Combined use of close-talk and throat microphones for improved speech recognition under non-stationary background noise.

Stéphane Dupont, Christophe Ris and Damien Bachelart

Multitel & FPMS-TCTS, Avenue Copernic 1, B-7000 Mons, Belgium
dupont,ris@multitel.be

Abstract

This paper intends to summarize recent developments and experimental results related to Automatic Speech Recognition (ASR) using signals captured with a throat-microphone. Due to the proximity of the sensor to the voice source, the signal is naturally less subject to background noise. This however yields speech sounds that have different frequency contents than with traditional microphones, and requires having specific acoustic models. We propose to use the information from both signals by combining the probability vectors provided by both acoustic models.

The systems are evaluated on a connected digit recognition task in French. A database has been recorded for both training the acoustic models and for testing the whole setup. It contains both throat and "ordinary" close-talk signals. To avoid any possibly unrealistic assumption on the effect of noise on each signal, the test portion has been acquired using a background noise played back through loudspeakers.

The ASR experiments that we achieved demonstrate the benefit of using alternative microphones. Relative recognition improvements as high as 80% were obtained on sequences of digits recorded in loud musical environment.

1. Introduction

Robustness to environmental noise or channel distortions is currently one of the main challenges in Automatic Speech Recognition (ASR), especially, when more and more commercial applications are emerging, confronting the technology to practical usage conditions. Currently, the most popular techniques deal with the noise degradations by trying to remove their effect in the power-spectral or feature domain. Other techniques try to adapt the acoustic models used in ASR to the particular noisy environmental context to deal with. Microphone arrays are also being studied, where a clean speech signal is infered from noisy signals captured by several microphones.

As an orthogonal approach to these, it also makes sense to work on getting the cleanest signal right from the acquisition system itself. Depending on the application that is targeted, directional microphones or close-talk microphones can be used. For the applications where carrying a device is accepted, one can also think about using signals captured by *non-conventional* microphone technologies. Several specific sensors technologies (like bone-, myographic- and throat-microphones) can provide the benefit of even stronger immunity to ambient "noise" sources.

In this paper, we introduce some recent work with a throat-microphone. The captured signal is less subject to noise due to its proximity with the source of useful speech signal. The transfer of different speech sounds is however different that from the mouth to a microphone. This, combined with variations implied by the positioning of the sensor, yields a baseline performance that is significantly worse than with a traditional close-talk microphone. The method that is proposed here takes advantage of the use of both a close-talk and a throat microphone.

In section 2, we provide some references on using "non-conventional" sensors for improving speech recognition in noise. We also introduce our work with the throat-microphone, the database that has been recorded, a few qualitative considerations as well as an assessment of the noise immunity property of this sensor. In section 3, a description of the feature extraction and acoustic modeling systems that have been used is presented. Also, the different variants for combining both sources of information are described. In section 4, we compare both the close-talk microphone and throat-microphone alone, as well as a in combination. In section 5, we discuss the results and propose some ideas for future improvements.

2. Alternative acoustic sensors

We recently initiated a study of alternative audio acquisition devices as an approach to the noise robustness problem in ASR.

The set of *alternative acoustic sensors* we address in this paper, can be referred to as the class of signal acquisition devices that attempt to capture the voice signal by other means than air conduction.

Few previous publications have been proposed that try to make use of noise-immune sensors attached to the body for improving ASR performance. In [1], a so-called Non-Audible Murmur (NAM) microphone is used just by itself, as an alternative to an ordinary microphone. The NAM device is based on a stethoscopic system and is attached to the skin of the user, for instance behind the ear. The work focus on one major issues in using alternative devices: the need to estimate an HMM model that is valid for the new signals. The paper proposes to use an adaptation procedure based on Maximum Likelihood Linear Regression (MLLR). Although noise robustness is presented as a potential advantage of the NAM device, reported results do not show any benefit yet.

