Noise-robust whispered speech recognition using a non-audible-murmur microphone with VTS compensation

Chen-Yu Yang, G. Brown, Liang Lu, J. Yamagishi, and S. King. Noise-robust whispered speech recognition using a non-audible-murmur microphone with vts compensation. In Chinese Spoken Language Processing (ISCSLP), 2012 8th International Symposium on, pages 220-223, 2012.

NOISE-ROBUST WHISPERED SPEECH RECOGNITION USING A NON-AUDIBLE-MURMUR MICROPHONE WITH VTS COMPENSATION

Chen-Yu Yang[2], *Georgina Brown*[1], *Liang Lu*[1], *Junichi Yamagishi*[1], *Simon King*[1]

[1]The Centre for Speech Technology Research, The University of Edinburgh, U.K.
[2]iFLYTEK Speech Lab, University of Science and Technology of China, P.R.China
yangcy@mail.ustc.edu.cn, liang.lu@ed.ac.uk, jyamagis@inf.ed.ac.uk, Simon.King@ed.ac.uk

ABSTRACT

In this paper, we introduce a newly-created corpus of whispered speech simultaneously recorded via a close-talking microphone and a non-audible murmur (NAM) microphone in both clean and noisy conditions. To benchmark the corpus, which has been freely released recently, experiments on automatic recognition of continuous whispered speech were conducted. When training and test conditions are matched, the NAM microphone is found to be more robust against background noise than the close-talking microphone. In mismatched conditions (noisy data, models trained on clean speech), we found that Vector Taylor Series (VTS) compensation is particularly effective for the NAM signal.

Index Terms— whisper recognition, non-audible murmur(NAM), silent speech interface (SSI), vector Taylor series (VTS), noise robustness

1. INTRODUCTION

Although Automatic Speech Recognition (ASR) is well developed and highly effective in carefully-targeted applications, here remain situations in which people have difficulty in using ASR technology. It may sometimes be socially-unacceptable or embarrassing to speak to a machine loudly and clearly in the presence of others. Or, the performance of ASR systems may degrade below useful levels in noisy conditions, particularly if the user cannot speak loudly. One approach to these problems is to perform ASR from signals other than the conventional acoustic wave acquired using a microphone; when the alternative signal can be acquired without the user speaking in the normal way, this is some called a silent speech interface (SSI).

Silent speech interfaces enable the human-machine speech communication to take place without the necessity of emitting an audible acoustic signal. To date, several different types of technology have been used for the SSI systems [1, 2]. These can be very effective in situations where normal microphones may not work very well.

A non-audible murmur (NAM) microphone is a kind of special microphone which can be used as the sensing device of a SSI system. The NAM microphone is a special body-conductive microphone [3]. It can be used to detect extremely quiet speech (NAM), that even listeners around the speaker can hardly hear. NAM speech tends to be unvoiced, like whispering. The best position to place the NAM microphone is just behind the ear [4]. It can be used to detect various kinds of speech, including whispering and normal speech, conducted through the soft tissue of the head [5]. It is more robust to environmental noise than an ordinary microphone, because of its noise-proof structure. Compared to other kinds of SSI systems, which may involve electrodes or other sensing devices, a NAM microphone-based SSI system is non-intrusive, cheap and convenient.

In this paper, we compare the performance of the whispered speech recognition for a NAM microphone and an ordinary close-talking microphone. To do this, we recorded a new whispered speech database in English, which has been publicly released. We created ASR benchmarks for this corpus using standard techniques such as heteroscedastic linear discriminant analysis (HLDA) projection and minimum phone error (MPE) training. Although the NAM microphone has a noise-proof construction, it may be still somewhat affected by environmental noise and the resulting acoustic mismatch could decrease ASR accuracy, if the acoustic models were trained on clean data. Therefore we also applied vector Taylor series (VTS) compensation [6, 7] to the NAM clean-speech model. The experimental results show that VTS compensation is highly effective for compensating the mismatch between the NAM clean model and the NAM noisy test data.

