A Multi-Sensor Speech Database with Applications towards Robust Speech Processing in Hostile Environments

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

[1] Vrije Universiteit Brussel, Institute for Broadband Technology, dept. IBBT-ETRO-DSSP, Brussels, Belgium
[2] Voice Insight, bat. EEBIC, avenue J. Wybran 40, 1070 Brussels, Belgium
[3] Thales Communications, Signal Processing and Multimedia dept., Colombes, France
E-mail: {tdekens, gpatsis, wverhels}@etro.vub.ac.be, frederic.beaugendre@voice-insight.com,
francois.capman@thalesgroup.com

Abstract

In this paper, we present a database with speech in different types of background noises. The speech and noise were recorded with a set of different microphones and including some sensors that pick up the speech vibrations by making contact with the skull, the throat and the ear canal, respectively. As these sensors should be less sensitive to noise sources, our database can be especially useful for investigating the properties of these special microphones and comparing them to those of conventional microphones for applications requiring noise robust speech capturing and processing. In this paper we describe some experiments that were carried out using this database in the field of Voice Activity Detection (VAD). It is shown that the signals of a special microphone such as the throat microphone exhibit a high signal to noise ratio and that this property can be exploited to significantly improve the accuracy of a VAD algorithm.

1. Introduction

Many speech related applications, such as speech recognition or speech coding, are very sensitive to ambient noises. Special microphones such as bone-conduction microphones can exhibit a high degree of noise robustness. It could thus be interesting to use the signal of these microphones as input signals for such noise-sensitive applications. However, most of these applications also require an input speech signal of sufficiently high quality and clarity. As the signals picked up by those special microphones are the speech signals after propagating through bones and tissues, they are generally of limited bandwidth. That is why, more often than not, using such special microphones will not by itself solve the problem. Nevertheless they could contribute to a more noise robust solution.

As a means for investigating this, a database was recorded containing the signals of some special contact microphones together with the signals of some regular air microphones. This database can be used as a tool to investigate the properties of the microphones as such and to determine whether and how contact microphones can help to make speech applications more noise robust.

In sections 2, 3 and 4 of this paper the recording protocol and the resulting database are described. In section 5, experiments with Voice Activity Detection (VAD) using the new database show that it can indeed be beneficial to use contact microphones in noise sensitive applications in general, and to use the throat microphone signal in this particular VAD task.

2. Database recording

2.1 Field recordings of noise sources

Since we wanted a realistic scenario for the database recordings, we needed to record some real life noise sources. This was done at the BASF site in Ludwigshafen, where different kinds of noise sources were recorded during an exercise of the plant's fire brigade.

All recordings were mono recordings and were performed using 24 bit precision and 48 kHz sampling frequency. Sound pressure level (SPL) measurements were taken with a Testo 815 SPL meter [Testo]. All sound sources have been measured and recorded at a distance of less or equal to 1m. Only a fire truck equipped with jet engines (referred to as a Turbolöscher in German) was measured and recorded at 10m.

Name	Description	SPL (dBA)
Noise type 1	Turbolöser	116
Noise type 2	Diesel engine of a large fire brigade truck	83
Noise type 3	Driving with a fire brigade van with the siren on	88
Noise type 4	Water pump of a large fire brigade truck	92
Noise type 5	Interfering speaker	NA
Noise type 6	Babble noise	NA

Table 1: The 6 noise profiles used for the database

From these recordings we selected four noise profiles that were used for the recordings. We also used two additional types of noises: interfering speech of a single speaker and multi-speaker babble noise. In the file names of the database we indicated with numbers 1 to 6 which of the 6 different noise types was used during the recording. Table 1 describes the 6 noise profiles used and the SPL level that was measured during the recording of that noise profile.

