interspeech 2008 incorporating SST 08

Feature Vector Normalization with Combined Standard and Throat Microphones for Robust ASR

Luis Buera, Antonio Miguel, Óscar Saz, Alfonso Ortega, Eduardo Lleida

Communication Technologies Group (GTC), I3A, University of Zaragoza, Spain.
{lbuera,amiguel,oskarsaz,ortega,lleida}@unizar.es

Abstract

We propose on-line unsupervised compensation technique for robust speech recognition that combines standard and throat microphone feature vectors. The solution, called Multi-Environment Model-based LInear Normalization with Throat microphone information, MEMLINT, is an extension of MEMLIN formulation. Hence, standard microphone noisy space and throat microphone space are modelled as GMMs and a set of linear transformations are learnt from data associated to each pair of Gaussians (one for each GMM) using training stereo data. On the other hand, to compensate some kinds of degradation which are not considered in MEMLINT, we propose to use jointly an on-line unsupervised acoustic model adaptation method based on rotation transformations over an expanded HMM-state space (augMented stAte space acousTic dEcoder, MATE). Some experiments with an own recorded database were carried out, showing that the proposed approach significantly outperforms the single microphone approach.

Index Terms: Throat microphone, robust speech recognition, feature vector normalization.

1. Introduction

When training and testing acoustic conditions differ, the accuracy of speech recognition systems rapidly degrades. To compensate this mismatch, classical solutions have been developed, such as feature vector normalization methods or acoustic model adaptation methods. However, another possible approach consists on complementing the standard microphone signal with robust additional signals. For this purpose, cameras [1] and different multi-sensory microphones [2] [3] [4] have been used. In this work, we propose to combine standard and throat microphones signals in an on-line unsupervised technique to provide robustness to ASR systems.

A throat microphone, also laryngophone, records the speech directly through sensors in contact with skin. It is placed closed to the Addams apple and captures the vibrations of the body tissues (skin, bone...). Thus, the main advantage of throat microphone is the robustness in adverse conditions due to its position and characteristics [4]. However, the recorded speech signal has a poor frequency content (low bandwidth), so that important degradation performance is appreciated when it is directly used in ASR. Hence, different kinds of combination with standard microphone signals have been proposed.

In [4], several recombination strategies were considered to estimate the HMM state emission probabilities. A SPLICE extension was developed in [3], where some linear transformations were learnt to model the mismatch between the multi-sensory microphone space and the close talk microphone space. Also the Probabilistic Optimum Filter, POF, technique was modified in [2] by using as source feature vector the concatenation of standard and throat microphone signals. Note that these solutions, which are recommended in very high noise situations e.g. military applications: cockpits of aircrafts, tanks...[5], can have also important applications in those situations where carrying a device is accepted [3], such as removing the push-to-talk button, coding speech with variable-rate...

In this work we propose Multi-Environment Model-based LInear Normalization with Throat microphone information, MEMLINT, which is an extension of MEMLIN [6] formulation to map the noisy feature vectors from the standard and throat microphones to the corresponding feature vectors from the close talk microphone. Furthermore, an unsupervised acoustic model adaptation approach based on rotation transformations is applied over the normalized feature vectors [7].

This paper is organized as follows: a description of the spectral characteristics of the throat microphone signal in comparison with those of the standard microphone signal is studied in Section 2. In Section 3, the proposed MEMLINT extension is presented. In Section 4, the unsupervised acoustic model adaptation approach is briefly explained. The experimental setup and the database are introduced in Section 5, showing the results in Section 6. Finally, the conclusions and future lines are presented in Section 7.

2. Time frequency analysis

Figures 1(a) and 1(b) show the spectrogram of the standard microphone and throat microphone data respectively for a particular utterance in a clean environment. It is observed that the throat microphone presents a low-bandwidth pass filter effect with a cut frequency of less than 4KHz. Also some kind of distortion is included in the speech signal. On the other hand, although the recording is performed in a clean environment, the standard microphone introduces a low frequency noise which does not appear in the throat microphone channel, showing much more robustness.

3. MEMLINT

The proposed Multi-Environment Model-based LInear Normalization with Throat microphone information, MEMLINT, assumes a general MMSE-based framework by providing a GMM modelling of both throat microphone and standard microphone noisy spaces. Therefore, a bias vector transformation is associated with each pair of Gaussians from the throat microphone space and the standard microphone noisy space.

This work has been supported by the national project TIN 2005-08660-C04-01.