Sections 2 and 3 of this paper describe our whispered speech database recorded with two microphones in parallel; a brief overview of techniques examined is provided. In section 4, the experimental results are presented and a summary is given in section 5.

2. THE PARALLEL WHISPERED SPEECH DATABASE

We have created a parallel whispered speech database recorded simultaneously via a NAM microphone which uses urethane-elastomer to create a close contact with the skin [5] and an omni-directional headset-mounted condenser microphone (a DPA 4035). The database comprises 420 sentences (about

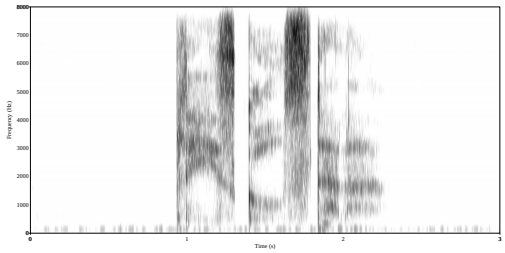

Fig. 1. *"Please call Stella", recorded in clean conditions by headset microphone.*

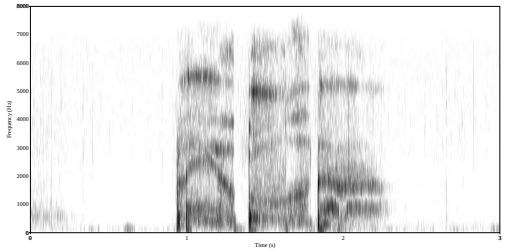

Fig. 2. *"Please call Stella", recorded in clean conditions by NAM microphone.*

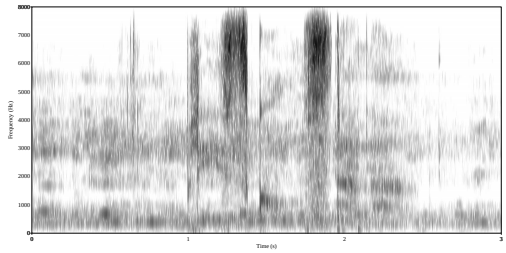

Fig. 3. *"Please call Stella", recorded in noisy conditions by headset microphone.*

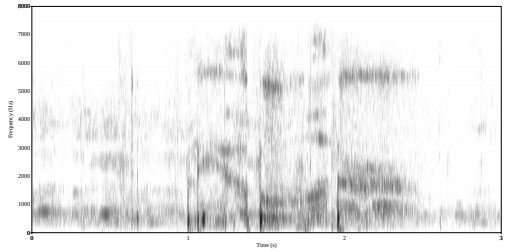

Fig. 4. *"Please call Stella", recorded in noisy conditions by NAM microphone.*

943 words), which were selected from newspaper text, uttered by a young female speaker. It is divided into two sections: one recorded in clean conditions and the other one in pre-recorded cafeteria noise played over a loudspeaker at 65 dB [A] (resulting in SNR of approximately 10 dB). Both sections of the corpus were recorded in a soundproof hemi-anechoic chamber (noise floor around 25 dB [A]) at 96kHz sampling rate and 24 bit sample depth into a Pro Tools HD system.

Figures 1 and 2 illustrate the general properties of the data recorded in clean conditions. From these spectrograms, it can be seen that the high frequency components captured by the NAM microphone are substantially attenuated compared to the headset microphone. Figures 3 and 4 show spectrograms for the noisy condition. It can be seen that the headset microphone captures much more background noise than the NAM microphone; the NAM microphone does capture some background noise, despite its noise-proof construction.

This corpus has been released and is available freely at http://homepages.inf.ed.ac.uk/jyamagis/release/CSTR-NAM-TIMIT-Plus-ver0.81.tar.gz The released corpus includes not only Herald sentences but also TIMIT sentences for further analysis and research.

3. WHISPERED SPEECH RECOGNITION

For benchmarking, standard algorithms including heteroscedastic linear discriminant analysis (HLDA) projection and minimum phone error (MPE) training were applied.

