APPLICATIONS OF NAM MICROPHONES IN SPEECH RECOGNITION
FOR PRIVACY IN HUMAN-MACHINE COMMUNICATION

Panikos Heracleous, Tomomi Kaino, Hiroshi Saruwatari, Kiyohiro Shikano
Nara Institute of Science and Technology, Japan
e-mail: {panikos,tomomi-k,sawatari,shikano}@is.naist.jp

Abstract

In this paper, we present the use of stethoscope and silicon NAM microphones in automatic speech recognition. NAM microphones are special acoustic sensors, which are attached behind the talker’s ear and can capture not only normal (audible) speech, but also very quietly uttered speech (non-audible murmur). As a result, NAM microphones can be applied in automatic speech recognition systems when privacy is desired. Previously, we presented speech recognition experiments for non-audible murmur captured by a stethoscope microphone. In this paper, we also present recognition results using a more advanced NAM microphone, the so-called silicon NAM microphone. Using adaptation techniques and a small amount of training data, we achieved a 93.9% word accuracy for non-audible murmur recognition. We also report experimental results in noisy environments showing the effectiveness of using a NAM microphone in noisy environments. In addition to a dictation task, we also present a keyword spotting experiment based on non-audible murmur.

1. Introduction

The NAM microphone [1] belongs to the acoustic sensor paradigm, in which speech is conducted not through the air, but within body tissues, bone, or the ear canal. The silicon and stethoscope NAM microphones were developed by Nakajima et al. in the Nara Institute of Science and Technology, Japan. The NAM microphone is attached behind the talker’s ear and speech is captured through body tissue. The bone-conductive microphone used in [2] and the throat microphone used in [3] are acoustic sensors similar to NAM microphones. Basically, in those studies a non-conventional acoustic sensor combined with a standard microphone was used to increase the robustness against noise. In [4] a prototype stethoscope NAM microphone and a throat microphone were used for soft whisper recognition in a clean environment. Our current research, focuses on the recognition of non-audible murmur using a NAM microphone in various environments. A speech recognition system able to recognize very quietly uttered speech can be used when privacy is preferable in human-machine communication. We should note, however, that the NAM microphone can also be used for audible speech recognition. Moreover, non-audible murmur recognition and audible speech recognition using NAM microphone can be integrated very effectively [5]. Speech captured by a NAM microphone has different characteristics compared with air-conducted speech. However, body tissue acts as a low-pass filter and the high frequencies are attenuated.

Figure 1 shows the power spectrum of a Japanese syllable captured by a NAM microphone and figure 2 shows the power spectrum of the same syllable captured by a close-talking microphone. Figures show the spectrum similarities between the two speeches up to 1kHz. After 1kHz, the NAM frequency components are attenuated, so that at 5kHz they are attenuated by almost 36dB compared with 1kHz components. Due to these differences, normal-speech hidden Markov models (HMMs)
In this experiment, both training and test data were recorded in a clean environment by a male speaker using NAM microphones. For training, 350 and for testing 48 non-audible murmur utterances were used. Figure 3 shows the achieved results. As the figure shows, the results are very promising. Using a small amount of data and adaptation techniques, we achieved a word performance comparable to normal-speech recognition (96.2% word accuracy). More specifically, using a stethoscope microphone we achieved an 88.9% word accuracy and using a silicon NAM microphone we achieved a 93.9% word accuracy for non-audible murmur recognition. The results also show the effect of the multi-iteration adaptation scheme. As can be seen, with increasing number of adaptation iterations, the word accuracy was markedly increased.

2. Speaker-dependent non-audible murmur recognition

In this section, we present experimental results for speaker-dependent non-audible murmur recognition using NAM microphones. The recognition engine used was the Julius 20k vocabulary Japanese dictation toolkit. The initial models were speaker-independent, gender-independent, 3000-state phonetic tied mixture (PTM) HMMs, trained with the JNAS database and the feature vectors were of length 25 (12 MFCC, 12 ΔMFCC, ΔE).

The non-audible murmur HMMs were trained using a combination of supervised 128-class regression tree MLLR and MAP adaptation methods. Using, however, the MLLR and MAP combination, the parameters are initially transformed using MLLR, and the transformed parameters are used as priors in MAP adaptation. In this way, during MLLR the acoustic space is shifted and the MAP adaptation performs more accurate transformations. Moreover, due to the use of a regression tree in MLLR, parameters which do not appear in the training data, and therefore are not transformed during MAP, are transformed initially during MLLR.

Due to the large difference between the training data and the initial models, single-iteration adaptation is not effective in non-audible murmur recognition. Instead, a multi-iteration adaptation scheme was used. The initial models are adapted using the training data and the intermediate adapted models were trained. The intermediate models were used as initial models and were re-adapted using the same training data. This procedure was continued until no further improvement was obtained. Results showed, that after 5-6 iterations significant improvement was achieved compared with the single-iteration adaptation. This training procedure is similar to that proposed by Woodland et al. [8], but the object is different.