2.2 Database recording setup

The recordings took place in a semi-anechoic recording room. While the speakers were reading a given text, the selected noise profile (see section 2.1) was played through loudspeakers at a given sound pressure level in order to simulate realistic conditions. In this section, we describe the procedure that we followed and the resulting specification of the database. Figure 1 shows a schematic representation of the recording setup.
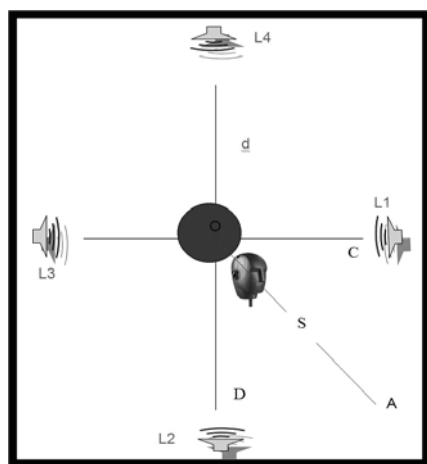

Figure 1: Schematic diagram of the recording setup

2.2.1.   Room
For the recordings a semi-anechoic chamber is used with acoustic absorption on the walls and ceiling but with a concrete floor.

2.2.2.   Speaker position
The speaker was sitting on a comfortable chair, with four loudspeakers surrounding him in a cross-like symmetry. The speaker's position was slightly out of the centre area (circle in figure 1), facing point A (the circle is an area of signal cancellation and therefore has to be avoided). Line OA is the bisector of the right angle OCD. The distance **d** between the cross centre O and any of the loudspeakers was 1 m.

2.2.3.   Control room
While the recording is done in a semi-anechoic chamber, equipment that produces noise (PC or laptops) are placed in a separate room (control room), outside the semi-anechoic chamber. All audio connections between the two rooms are done with multi-core cable.

2.2.4.   SPL meter
To measure the loudness level (in dBA) of the noise profiles played through the loudspeakers, we used a Testo 815 SPL meter [Testo]. This way, by controlling the volume of the sound produced by the loudspeakers, we can record each speaker at different noise SPL levels. The SPL meter's microphone was positioned as closely as possible to the speaker's left ear.

2.2.5.   Loudspeakers
Loudspeakers were placed on stands whose height was adjusted so that the centre of the loudspeaker is on the same level as the speaker's ears. Noise profiles were reproduced by all 4 loudspeakers at the same volume setting. Table 2 gives some information on the loudspeakers used.

Loudspeakers	Description	Model
L1, L2	2-way, 8 inch Active near-field monitor	EVENT Studio Precision 8 [Event]
L3,L4	2-way, 6.5 inch Active near-field monitor	GENELEC 1030A [Genelec]

Table 2: Loudspeaker information

2.2.6.   Microphones used for the speech recordings
Seven different kinds of microphones were used during the recordings (see Table 3). The overhead microphone was placed on a table top 50cm from the speaker's head (point S in figure 1) and pointing towards the speaker's mouth.

Mic Nr	Description	Output	Model
1	Close talk	XLR Balanced	AKG-C444L (discontinued, replaced by C555L [AKG])
2	Close talk (Bluetooth)	Mini-Jack Unbalanced	Voice Insight BlueVQL headset
3	Throat	Mini-Jack Unbalanced	Clearercom Stryker PC [Clearercom]
4	Bone contact (in ear)	Jack Unbalanced	Invisio Bone Mic Headset microphone [Invisio]
5	Overhead	XLR Balanced	Behringer XM1800S [Behringer]
6	Skull	Mini-Jack Unbalanced	MSA Gallet [MSA]
7	Close talk	Mini jack	Sennheiser PC141 [Sennheiser]

Table 3: Microphone information

2.3 Recording procedure

Before the recordings began, the speaker was told that he or she has to utter what will be prompted on a display in front of him or her while different kinds of noises will be played in the recording room. They were told to speak in a way that feels most natural to them if they know that they

are talking to a computer, which is going to recognize their speech.