Although the NAM microphone is relatively insensitive to the external noise [8], some background noise may be still captured by the NAM microphone and it will seriously impact the accuracy of the recogniser, especially in the mismatched condition. Techniques to improve the robustness of ASR engines can be applied to both feature and model domains. As a preliminary study, here we only use model-based vector Taylor series (VTS) compensation to handle the mismatch introduced by the noise. In this work, the mismatch function for static features in cepstral domain used by VTS is [7]

$$\begin{aligned}\mathbf{y} &= \mathbf{x} + \mathbf{h} + \mathbf{C}\log\left(\mathbf{1} + \exp\left(\mathbf{C}^{-1}(\mathbf{n} - \mathbf{x} - \mathbf{h})\right)\right) \\ &= \mathbf{f}\left(\mathbf{x}, \mathbf{n}, \mathbf{h}\right) \end{aligned} \quad (1)$$

where $\mathbf{x}$ and $\mathbf{y}$ are the clean and noise distorted speech, $\mathbf{h}$ and $\mathbf{n}$ stand for the channel distortion and additive noise respectively. By applying the first order VTS expansion and taking expectation with respect to the static parameters of Gaussian component m, the updated static parameters are obtained as

$$\boldsymbol{\mu}_y^m = \mathbf{f}(\boldsymbol{\mu}_x^m, \boldsymbol{\mu}_h, \boldsymbol{\mu}_n) \quad (2)$$

$$\boldsymbol{\Sigma}_y^m = \mathbf{G}_m \boldsymbol{\Sigma}_x^m \mathbf{G}_m^T + \left(\mathbf{I} - \mathbf{G}_m\right) \boldsymbol{\Sigma}_n \left(\mathbf{I} - \mathbf{G}_m\right)^T \quad (3)$$

where $\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n$ are the mean and covariance of additive noise, and $\boldsymbol{\mu}_h$ denotes the mean of channel distortion. $\mathbf{G}_m$ is the Jacobian matrix as $\frac{\partial \mathbf{f}(\cdot)}{\partial \mathbf{x}}|_{\boldsymbol{\mu}_x^m, \boldsymbol{\mu}_n, \boldsymbol{\mu}_h}$. The dynamic parameters are derived by the continuous time approximation [9].

4. EXPERIMENTS

4.1. Experimental conditions

Since the database is small, we used 5-fold cross validation in all of our experiments: in each fold, 336 sentences were used

as the training set and 84 sentences as the test set. The original recordings were at 96kHz but were down-sampled to 8kHz in the experiments. The reasons for trying an 8kHz sampling rate include severely attenuated high-frequency components (above 4 kHz) of the signal captured by the NAM microphone due to the lack of lip radiation and the low-pass characteristics of soft tissue [10].

In all experiments, HTS tools [11] version 2.2 were used for training acoustic models and decoding. When comparing NAM and headset microphones, we used two types of feature vector: 12th-order MFCCs and log energy plus their delta and acceleration coefficients; 12th-order PLPs and the 0th PLP coefficient plus their delta and acceleration coefficients. Cepstral Mean Normalisation (CMN) and pre-emphasis were also applied to both feature types. However, for VTS experiments, we only use 12-order MFCC features with the 0th MFCC coefficient appended their delta and acceleration coefficients without CMN. [1]

The acoustic modelling unit was triphones, each modelled by a left-to-right continuous density hidden Markov model (CDHMM), with 3 emitting states. A decision tree-based clustering method using the minimum description length (MDL) stopping criterion [12] was applied to control model complexity and deal with sparsity.

Given that our data are limited, the number of Gaussian mixture components in the acoustic models for recognition and VTS compensation was set at 3. Before estimating the HLDA transform, an acoustic model with 52-dimension feature vectors (i.e., with third differential coefficients appended) was estimated first. After applying the estimated HLDA transform, the acoustic model was projected back to 39 dimensions. MPE training was based on the model obtained after HLDA projection.