2.1. Non-Audible murmur recognition using clean data

In this experiment, both training and test data were recorded in a clean environment by a male speaker using NAM microphones. For training, 350 and for testing 48 non-audible murmur utterances were used. Figure 3 shows the achieved results. As the figure shows, the results are very promising. Using a small amount of data and adaptation techniques, we achieved a word performance comparable to normal-speech recognition (96.2% word accuracy). More specifically, using a stethoscope microphone we achieved an 88.9% word accuracy and using a silicon NAM microphone we achieved a 93.9% word accuracy for non-audible murmur recognition. The results also show the effect of the multi-iteration adaptation scheme. As can be seen, with increasing number of adaptation iterations, the word accuracy was markedly increased.

2.2. Non-audible murmur recognition using simulated noisy data

In this experiment, the same clean 350 utterances were used for adaptation. For testing, 48 noisy non-audible murmur utterances were used. Noise recorded in an office was played back at 50 dBA, 60 dBA and 70 dBA
levels and was recorded using NAM microphones. The recorded noises were superimposed onto clean data to create the noisy test data.

Figure 4 shows the obtained results. As can be seen, for the 50 dBA and 60 dBA noise levels the performance was almost equal to that of the clean case. When the noise level became 70 dBA, the performance decreased, however, still non-audible murmur recognition with reasonable results was possible. Note, that no additional noise reduction approaches were used, and that the HMMs were trained using clean data. Results show that stethoscope NAM microphone is less robust against noise, particularly at the 70 dBA noise level.

Figure 5 shows the long-term spectrum of the noise used in our experiments. The noises captured by NAM microphones were superimposed onto the clean test data to simulate the noisy test data. Figure 6 shows the spectrum of the noise recorded using NAM microphones at 70 dBA level. The figure shows the similarity in the spectra of the two captured noises. Differences appear between 3kHz and 5kHz, where noise captured by the stethoscope microphone shows a higher spectral content. This might explain the significant decrease in word accuracy at 70 dBA when using the stethoscope microphone.

3. A keyword-spotting experiment using non-audible murmur

In this section, we present a keyword-spotting experiment for non-audible murmur. A non-audible murmur-based keyword spotting system, however, can be applied to extract a specific number of keywords from unconstrained

input speech in privacy conditions. In some applications, when only a small number of keywords is required, a keyword-spotting system, with lower complexity and faster decoding, might be more reasonable than a dictation system.

In a keyword-spotting approach, not only the keywords, but also the non-keyword intervals must be modeled explicitly. Our approach, was based on phonemic garbage models [9]. The keywords were modeled using context-dependent HMMs, and monophone HMMs were used to model the non-keyword portions. Both HMM sets were trained with non-audible murmur data recorded using a silicon NAM microphone. Fourty-three monophone HMMs were connected as to allow any sequence. The vocabulary consisted of 25 keywords randomly selected from JNAS database. Figure 7 shows the grammar used in our experiment, which allowed at most one keyword per utterance.

In our experiment, the following evaluation measures were used:

- Detection rate. The percentage of keywords detected.
- Rejection rate. The percentage of non-keywords rejected.
- Receiver Operating Characteristics (ROC) and Figure of Merit (FOM). The putative hits are sorted with respect to their scores, and the probability of detection at each false alarm is computed. The FOM is calculated as the average probability of detection between 0 and 10 false alarms per keyword.

For testing, we used 18 utterances, which included one keyword, and 24 utterances which did not include any keyword. Figure 8 shows the ROC curves. The figure shows, that by allowing 4 alarms per keyword we achieved 88.2% detection rate. The achieved FOM was 85.6%, which is promising result. The figure also shows, that using duration normalized scores the performance was decreased. Figure 9 shows the detection and rejection rates. To achieve higher detection and rejection rates, a word insertion penalty is tuned to decrease the likelihood of the garbage models. Without this tuning, however, a large number of false rejections (keywords are hypothesized as garbage models) appears, and as a result the detection rate decreases. With word insertion penalty
tuning, we achieved a 82.5% equal rate (equal detection and rejection rates).

4. Conclusions

In this paper, we presented applications of non-audible murmur using NAM microphones. NAM microphones belong to a new paradigm of body-conductive acoustic sensors which can be used in automatic speech recognition systems when privacy in human-machine communication is desired. Using NAM microphones, we recognized very quietly uttered speech with a performance comparable to that of normal-speech recognition. More specifically, with a small amount of clean training, test data and adaptation methods we achieved a 93.9% word accuracy using a silicon microphone and an 88.9% word accuracy using a stethoscope microphone. We also carried out experiments using simulated noisy test data with very promising results. At 50 dBA and 60 dBA noise levels, the performance remains almost equal to that in the clean case. We also presented a keyword spotting experiment based on non-audible murmur. Using a 25-keyword vocabulary, we achieved a 88.2% detection rate when allowing 4 false alarms per keyword. By tuning word insertion penalty as a universal threshold, we achieved a 82.5% equal rate. Both evaluation methods show that keyword spotting can be applied effectively in non-audible murmur paradigm. Since non-audible murmur recognition is a new phenomena, further analysis and investigation are necessary. In addition, we also plan to carry out speaker-independent experiments for non-audible murmur and to conduct keyword-spotting experiments using larger vocabularies and in various environments.

5. References


