ABSTRACT

This paper presents experimental results for a novel microphone solution to ASR in very noisy environments. PARAT is an intelligent hearing protector and communication terminal implemented as a lightweight earplug. The inner microphone in PARAT captures the speech signal behind a seal in the auditory canal that impedes contamination from external noise sources. Experiments on a stereo speech database contaminated with noise show that the PARAT microphone compares favourably with a noise cancelling close-talking microphone of high quality on a speaker dependent connected digit task.

1. INTRODUCTION

Automatic speech recognition (ASR) may facilitate hands-free and eyes-free operation of machines and computer applications. However, many real-world contexts where speech input would be valuable involve adverse acoustic environments, and the ASR performance is sensitive to noise contamination of the speech signal [1]. Applications that would be very useful if the speech recognition was accurate may therefore be useless in field deployment.

This problem has motivated noise-robust strategies such as signal and feature compensation [2] and model compensation [3]. Popular methods include non-linear spectral subtraction (NSS) [4] and parallel model combination (PMC) [5], both of which are quite effective when the noise is stationary or varying slowly. NSS and PMC are however much less robust when the noise is changing rapidly or the noise level is severe.

A complementary approach to noise robustness focuses on the signal acquisition level of the ASR system. Some applications allow the designer (or the user) to select a microphone technology that can reduce signal contamination at the source, and thus improve ASR performance. Two established noise-robust technologies are microphone arrays and noise cancelling close-talking microphones (CTMs). The first uses beamforming to capture sound from the direction of the speaker, while attenuating background noise from other directions [6]. Microphone arrays must be mounted on a surface facing the speaker, and this constraint limits the applicability. While suitable for car, cockpit and desktop applications, microphone arrays are impractical for mobile users in the field. CTMs are popular for ASR applications, and have proved effective in noisy environments [6]. The noise cancelling microphone sits at the end of a boom mounted on a headset or helmet, and its performance depends on a correct position relative to the mouth. These constraints may be impractical in some applications.

2. THE PERSONAL ACTIVE RADIO/AUDIO TERMINAL

PARAT [7][8] is an intelligent hearing protector and audio interface for communication systems. A problem with traditional hearing protection is that people tend not to use it continuously, even in adverse acoustic environments, because it hampers communication with other people. The primary idea behind the PARAT concept is to integrate communication solutions with hearing protection, in a way that encourages continuous use.

Figure 1: The PARAT communication terminal.

A principle sketch of PARAT is shown in Figure 1. An earplug with a seal sits in the ear. The inner microphone is used to capture the user’s voice. In quiet environments the sound captured by the outer microphone is reproduced unaltered by the loudspeaker on the inside of the earplug. When the noise becomes severe, the trans-
mission from the outer microphone is processed and gradually reduced. In determining the appropriate attenuation the system distinguishes between speech and noise. As all the processing happens on a sample-by-sample level, PARAT can handle even strong impulse noise, like gun shots. In addition to the passive noise attenuation by the earplug, PARAT can utilise active noise reduction (ANR) for further dampening of low-frequency noise components.

From previous experience it is clear that the speech signal captured inside the earplug is relatively clean and thus well suited for radio communication even in very noisy surroundings. In this work we are interested in the properties of the inner microphone used with ASR. In the next section we describe the test bench that has been set up for these experiments.

3. PARAT TEST BENCH

Our experiments are based on a two-speaker stereo database collected with two microphones, PARAT and H-374(V)5. The latter is a high-quality noise cancelling CTM used in military communication systems. The recordings were done with custom-made recording and prompting software over several days in an anechoic chamber. The speakers employed a loud voice level that was kept approximately constant relative to a target level on a digital sound level meter. Sound files are sampled at 16 kHz and 16 bits/sample. Note that ANR was not used with PARAT.

The domain of the database is speaker dependent recognition of Norwegian digit strings of known length. The training set consists of 1700 digits per speaker, realised as 100 isolated digits, 200 3-digit strings, and 200 5-digit strings. Each digit model is thus trained on 170 examples from a variety of (digit string) contexts. The test set consists of 1600 digits per speaker, in the form of 200 3-digit strings, and 200 5-digit strings.