A bigram language model for a 126k-word vocabulary (this is the size of the dictionary we used) trained on the complete text of the parallel database by SRILM toolkit [13] was used in our system. The dictionary which we used was an English lexicon called Combilex [14].

4.2. Whispered Speech Recognition using NAM and headset microphones

The experiments in this section are for matched conditions, where the training data and test data are either both clean, or both noisy. All results are the averages over 5 folds of cross-validation.

Table 1 shows word accuracy in clean conditions. We can see that HLDA projection and MPE training are very effective in whispered speech recognition using either microphone type. The best performance under clean conditions using the headset microphone is slightly better than for the NAM microphone, 3–4 percent absolute higher.

[1] This feature configuration is chosen to fit into the VTS code.

Table 1. *Word accuracy (%) in clean conditions.*

	MFCC_E_D_A_Z(_T)		PLP_0_D_A_Z(_T)	
Method	headset	NAM	headset	NAM
Baseline	80.2	75.4	76.7	76.5
+ HLDA	80.4	76.6	79.8	76.6
+ MPE	80.9	77.3	81.4	78.3

Table 2. *Word accuracy (%) in noisy conditions.*

	MFCC_E_D_A_Z(_T)		PLP_0_D_A_Z(_T)	
Method	headset	NAM	headset	NAM
Baseline	58.6	65.2	53.3	68.9
+ HLDA	61.2	72.9	59.2	74.9
+ MPE	66.5	74.7	65.2	76.3

Table 2 shows the performance in noisy conditions. Again, HLDA projection and MPE training all improve word accuracy. We now see that the headset microphone is more sensitive to background noise than the NAM microphone. PLP features provide the best word accuracy for the NAM microphone. The most striking result is that the word accuracy of the NAM microphone using PLP features, HLDA projection and MPE training is comparable to the accuracy in clean conditions (76.3% vs. 78.3%) .

4.3. VTS compensation of the NAM clean acoustic model

The experiments reported above used training data and test data in matched conditions for both the headset and NAM microphones. The next experiment concerns robustness to mismatch between training and test conditions. We use an acoustic model trained on NAM speech from clean conditions, to recognise NAM speech from noisy conditions, and examine the benefits of applying VTS compensation to the clean model. Note that the noisy speech data in this experiment is not artifically corrupted one but real noisy data In this experiment, waveforms with a sampling rate of 8kHz were used, with MFCC features. Note that 'MFCC_0_D_A' was used for the VTS compensation instead of 'MFCC_E_D_A_Z', because of VTS assumes this feature theoretically.

Table 3 shows results with and without VTS compensation. Comparing the first and second rows, we can see that the NAM clean acoustic model is still very much affected by acoustic mismatch, despite its noise-proof construction, with a decrease in recognition accuracy from 76.1% to just 15.1%. We then applied VTS noise compensation for the mismatched system. We initialised the noise model parameter $\boldsymbol{\mu}_n, \boldsymbol{\mu}_h$ and $\boldsymbol{\Sigma}_n$ by the first and last 20 frames of each utterance which were assumed to be silence, and the first round of decoding was performed. The hypothesis was then used to update the noise model, and another decoding pass was conducted. The procedure was repeat which gave the final results of VTS in

Table 3. *Results of applying model compensation to the NAM clean acoustic model. 'MFCC_0_D_A' was used for the VTS compensation instead of 'MFCC_E_D_A_Z', because of VTS assumes this feature theoretically.*

	MFCC_0_D_A	
Test data	Compensation.	Word Accuracy (%)
clean data	no	76.1
noise data	no	15.1
noise data	VTS	64.9

Table 3 as 64.9%. This result shows that for NAM microphone, VTS is still very effective in the mismatch condition introduced by noise.

5. CONCLUSIONS AND FUTURE WORK

We have introduced a new corpus of whispered speech recorded simultaneously via a close-talking headset microphone and a non-audible murmur (NAM) microphone, under both clean and noisy conditions. We have provided benchmark recognition results for this corpus.