MLLR adaptation is also used in [2]. Here, the sensor is an inner microphone that captures the speech signal behind an acoustic seal (or earplug) in the auditory canal. The ASR performance and robustness based on the inner microphone and a close-talk microphone are compared. It is shown on data recorded in real noisy environments that the inner microphone allows to get the same performance as the standard microphone at 15 to 20 dB higher noise level.

In [3], the combined use of a standard and a throat microphone is investigated. The combination technique is based on piecewise-linear transformation of a feature space to another. The source feature space is defined by a feature vector that con-

tains both standard and throat microphone features in noisy conditions and the target space is defined by the clean speech features. At a Signal-to-Noise Ratio (SNR) of 6 dB, the word error rate reduction was close to 30%.

In [4], the sensor is a bone-microphone coupled with a regular close-talk microphone. The device has the look and feel of a regular headset, but a bone-microphone component is applied just above the zygomatic bone. Here, the proposed technique also consists in estimating the clean speech features from the features computed from both microphones coupled with a SPLICE-like mapping [5] from the bone-microphone features to the clean speech features. Current results are based on a mean opinion score (MOS) and show some benefit at 0 to 10 dB SNR.

These papers illustrate the variety of devices that can be used as a replacement or a complement to regular microphones. In addition to conduction through the air, speech is also naturally conducted within the body tissues and up to the skin, bones and ear canal. This property can be used to obtain alternative signals. Sensors that are not relying on acoustic signals could also be considered. For instance, radar-like technology applied around the vocal tract [6], and electromyography applied on the muscles of the jaws [7] have also been proposed in the literature for improved speech enhancement or recognition.

The reported results illustrate that applications using ASR can probably benefit from these alternative devices, mostly in terms of noise robustness. It is also important to notice that the techniques proposed allow to handle non-stationary noises (and even background speech [4]). Current speech enhancement or adaptation techniques still have a hard time to deal with these.

However, other issues naturally appear and are discussed in some of these papers. First, if these devices are less subject to external noises, as well as breath noise, they become however more sensitive to body-internal noises, like swallowing or whispering for instance. Then, some of them pose the problem of practical attachment, comfort and ergonomics. Besides, of same crucial importance, few data is available for developing the acoustic models for these sensors, as well as to design the algorithms that combine both acoustic streams.

Our paper presents preliminary experiments conducted on the exploitation of the *throat* microphone in a digit recognition task. The throat microphone can be seen as a transducer of the vibrations of the body tissues (skin, bone, cartilage, ...) produced when one is speaking. Placed directly on the skin closed to the Adam's apple (see Fig. 1), the device captures the vibrations of the vocal cords transmitted through the throat.

Most of the papers discussed above rely on adaptation techniques on a few recordings. In our work, we have focused on the combination of independent acoustic models, for which a complete speech database was recorded. For pratical reasons, we choose a connected digit recognition task.

Figure 1: *How to wear the throat-microphone.*

2.1. Database

As a first step, a speech database was collected. It contains records of sequences of digits in French, with both the throat-microphone and a headset (air-conductive) microphone, in clean and noisy environments. The two signals were recorded simultaneously in a synchronized way. The database has been designed in the purpose of training acoustic models for ASR (sufficient amount of clean data) and testing in difficult noisy conditions (loud background music). So, 51 speakers have pronounced 5750 utterances (sequences of digits), among which 650 utterances were recorded in a noisy environment for testing purpose. Two noise conditions were created, music diffused at 84 dBA and 96 dBA (measured with a soundmeter). Corresponding average Signal-to-Noise Ratios (SNR) related to both the headset microphone and the throat microphone are presented in Table 1. Note that the headset microphone, as a close-talk directional microphone, is already inherently robust to the background noise.

Noise level	Headset mic.	Throat mic.
84 dBA	18 dB	34 dB
96 dBA	12 dB	25 dB

Table 1: *SNR for the two noise conditions and for the two microphones.*

The Lombard effect that we observed in the recordings may explain the reduced SNR gap as compared to the noise level gap between the two noise conditions. In our experiments, 4100 utterances were used to train the acoustic models of the speech recognition systems on both the "classical" speech signal and the throat signal. Test data consist in 1000 clean utterances, 500 noisy utterances (84 dBA) and 150 highly noisy utterances (96 dBA).