Noisy speech was simulated by contaminating the clean speech signals with additive noise signals corresponding to specified noise levels in the room. Our approach was to estimate the transfer functions from a calibrated reference microphone placed in the vicinity of the speaker, to each of the two microphones used to capture speech. We created a controlled noise environment, in the form of a diffuse sound field, by playing pink noise through a loudspeaker in an echo chamber. For each speaker the power in 1/3 octave bands were measured for all three microphones. Notice that the upper cut-off frequency is below 3 kHz for PARAT.

In order to establish the absolute levels, the 1 kHz 1/3 octave dB SPL and dBA levels were also measured with the reference microphone for each of the noise types used in our experimental work. These are described in some detail in Section 4.2.

4. PARAT SIGNAL CHARACTERISTICS

4.1 The PARAT clean speech signal

The PARAT speech signal differs from speech signals captured by an ordinary microphone in several ways. One of the more noticeable differences is a marked increase in signal intensity for the nasals /m/, /n/ and /ng/, compared to the CTM signal. Some other consonants, such as the liquids /l/ and /r/, are affected similarly. Another difference is found in the closure portions of the voiced stops /b/, /d/ and /g/. Here the CTM signal contains weak periodic energy in the low-frequency region, whereas the PARAT signal contains wide-band energy in formant-like structures. Figure 2 gives examples of nasals, liquids and stops.

![Figure 2: Example of the utterance ‘adrenalin’ captured simultaneously on PARAT and the CTM.](image)

The effects above may be positive in the sense that the affected parts of the PARAT speech signal become more prominent in the presence of noise. It is however not known how much (phonetically) discriminative information these parts contain. In a more negative vein we have observed that non-speech vocalisations caused by for instance breathing, coughing, swallowing and clicking of teeth result in comparatively strong PARAT signals. Also, the PARAT speech signal has a limited bandwidth, due to a strong attenuation of high frequencies in the acoustic channel between the vocal tract and the ear canal. Figure 3 shows estimates of the long-term power spectral densities of speech for both PARAT and the CTM. Notice that the upper cut-off frequency is below 3 kHz for PARAT.

A systematic investigation of general acoustic-phonetic distinctions in the PARAT signal, and their effect on ASR, depends on speech data from a larger number of speakers.

![Figure 3: Long term power spectral density estimates for speech captured with PARAT and the CTM.](image)
4.2. The PARAT speech signal in ambient noise

Three kinds of environmental noise were used, here called CV90, Babble and Factory1. The first is an in-house recording from the interior of a military combat vehicle. The last two was taken from the NOISEX database [10]. Babble is background speech from approximately 100 people in a canteen, while Factory1 was recorded near plate cutting and electrical welding equipment. Of these CV90 is the most stationary and Factory1 the least stationary. Figure 4 shows power spectral density estimates for the noise signals. The curves are arbitrary set equal at 1 kHz. The Babble signal has a speech-like characteristic, while the CV90 signal has most energy in the low-frequency region. Compared to these the Factory1 signal has its energy more evenly distributed.

Figure 4: Power spectral density estimates for the CV90, Babble and Factory1 noise signals.

We view the transfer functions from the location of the reference microphone to the digitised PARAT and CTM microphone signals as normalised transfer functions multiplied by scaling factors. The normalised transfer functions were found as described in Section 3. Linear phase FIR filters were designed using a simple frequency sampling procedure, and applied to the noise signals. The scaling factor depend on the desired noise level. As the 1 kHz 1/3 octave dB SPL and dBA levels are known for both the pink noise and the ambient noise, the filtered ambient noise can easily be scaled to correspond to an arbitrary noise level in the room and then added to our speech database.

