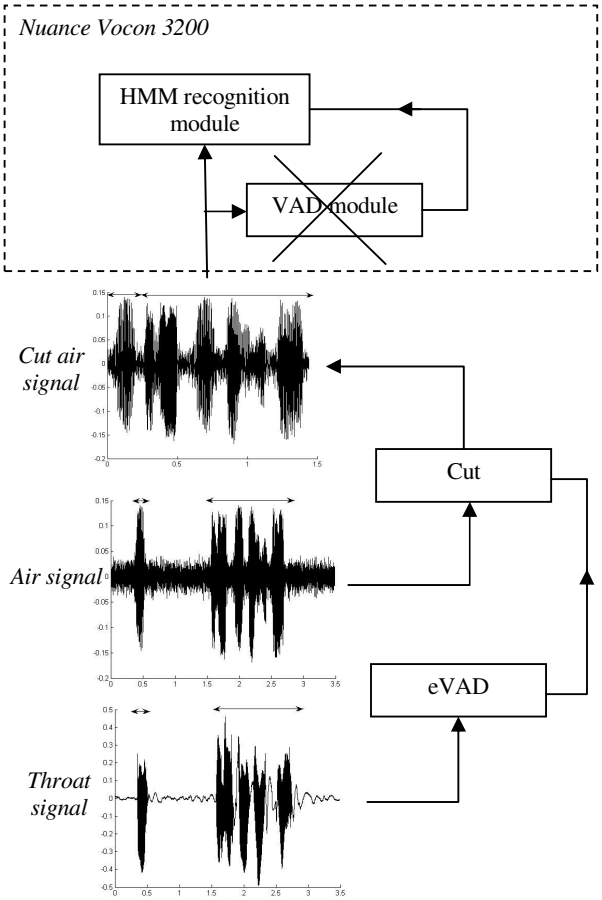

Figure 2-Schematic representation of experimental set-up

speech recognizer. This means that the internal VAD of the recognizer is assigned with the job of determining when the speaker is actually talking using only the air microphone signal as input. This VAD analyses the signal by looking at the energy levels, duration and frequency content of events to decide whether the event corresponds to a talking user. All parameters of the internal VAD where set to their default values as these should provide the best general performance. In a second set-up we used the synchronized throat microphone signal to determine with the help of our energy based VAD the speech start and stop positions. For this we used 32ms long Hamming windows with 50% overlap. The number of FFT points was 512. Equation 1 was used to calculate the energy, using a smoothing parameter $N = 6$ and a range $[k_1, k_2]$ corresponding to the frequency band [250 5000] Hz. The parts where speech was detected in each of the recordings were cut out from the corresponding air channel signal and consolidated into separate sound files. Note that in our isolated numbers task the VAD should ideally detect only one speech part in each recording: the one corresponding to the uttered number. The sound files that were obtained in this way were then used as input files for the speech recognition system. Since the recognizer uses the first 100ms of the input

signal to execute noise estimation, we ensured that the sound files started with 100ms of background noise, in order to not disturb this noise estimation process. The internal VAD of the recognizer was switched off so as to make sure that all parts of these signals would be analysed by the recognizer itself. Figure 2 gives a schematic representation of this second set-up.

The Nuance Vocon 3200 requires a Backus-Naur Form grammar in order to work properly. We provided the system with a grammar that contained all numbers from 0 to 9999. This means that during our tests the recognizer could recognize 10000 different numbers.

Table 1 compares the word accuracies that were attained from the Nuance Vocon 3200 when the internal VAD was used with the situation in which the throat microphone signal was used as input signal for eVAD voice activity detection. Also an estimation of the signal to noise ratio is given for each noise type and speaker.

The Word Accuracy (WA) is defined as:

$$WA = \frac{H - I}{N} \qquad (6)$$

where H is the number of correctly recognized words, which is equal to the total number N of words to be recognized minus the sum of the number of deletions and substitutions; I represents the number of insertions. Note, however, that since in our case only one word needs to be recognized, no deletions or insertions are possible.

First of all, it can be clearly seen that the use of the throat based eVAD leads to important WA improvements in general. It can also be seen that when stationary noise such as jet or diesel engine noise is present, relatively smaller improvements in Word Accuracy are obtained compared to the other noise types. This is due to the fact that a Voice Activity Detector is more robust against stationary noises, and so is a speech recognizer. The fact that the recognizer's WA for the French speaker is inferior to that of the German speaker could be explained by the fact that this French speaker was less influenced by the Lombard effect; the speaker kept the loudness of his voice at a relatively low level, even at this very high noise level of 80 dBA. This resulted in air signals with very low SNR, which makes the recognition task more difficult. While the German speaker did produce Lombard speech, it seems the mismatch with normal speech was not big enough to affect the recognition performance much.