As a starting point, the first noise type is played through the loudspeakers, the SPL level is measured and the volume is adjusted until the SPL level at the speaker's ear reaches the desired value. Then the first sentence of the speech corpus is displayed on the monitor in front of the speaker. Next, the signals of the different microphones are recorded. After the record button is pushed, three seconds go by before the speaker gets the sign to start speaking. While the speaker talks, a recording assistant listens to the most intelligible signal of the different microphones (the AKG444L) to ensure that the speaker makes no mistakes and that fluent speech of good quality is being recorded. If the result is not satisfactory, the current sentence will be re-recorded. After that, the next sentences in the corpus are recorded, after which the next noise type is selected, the level is adjusted and again all the sentences are recorded. When all the noise types have been used, the next SPL level is chosen. The protocol planned to start with the lowest noise SPL and to gradually increase the SPL.

2.4 The raw data recordings

As the skull microphone required that a firefighter's helmet was worn in order to have a realistic signal, and since the Bluetooth microphone could not be worn together with the helmet, we had to do two recording sessions per speaker; one with and one without helmet. This did not leave enough time to record a same speaker at multiple noise SPL levels and one SPL level was chosen per speaker.

Tables 4 and 5 give some more information about the recordings. For each recording the speaker, the noise level, the noise type and the microphones are listed. The noise types and microphones used are split up into two columns: one corresponding to the recordings without the helmet and the other one to the recording with the helmet. During the recording of the last speaker (French male 2) no SPL meter was used, but the levels of the two noise types used (background speaker and babble noise) were chosen in a way that the noise sounded realistic. For the background speaker, the recorded files of speaker 4 (French male) were used. More information about the noise type numbers can be found in table 1, the microphone numbers can be found in the section about the microphones.

Reference number	Speaker	Noise level (SPL)
1	German female	85dBA
2	German male	80dBA
3	Dutch male	75dBA
4	French male	80dBA
5	American male	75dBA
6	French male 2	NA

Table 4: Recording conditions part 1: speakers and background noise levels

Reference number	Noise type		Microphones	
	NH	H	NH	H
1	1,2,3,4	2	1,2,3,4,5	4,6
2	1,2,3	2	1,2,3,4,5	4,6
3	1,2,4	1,2,4	1,2,4,5	1,3,4,5,6
4	1,2,4	1,2,4	1,2,3,4,5	1,3,4,5,6
5	1,2,4	1,2,4	1,2,3,4,5	1,3,4,5,6
6	5,6		2,3,7	

Table 5: Recording conditions part 2: noise types and microphones (NH = without helmet, H = with helmet)

3. Metadata

3.1 Text corpus

The text corpus we used for the recordings consists of three parts, corresponding to four different tasks: commands for 2 distinct speech recognition tasks (App1, App2), numbers and phonetically balanced sentences. Table 6 gives information about the amount of utterances per type that were used for the different languages.

	App 1	App 2	Nrs	Phon bal sentences
German	50	30	25	23
Dutch	50	30	25	20
French	0	68	25	40
American	0	74	25	31

Table 6: Nr. of utterances per task and per language

3.2 Manual segmentation

For all multi-channel sound files containing uttered numbers (a total of 675 files), metadata is available that gives the start and stop positions of speech. For this we manually determined the sample where speech starts and where it stops in one channel. Since all the other channels are synchronized with this channel (except for the channel corresponding to the Bluetooth microphone, which has a small delay of about 80ms), these start and stop positions should hold for those channels as well. This information can for example be used as a reference in studies on Voice Activity Detection (VAD).

3.3 Speaker information

Six recording sessions with six different speakers were conducted. Table 7 shows some information about these speakers.

	Gender	Country	Region	Age
1	Female	Germany	Baden-Wuertemberg	22
2	Male	Germany	Dresden	27
3	Male	The Netherlands	Eindhoven	28
4	Male	France	Paris	40
5	Male	USA	Seattle, Washington	33
6	Male	France	Paris	42

Table 7: Speaker information

4. Database verification

Database verification was done in two steps. As explained in section 2.3, a recording assistant listened to the most intelligible signal of the different microphones during the recordings. In this way, we could make sure that the speaker made no mistakes and the volume of his/her voice was within acceptable limits. After the recordings, all of the recorded signals were listened to in order to verify the quality of the recorded files.