2.2. A few qualitative observations

As we might expect, the signal captured by the throat microphone suffers from some "distortions" compared to the air-conducted signal. First, as throat vibrations are mainly generated by the vibrations of the vocal folds, unvoiced sounds (fricative, unvoiced plosives, ...) are strongly attenuated while other sounds like nasals or voiced posives are relatively amplified. Similarly, other physiological sounds like swallowing or whispering are amplified while, on the other side, breathing sounds or whistling are just not present in the throat signal. The same way, we can, intuitively, expect that the frequency representation of the signal will be limited by the intrinsic elasticity properties of the human tissues. From a spectral analysis of audio recordings, we have observed that the propagation channel (throat, skin, ...) acts as a low-pass filter with a 3 kHz cutoff frequency.

2.3. Noise robustness

We have conducted some experiments in order to assess the claimed robustness of the throat microphone to environmental noise. In that purpose, we have recorded, simultaneously with the headset mic and the throat mic, different noises, played back with increasing intensity levels (from 45 dBA to 105 dBA, step 5 dBA) [1]. Similarly, a few sentences have been recorded

---

[1] Sound intensity was measured with a soundmeter close to the experimenter. The "silence" intensity in the recording room was 40 dBA.

with the two microphones for normalization purpose. Figure 2 shows the evolution of the normalized noise level recorded on the headset microphone and the throat microphone as a function of the noise intensity played back in the room.
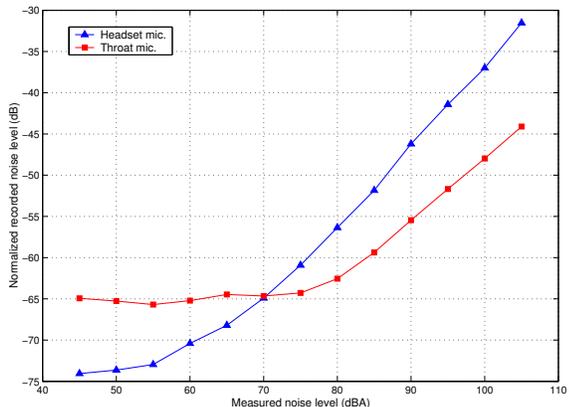

Figure 2: *Noise level of recordings (normalized) vs. measured noise level for both microphones.*

We observe that the estimated noise level is rather constant (-65 dB normalized) for the throat microphone up to a noise level of 75 dBA (which correspond to a very bustling street), which means that, up to this point, background noise is practically not captured by the microphone (at least at a level lower than the device intrinsic noise level). Beyond this point, the noise level estimated from the throat microphone is increasing, however, not as fast as the noise level estimated from the close-talk microphone. The noise robustness performance is therefore clearly demonstrated. The next sections will discuss the opportunity to use the throat microphone for ASR.

3. ASR systems

The ASR systems are based on the hybrid HMM/ANN architecture, which combines Artificial Neural Networks (ANN) and Hidden Markov Models (HMM) [8]. As a first step, we have trained two acoustic models (ANN), one for each microphone, on a subset of the recorded sentences (4100 utterances pronounced by 41 speakers). The ANN are trained as phoneme classifiers with the same set of phonetic units for the two systems (36 phonemes in French). Of course, as all the phonemes are not present in the ten digits, only a few of them are actually used.

The acoustic features were the J-RASTA PLP coefficients [9] with first and second derivatives. The J-RASTA processing provides some noise filtering. The goal being to start with a fairly good system (close-talk microphone, noise robust acoustic features), and investigate the benefit of using alternative microphones.

3.1. Combined use of throat and close-talk microphones

We first compare the performance of the two ASR systems independently. It is, indeed, a crucial point to assess the usability of the throat microphone for ASR purpose or the suitability of classical acoutic analysis and acoustic modeling for this particular signal.