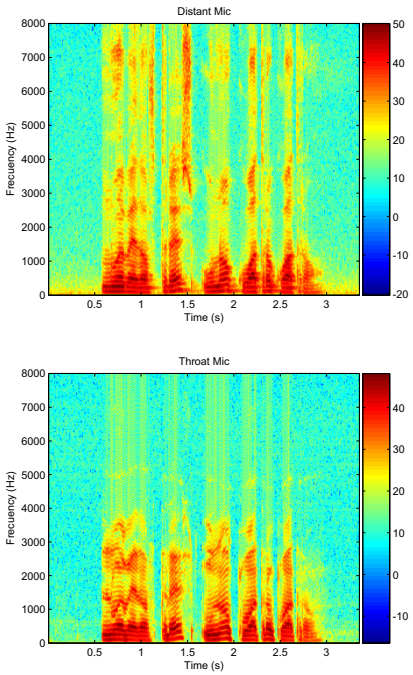

Figure 1: Time frequency representation for one utterance (sequentially from top to bottom): a) recorded with standard microphone. b) recorded with throat microphone.

3.1. MEMLINT approximations

• Standard microphone noisy space is divided into a combination of several basic environments, e, and the standard microphone noisy feature vectors, $\mathbf{y}_t$, are modelled as a GMM for each basic environment ($GMM_{y,e}$)

$$p_e(\mathbf{y}_t) = \sum_{s_y^e} p(\mathbf{y}_t|s_y^e)p(s_y^e), \tag{1}$$

$$p(\mathbf{y}_t|s_y^e) = \mathcal{N}(\mathbf{y}_t; \mu_{s_y^e}, \mathbf{\Sigma}_{s_y^e}), \tag{2}$$

where s_y^e denotes the corresponding Gaussian of the noisy model for the e basic environment, $\mu_{s_y^e}$, $\mathbf{\Sigma}_{s_y^e}$, and $p(s_y^e)$ are the mean vector, the diagonal covariance matrix, and the a priori probability associated with s_y^e.

• Throat microphone feature vectors, $\mathbf{l}_t$, are modelled using a GMM (GMM_l). Observe that in this case it is not necessary to divide the space into basic environments because it can be considered that the feature vectors are not affected by background noise [4].

$$p(\mathbf{l}_t) = \sum_{s_l} p(\mathbf{l}_t|s_l)p(s_l), \tag{3}$$

$$p(\mathbf{l}_t|s_l) = \mathcal{N}(\mathbf{l}_t; \mu_{s_l}, \mathbf{\Sigma}_{s_l}), \tag{4}$$

where s_l denotes the corresponding Gaussian of the model, μ_{s_l}, $\mathbf{\Sigma}_{s_l}$, and $p(s_l)$ are the mean vector, the diagonal covariance matrix, and the a priori probability associated with s_l.

• Standard microphone clean feature vectors, $\mathbf{x}_t$, can be approximated as a linear function of $\mathbf{y}$ which depends on s_l and s_y^e: $\mathbf{x}_t \approx \Psi(\mathbf{y}_t, s_l, s_y^e) = \mathbf{y}_t - \mathbf{r}_{s_l,s_y^e}$, where $\mathbf{r}_{s_l,s_y^e}$ is the bias vector transformation for each pair of Gaussians, s_l and s_y^e.

3.2. MEMLINT enhancement

With those approximations, the MMSE estimation for $\mathbf{x}_t$ is

$$\hat{\mathbf{x}}_t = \mathbf{y}_t - \sum_e \sum_{s_y^e} \sum_{s_l} \mathbf{r}_{s_l,s_y^e} p(e|\mathbf{y}_t)p(s_y^e|\mathbf{y}_t, e)p(s_l|\mathbf{l}_t), \tag{5}$$

where $p(e|\mathbf{y}_t)$ is the a posteriori probability of the basic environment; $p(s_y^e|\mathbf{y}_t, e)$ is the a posteriori probability of the noisy model Gaussian s_y^e, given the feature vector $\mathbf{y}_t$ and the basic environment, e. Those two terms are computed for each frame applying (1) and (2) as it can be observed in [6]. On the other hand, $p(s_l|\mathbf{l}_t)$, which is the probability of the Gaussian s_l, given the feature vector $\mathbf{l}_t$, can be obtained directly with (3) and (4)

$$p(s_l|\mathbf{l}_t) = \frac{p(s_l)\mathcal{N}(\mathbf{l}_t; \mu_{s_l}, \mathbf{\Sigma}_{s_l})}{\sum_{s_l} p(s_l)\mathcal{N}(\mathbf{l}_t; \mu_{s_l}, \mathbf{\Sigma}_{s_l})}. \tag{6}$$

Finally, the bias vector transformation, $\mathbf{r}_{s_l,s_y^e}$, is estimated in a previous training phase using stereo data.