5. Application to noise-robust VAD

Since the non-traditional microphones rely on the fact that the human voice can also be transmitted through other media than air and since the ambient noises travel to the microphones through the surrounding air, these special microphones should exhibit more noise robust properties. However, because of the fact that they pick up the speech signal that propagates through bones and tissues, higher frequency components are attenuated and only a low pass speech signal is retained. These two properties make these microphones as such interesting in applications that are sensitive to noise, but where the speech quality is of less importance. VAD is an example of such an application.

5.1 Experiments

In our experiments, we used the part of the database that was manually segmented (numbers) to have a reference for the speech-pause detection. The VAD algorithm used was the INNL VAD [Dekens et al., 2007]. We applied this VAD to the selected sound files using 32ms long speech frames with 75% overlap and several threshold values. For each threshold the speech detection probability (SDP) was calculated by dividing the number of frames that were correctly classified as speech by the total number of speech frames, and the false alarm probability (FAP) was calculated as the number of pause frames that were classified as speech frames divided by the total number of pause frames.

We used two types of microphones in our experiments: the Bluetooth close talk microphone and the throat microphone. The recorded files were downsampled to 16 kHz and converted to 16 bit precision. Since some low frequency noise is present in the throat microphone signal and since this signal contains no high frequency energy, only the frequency band [250, 5000] Hz of this signal was used for the VAD. For the minimum length of a speech and a pause segment we used 250 ms and 100 ms, respectively. Detected speech regions were extended 50 ms before and after the detected region. This ensures that when speech begins or ends with certain sounds that are hard to detect in the presence of noise (e.g. /s/), these sounds can still be classified as speech while using relatively high threshold values.

Figure 2 shows an example of the power (in the frequency band used) of the signals of the two microphones in the case of an interfering speaker. The vertical lines indicate the speech start and stop positions. It can be seen that the throat microphone's signal exhibits a much higher signal to noise ratio, which is why we expect this signal to be much more suitable for a reliable Voice Activity Detection.
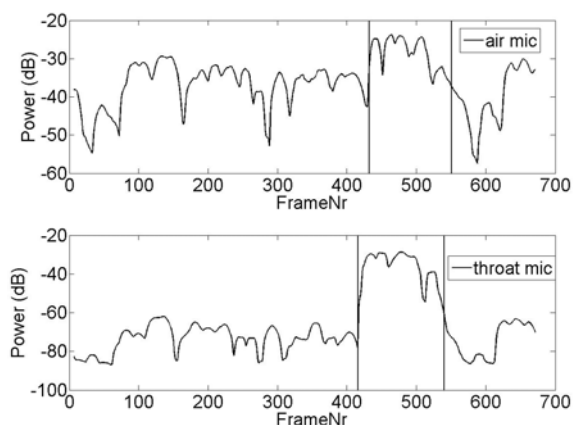

Figure 2: Signal power, top: air signal, bottom: throat signal (vertical lines indicate start and end of speech)