From the qualitative observations of the throat signal, we can expect some degradation on clean speech due to the

poor frequency content while performance in noisy conditions should be better. These observations led us to consider the combination of the throat microphone with a close-talk microphone. We introduce, hereafter, the recombination strategies that we have investigated. They are all based on the re-estimation of the HMM state emission probabilities from the probabilities estimated by the two independent acoustic models (outputs of the ANNs) (see Fig. 3).
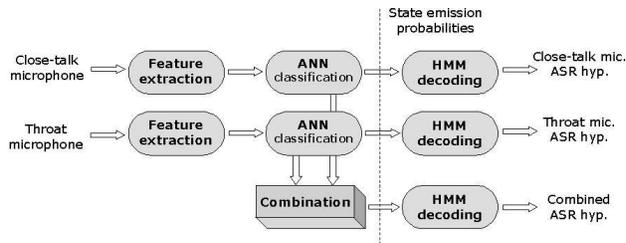

Figure 3: *HMM state emission probabilities combination.*

We investigated three combination paradigms:

1. As the environmental noise is almost not captured by the throat microphone, we suggest to use this signal for voice activity detection only. We know, indeed, that an accurate silence/speech labeling improves a lot the recognition performance. Practically, this comes to replace the silence state emission probability of the close-talk ASR system by the one of the throat ASR system. This approach is denoted *VAD combination* hereafter.

2. In this second approach, we consider that the two signals convey potentially useful information. In that sense, we combine the whole emission probability vectors, according to a weighted geometric average rule:

$$P(q_k|x_t)_{combi} = P(q_k|x_t)^{\alpha}_{close-talk} \cdot P(q_k|x_t)^{(1-\alpha)}_{throat}$$

Where:

$P(q_k|x_t)$ is the emission probability for state q_k given acoustic observation x_t at time t

α is the weighting factor.

This approach is denoted *Expert combination* hereafter.

3. We combine the first two methods, the emission probability vectors are combined according to the weighted geometric average rule, while the speech/silence decision is taken from the throat microphone only. Hence, we expect to take advantage of the high reliability of the silence probability coming from the throat microphone system. This approach is denoted *Hybrid combination* hereafter.

4. Experiments

4.1. Independent ASR systems

Table 2 shows the ASR error rates for both the close-talk and the throat microphone systems independently. The results confirm the expectation analysis, i.e. lower performance of the throat microphone for clean speech (denoted 40 dBA in the table, which corresponds to average background noise level in a very quiet room) and almost 50% relative improvement in noisy conditions. Moreover, it validates the suitability of the throat microphone signal for the ASR task.

Data	Headset mic.	Throat mic.
40 dBA	1.1%	4.3%
84 dBA	13.8%	7.6%
96 dBA	34.6%	18.4%

Table 2: *Error rate for both ASR systems performing independently.*

Performance of the headset microphone on noisy speech may seem quite poor (18 dB and 12 dB SNR, see Table 1) compared to what can be found in the literature (e.g. AURORA tasks [10]) but, note that, a) the noise is background music which is non-stationary. The J-RASTA processing, as well as speech enhancement techniques, may have little effect on that type of noise, b) the utterances have been recorded in real noisy environment (not artificially added), some side-effects such as the Lombard effect may have occured.

4.2. Combined ASR systems

Table 3 gives the error rate for the *VAD combination* method and the relative improvement compared to the close-talk microphone only. The first interesting observation is that there is no degradation of performance on clean speech. Moreover, we confirm the importance of an accurate speech/silence detection in the recognition process. The error rate is always better or similar than the error rate of one or the other sensor alone.

Data	Combined (VAD)	Relative imp.
40 dBA	1.1%	0.0%
84 dBA	5.5%	61.1%
96 dBA	18.9%	45.4%

Table 3: *Error rate for the VAD combination method and relative improvement regarding the close-talk microphone system.*

Table 4 gives the error rate for the *Expert combination* method, the relative improvement compared to the close-talk microphone only and the optimal value of the weighting factor α for each test condition. The *Expert combination* yields very good improvements in noisy conditions and keeps the good performance of the close-talk microphone on clean speech. A drawback of this approach is the rather important sensitivity of the performance to the weighting factor α, although the detailed results are not reported in this paper.