5. EXPERIMENTAL RESULTS

The goal of this work was to establish the performance of speaker dependent (SD) whole-word acoustic models of low complexity trained on clean speech. Such models are relevant to some of the scenarios where PARAT is considered for ASR. Compact SD ASR engines can run on mobile devices, and professional users of small vocabulary ASR applications are often willing to provide a small number of training samples of each vocabulary unit. Users are moreover free to employ their own dialect.

Our speech recognition system is similar to that defined in the Aurora project for recognition of connected digits in noise [11]. The front end calculates 12 mel-frequency cepstral coefficients and log-energy every 10 ms from 25 ms speech frames, extended by 1st and 2nd order derivatives. Each digit is modeled by a left-to-right HMM with 8 states (compared to 16 in the Aurora recogniser) and no skips. Two pause models are used. One called "sil" has 3 states, and shall model silence before and after the utterance. The other called "sp" shall model pauses between digits. This is a tee model [12], which consists of a single state tied to the middle state of the "sil" model.

State observation densities are modeled by Gaussian mixture pdfs with diagonal covariance matrices. The digit models use 3-element mixtures while the pause models use 6-element mixtures. Note that in contrast to the Aurora recogniser all Gaussians share a common (global) variance vector. This is appropriate when the training data is limited, but is also known to improve the performance in adverse conditions [13]. The training is done in several steps by applying the embedded Baum-Welch algorithm as described in [11]. The recognition grammar was restricted to the correct number of digits. This was done to avoid the need to tune a recognition system parameter, the word insertion penalty [12], to our recognition task. The recogniser is implemented in HTK [12].

Testing on clean speech gave about 100% accuracy for both speakers and both microphones. This indicates that the recognition task is too easy to differentiate the microphones in clean conditions.

Figure 5 shows the normalised transfer functions for the two cases for speaker A. The absolute scaling is arbitrary, but not the relative scaling. The magnitude response for the PARAT microphone lies below that of the CTM, except around 2 kHz. This peak in the PARAT transfer function is due to a high-pass microphone filter that is used to counteract the attenuation of high frequencies between the vocal tract and the ear canal.

Figure 5: Transfer functions from a reference microphone in the room to PARAT and the CTM.

Figure 6 and 7 show the recognition results for both microphones and all noise types and levels for speaker A and B, respectively.
For speaker A the PARAT recognition accuracy is close to 100% at 100 dBA for Babble and Factory1 noise and at 105 dBA for CV90 noise. The CTM achieves comparable results at noise levels that are 5-10 dB lower. We observe that the recognition results are far worse for speaker B than for speaker A in all conditions. The cause is probably a much stronger tendency to coarticulation between the digits for speaker B. We also note that the performance difference is smaller between the two microphones. An analysis of the transfer function from the room to PARAT for speaker B revealed that the attenuation of external noise was not as good as for speaker A. This indicates that there may be significant speaker dependent effects with the noise attenuation for the PARAT microphone.

6. CONCLUSIONS AND FURTHER WORK

In this paper a novel platform for ASR in adverse acoustic environments has been benchmarked on a two-speaker stereo database. PARAT is a hearing protector and communication terminal implemented as a lightweight earplug. On a connected digit task the PARAT inner microphone has shown robustness to noise levels up to 105 dBA with speaker dependent whole-word models. This was compared to a close-talking microphone of high quality. PARAT performed better than the CTM in all noisy conditions. The difference in performance between the two speakers can be explained in part by more coarticulation between digits and in part by less attenuation of external noise with PARAT for speaker B.

Our focus has been on comparing performances achieved with the two different microphone solutions, under different noise conditions. The system performance has not yet been optimised, but experiments with noise robust feature extraction are under planning.

ACKNOWLEDGEMENTS

The work was in part financed by the SINTEF Telecom and Informatics strategic project AMBiLAB (Ambient
Laboratory) and the strategic project Acoustic Research Centre funded by the Research Council of Norway. We want to thank Sverre Stensby and Trym Holter at SINTEF Telecom and Informatics for their help with this project.

REFERENCES


[8] NACRE AS: Natural Communication in Rough Environments (http://www.nacre.no/)