5.2 Results

The results of the conducted experiments can be seen in figures 3, 4 and 5. Each point in these figures shows the FAP and SDP values that were obtained for that speaker and noise type with a certain threshold value. It can be seen that in all figures the curve corresponding to the throat microphone signal reaches an acceptable SDP at lower FAP than that corresponding to the air microphone. This shows that the throat microphone signal is more
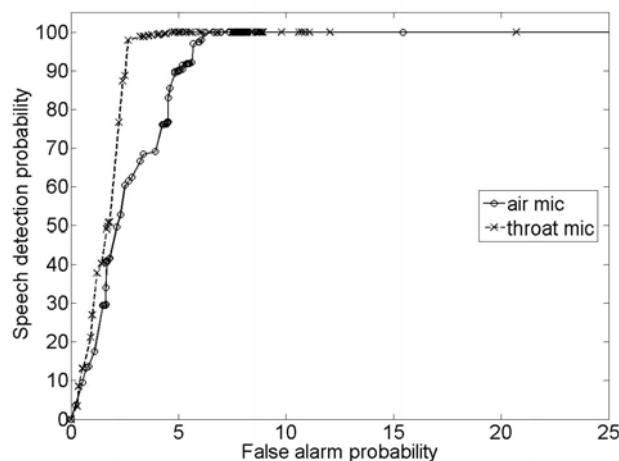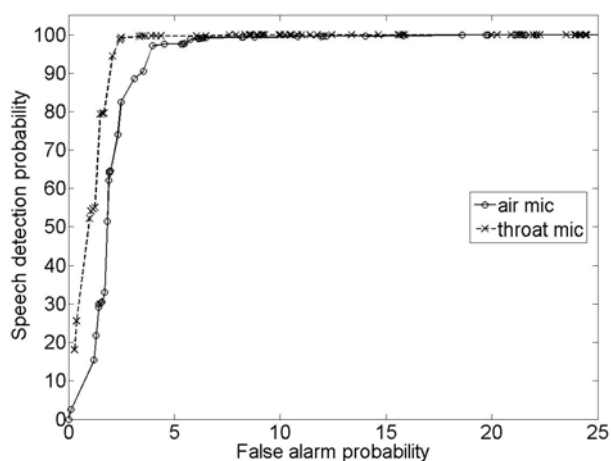

Figure 3: Speech detection probability vs False alarm probability. Left: Speaker 1, noisetype3, right: Speaker 4, noisetype1
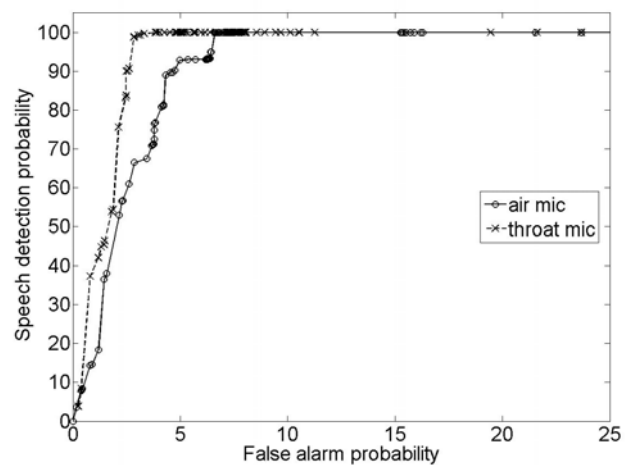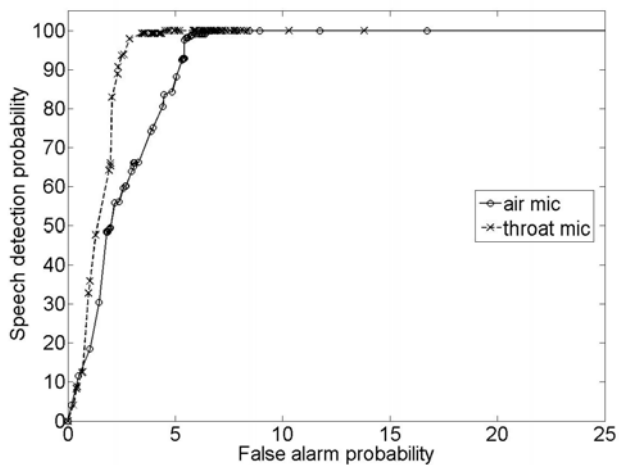
Figure 4: Speech detection probability vs False alarm probability. Left: Speaker 4, noisetype2, right: Speaker 4, noisetype4
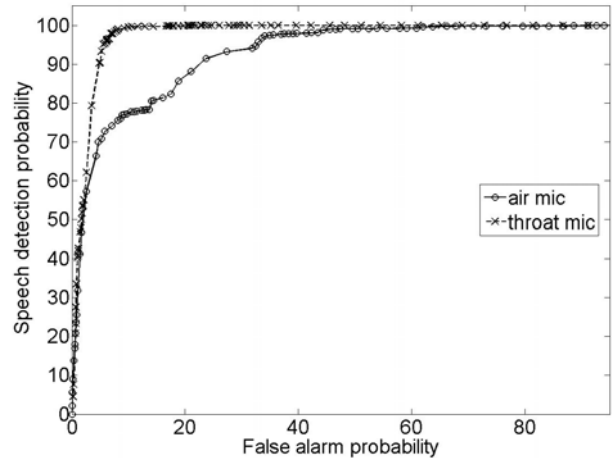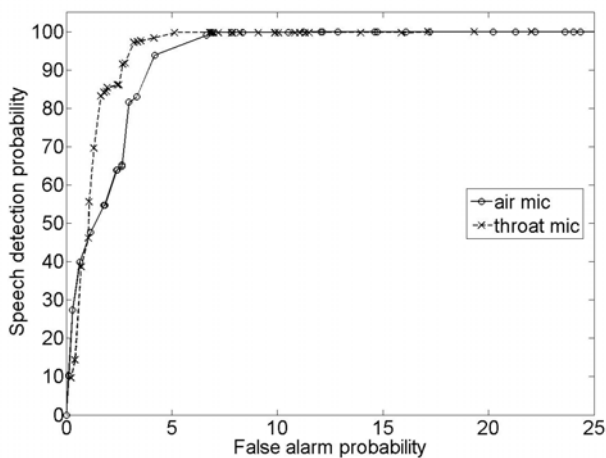