When the training and test conditions are matched, the NAM microphone was found to be more robust against background noise than the close-talking microphone (word accuracy of 76.3% vs 66.5%). This is consistent to the results reported in literatures. The recognition accuracies of NAM noisy data using a NAM clean model with and without vector Taylor series (VTS) compensation were compared to examine the impact of acoustic mismatch. It was found that, although acoustic mismatch has a very substantial impact on NAM recognition accuracy, VTS compensation could very effectively mitigate this. It implies that VTS is effective even for the body conductive noisy speech to be used for the SSI system, and this is the main contribution of this paper. Future work may include larger scale collection of NAM speech data uttered by many speakers in various types of noise at various SNRs.

6. ACKNOWLEDGEMENTS

The research leading to these results was partly funded from EPSRC grants EP/I031022/1 and EP/J002526/1, the National Natural Science Foundation of China - Royal Society of Edinburgh Joint Project (Grant No. 61111130120) and from seedcorn funding from the Euan MacDonald Centre for Motor Neurone Disease research. We are grateful to Dr Mark Gales and Dr Federico Flego for providing the VTS code. This work was performed whilst the first author was a visitor at the Centre for Speech Technology Research, University of Edinburgh, UK.

7. REFERENCES

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270 – 287, 2010.

[2] P. Heracleous, J. Even, C. Ishi, T. Miyashita, and N. Hagita, "Fusion of standard and alternative acoustic sensors for robust automatic speech recognition," in *Proc. ICASSP*, 2012, pp. 4837–4840.

[3] Y. Nakajima, H. Kashioka, N. Campbell, and K. Shikano, "Non-Audible Murmur (NAM) Recognition," *IEICE - Transactions on Information and Systems*, vol. E89-D, no. 1, pp. 1–4, Jan. 2006.

[4] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell, "Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin," in *Proc. ICASSP*, 2003, pp. 708–711.

[5] T. Toda, K. Nakamura, T. Nagai, T. Kaino, Y. Nakajima, and K. Shikano, "Technologies for processing body-conducted speech detected with non-audible murmur microphone," in *Proc. INTERSPEECH*, Brighton,UK, 2009, pp. 632–635.

[6] P.J. Moreno, B. Raj, and R.M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proc. ICASSP*. IEEE, 1996, vol. 2, pp. 733–736.

[7] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," in *Proc. ICSLP*, 2000, pp. 869–872.

[8] V. Tran, G. Bailly, H. Lovenbruck, and T. Toda, "Improvement to a nam-captured whisper-to-speech system," *Speech Communication*, vol. 52, no. 4, pp. 314 – 326, 2010.

[9] RA Gopinath, MJF Gales, PS Gopalakrishnan, S Balakrishnan-Aiyer, and MA Picheny, "Robust speech recognition in noise—Performance of the IBM continuous speech recogniser on the ARPA noise spoke task," in *Proc. ARPA Workshops Spoken Lang. Syst. Technol.*, 1995, pp. 127–130.

[10] D. Babani, T. Toda, H. Saruwatari, and K. Shikano, "Acoustic model training for non-audible murmur recognition using transformed normal speech data," in *Proc. ICASSP*, 2011, pp. 5224–5227.

[11] H. Zen, T. Nose, J. Yamagishi, S. Sako, and K. Tokuda, "The hmm-based speech synthesis system (hts) version 2.0," in *Proceedings of ISCA Workshop on Speech Synthesis (SSW-6)*, Bonn, Germany, 2007.

[12] Koichi S. and Takao W., "MDL-based context-dependent subword modeling for speech recognition," *Acoustical Science and Technology*, vol. 21, no. 2, pp. 79–86, 2000.

[13] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit," in *Proc. ICSLP*, Denver, Colorado, USA, 2002, pp. 901–904.

[14] K. Richmond, R. Clark, and Sue, "On generating combilex pronunciations via morphological analysis," in *Proc. INTERSPEECH*, Makuhari, Japan, 2010, pp. 1974–1977.