When a less stationary or more speech-like noise corrupts the speech signal, such as a siren or babble noise, absolute word accuracy improvements in the range of 20-30% are obtained. When the speech signal of the user is accompanied by the speech signal of a second, undesired, speaker, the most remarkable results can be noticed. In this case accurate air-based VAD becomes impossible, especially when there is no big difference in energy levels between the two speech signals, considering that the other properties of the two signals are comparable. In addition, when a signal fragment that contains only speech produced by the interfering speaker is sent to the recognizer, the recognizer will conclude that the signal contains speech and will make an attempt to recognize this speech, which will lead to incorrect transcriptions. Table 1 shows that these effects cause a Word Accuracy as low as 0%. Since the speech recognition task consisted of recognizing one word per run, no deletions or insertions were possible and a WA of 0% means that not a single word was recognized correctly. Using the throat microphone based eVAD for telling the speech recognizer where the useful speech can be found, pushed the recognition accuracy up to 80%.

Table 1 ASR results

Language	Noise type	SNR (dB)	Word accuracy (%)	
			Internal VAD	Throat eVAD
German	Jet engine	10.66	96	96
German	Diesel engine	8.88	92	96
German	Siren	10.75	60	88
French	Jet engine	1.00	48	64
French	Diesel engine	4.50	72	76
French	Babble noise	1.75	56	80
French	Interfering speaker	5.00	0	80

4. CONCLUSION

In this paper we used a throat microphone to improve the accuracy of an off-the-shelve state-of-the-art speech recognizer. To this end, the recognizer's internal VAD was deactivated and our eVAD algorithm was used on the throat microphone signal to eliminate the non-speech segments from the recognizer's close talk microphone input (100ms of leading background noise was left for the recognizer's noise suppression algorithm). Our experiments showed a systematic improvement in word accuracy and with non-stationary noise sources remarkably higher word accuracies were achieved, especially when the noise consisted of competing speech.

5. ACKNOWLEDGMENT

Part of this work was performed in the context of the EU-FP6 project SAFIR (IST-2002-507427).

REFERENCES

[1] Ishmitsu S, Kitikaze H, Tsuchibushi Y, Yanagawa H, Fukushima M, "A noise-robust speech recognition system making use of body-conducted signals" , Acoust. Sci. & Tech., vol.25, no.2, pp 166-169, 2004.

[2] V.T. Thang, S. Germine, M. Unoki, M. Akagi, “Method of LP-based blind restoration for improving intelligibility of bone-conducted speech”, Interspeech 2007, Antwerp, Belgium, Aug. 2007.
[3] T. Shimamura, J. Mamiya and T. Tamiya, “Improving bone-conducted speech quality via neural network”, IEEE International Symposium on Signal Processing and Information Technology, Aug. 2006.
[4] Z. Liu, Z. Zhang, A. Acero, J. Droppo, and X. Huang, “Direct filtering for air-and bone-conductive microphones,” Proc. MMSP, pp. 363–366, Sept. 2004.
[5] 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. ASRU, pp. 249–254, Dec. 2003.
[6] J. Hershey, T. Kristjansson and Z. Zhang, “Model-Based Fusion of Bone and Air Sensors for Speech Enhancement and Robust Speech Recognition", Proc. ISCA Workshop on Statistical and Perceptual Audio Processing (SAPA2004), Jeju, Korea, Oct. 3, 2004.
[7] A. Subramanya, Z. Zhang, Z. Liu, J. Droppo and A. Acero, “A graphical model for multi-sensory speech processing in air-and-bone conductive microphones”, Proc. Eurospeech, pp. 2361-2364, Sept. 2005.
[8] A. Subramanya, Z. Zhang, Z. Liu and A. Acero, “Speech Modeling with Magnitude-Normalized Complex Spectra and its Application to Multisensory Speech Enhancement", IEEE International Conference on Multimedia & Expo (ICME), Toronto, Canada, July 9-12, 2006.
[9] M. Graciarena, H. Franco, K. Sonmez and H. Bratt, “Combining standard and throat microphones for robust speech recognition,” IEEE Signal Processing Letters, vol. 10, no. 3, pp. 72–74, March 2003.
[10] S. Dupont and C. Ris, “Combined use of close-talk and throat microphones for improved speech recognition under non-stationary background noise”, In Proc. of Robust 2004 (Workshop (ITRW) on Robustness Issues in Conversational Interaction), Norwich, UK, August 2004.
[11] T. Dekens, G. Patsis, W. Verhelst, F. Beaugendre and F. Capman "A Multi-sensor Speech Database with Applications towards Robust Speech Processing in hostile Environments," The International Conference on Language Resources and Evaluation (LREC), Marrakech, Morocco, May 28-30, 2008.
[12] http://www.clearercom.com/pc_throat_mic.htm
[13] http://www.voice-insight.com/
[14] http://www.nuance.com/vocon/3200/