Figure 5: Speech detection probability vs False alarm probability. Left: Speaker 6, noisetype5, right: Speaker 6, noisetype5 + 6

suitable as a VAD input signal. Moreover, the right hand panel of figure 5, where the FAP and SDP are calculated using two noise types, shows that when the throat microphone signal is used, the VAD is not as sensitive to the type of noise as when the air microphone is used, i.e. with one fixed set of VAD parameters good results can be obtained in different kinds of background noises.

6. Conclusion

In this paper a database of noisy speech, picked up by a set of different microphones was presented. Using this database, it should be very interesting to study the properties of some of these microphones and their usefulness for noise robust speech applications. In this paper it was shown that one aspect of speech processing, i.e. voice activity detection, could certainly benefit from the use of these microphones.

7. Acknowledgements

8. References

AKG, http://www.akg.com/site/products/powerslave,id,985,pid,985,nodeid,2,_language,EN.html

Behringer, http://www.behringer.com/XM1800S/index.cfm?lang=eng

Clearercom, http://www.clearercom.com/pc_throat_mic.htm

Dekens T., Demol M, Verhelst W. and Beaugendre F, (2007), "Voice Activity Detection based on Inverse Normalized Noise Likelihood Estimation," *proceedings of the XIII-th Convention of Electrical Engineering, CIE 2007, Santa Clara, Cuba*. Available at http://www.etro.vub.ac.be/Research/DSSP/Publications/Publications.htm

Event, http://www.eventelectronics.com/index/index.php?page=Products&product=SP8

Genelec, http://www.genelec.com/products/previous-models/1030a/

Invisio,
http://www.swatheadsets.com/tactical/invisio/invisio.h
tml
MSA, http://www.gallet.fr/index.php?id=188&L=5
Sennheiser,
http://www.sennheisercommunications.com/comm/ic
m_eng.nsf/root/05350
Testo,
http://www.testo.co.uk/online/abaxx-?$part=PORTAL.
GBR.Applications&$event=show-from-content&exter
nalid=opencms:/Products/MeasurementParameters/am
bientairquality/Messgeraete/testo_815/Englisch.produ
ct