Data	Opt. α	Combined (Expert)	Relative imp.
40 dBA	0.7	0.9%	18.2%
84 dBA	0.3	3.8%	72.5%
96 dBA	0.3	12.9%	62.7%

Table 4: *Error rate for the Expert combination method and relative improvement regarding the close-talk microphone system.*

Table 5 gives the error rate for the *Hybrid combination* method, the relative improvement compared to the close-talk microphone only and the optimal value of the weighting factor α for each test condition. This approach still improves the recognition performance in noisy condition. An interesting observation of the results that we obtained is the much lower sensitivity of the performance to the weighting factor α. A value of $\alpha = 0.5$ (i.e. unweighted geometric average) yields almost optimal performance (1.2% / 2.7% / 12.8%).

Data	Opt. α	Combined (Hybrid)	Relative imp.
40 dBA	0.4	0.9%	18.2%
84 dBA	0.5	2.7%	80.4%
96 dBA	0.3	12.6%	63.6%

Table 5: *Error rate for the Hybrid combination method and relative improvement regarding the close-talk microphone system.*

5. Conclusions

Although the combination rules that have been applied are very simple, these preliminary results suggest the strong benefit of using "non-conventional" speech sensors. Adoption will however require to address other issues such as the relative large size and discomfort of wearing the device during long time, as well as the difficulty of properly positioning the microphone. Acceptance is hence more likely to start from professional market segments where speech input would be a major benefit, for instance due to hands-busy situations.

The problem of database acquisition for more complex tasks has to be discussed too. Indeed, each type of sensor will affect the speech signal differently and creation of new databases would therefore be necessary. This is why, in the future, adaptation techniques as well as prior knowledge on the nature of the throat-microphone signal will rather be used to develop the models. Similarly, other combination strategies should be considered in order to limit the necessary amount of new data.

6. References

[1] P. Heracleous, Y. Nakajima, A. Lee, H. Saruwatari, K. Shikano, "Accurate Hidden Markov Models for Non-Audible Murmur (NAM) Recognition Based on Iterative Supervised Adaptation", in Proc. of ASRU 2003, U.S. Virgin Islands, Dec. 2003.

[2] O.M. Strand, T. Holter, A. Egeberg, S. Stensby, "On the Feasibility of ASR in Extreme Noise Using the PARAT Earplug Communication Terminal", in Proc. of ASRU 2003, U.S. Virgin Islands, Dec. 2003.

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

[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", in Proc. of ASRU 2003, U.S. Virgin Islands, Dec. 2003.

[5] J. Droppo, L. Deng, A. Acero, "Evaluation of the SPLICE on the Aurora2 and 3 Tasks", in Proc. of ICSLP 2002, Denver, Colorado, Sep. 2002.

[6] L. Ng, G. Burnett, J. Holzrichter, T. Gable, "Denoising of Human Speech using Combines Acoustic and EM Sensor Signal Processing", in Proc of ICASSP 2000, Istanbul, Turkey, May 2000.

[7] H. Manabe, A. Hiraiwa, T. Sugimura, "Unvoiced Speech Recognition using EMG - Mime Speech Recognition", in Proc. of CHI 2003, Fort Lauderdale, Florida, Apr. 2003.

[8] H. Bourlard and N. Morgan, "Connectionist Speech Recognition: A Hybrid Approach", Kluwer, 1994.

[9] H. Hermansky and N. Morgan, "RASTA processing of speech", IEEE Trans. on Speech and Audio Processing, vol.2, nr.4, pp. 578-589, Oct. 1994.

[10] D. Macho, L. Mauuary, B. Noe, Y. M. Cheng, D. Ealey, D. Jouvet, H. Kelleher, D. Pearce and F. Saadoun, "Evaluation of a noise-robust DSR Front-End on Aurora Databases", in Proc. of ICSLP 2002, Denver, Sep. 2002.