3.3. MEMLINT training

In order to estimate the bias vector transformation, a stereo data corpus for each basic environment is needed, $(\mathbf{X}_e, \mathbf{Y}_e, \mathbf{L}_e) = \{(\mathbf{x}_1^e, \mathbf{y}_1^e, \mathbf{l}_1^e); ...; (\mathbf{x}_{t_e}^e, \mathbf{y}_{t_e}^e, \mathbf{l}_{t_e}^e); \quad ...; (\mathbf{x}_{T_e}^e, \mathbf{y}_{T_e}^e, \mathbf{l}_{T_e}^e)\}$, with $t_e = 1, ..., T_e$. Observe that the stereo database is obtained by simultaneous recording of clean standard (close talk), noisy standard (far field) and throat microphones signals. Thus, $\mathbf{r}_{s_l,s_y^e}$, is estimated by minimizing the defined mean weighted square error, ξ_{s_l,s_y^e}, (7) with respect to $\mathbf{r}_{s_l,s_y^e}$ (8). Note that $p(s_l|\mathbf{l}_{t_e}^e)$ and $p(s_y^e|\mathbf{y}_{t_e}^e, e)$ can be estimated directly applying (1), (2) and (3), (4), respectively.

Comparing the proposed approach with respect to the conventional MEMLIN framework [6], it can be observed that no GMM is used to model the close talk microphone feature vectors. Instead of that, GMM_l is applied, and, a result of it, the cross-probability model, $p(s_x|s_y^e, e, \mathbf{y}_t) \simeq p(s_x|s_y^e, e)$, which is the probability of the clean model Gaussian, s_x, given the noisy model one and the basic environment, [6] does not have to be estimated in the previous training process.

3.4. Simple MEMLINT

Other option to include the throat microphone information in MEMLIN approach could be to substitute $GMM_{y,e}$ in the conventional MEMLIN framework by GMM_l, providing a similar solution to [3]. In that case the throat microphone feature vectors would be mapped directly to the close talk microphone feature vector space without using $\mathbf{y}_t$

$$\hat{\mathbf{x}}_t = \mathbf{y}_t - \sum_{s_x} \sum_{s_l} \mathbf{r}_{s_l,s_x} p(s_l|\mathbf{l}_t)p(s_x|s_l, \mathbf{l}_t). \tag{9}$$

Observe that a GMM is used to model the close talk microphone feature vectors and $\mathbf{x}_t$ can be approximated as $\mathbf{x}_t \approx \Psi(\mathbf{y}_t, s_l, s_x) = \mathbf{y}_t - \mathbf{r}_{s_l,s_x}$. On the other hand, the corresponding cross-probability model, $p(s_x|s_l, \mathbf{l}_t) \simeq p(s_x|s_l)$ and the bias vector transformation, $\mathbf{r}_{s_l,s_x}$, should be estimated in the previous training process with stereo database, $(\mathbf{X}_e, \mathbf{L}_e)$, in a similar way as [6]. As it will be exposed, the results for this solution are not so good as the ones obtained with the first approach. For simplicity, we refer to this approach as Simple MEMLINT (S-MEMLINT).

4. Acoustic model adaptation

Observe that MEMLINT estimates the clean feature vector by using a bias vector transformation (5), not taking into account

$$\xi_{s_l,s_y^e} = \sum_{t_e} p(s_l|\mathbf{l}^e_{t_e}) p(s^e_y|\mathbf{y}^e_{t_e}, e) Tr[(\mathbf{x}^e_{t_e} - \mathbf{y}^e_{t_e} + \mathbf{r}_{s_x,s^e_y})(\mathbf{x}^e_{t_e} - \mathbf{y}^e_{t_e} + \mathbf{r}_{s_x,s^e_y})^T], \quad (7)$$

$$\mathbf{r}_{s_l,s^e_y} = \underset{\mathbf{r}_{s_l,s^e_y}}{arg\,min}(\xi_{s_l,s^e_y}) = \frac{\sum_{t_e} p(s_l|\mathbf{l}^e_{t_e}) p(s^e_y|\mathbf{y}^e_{t_e}, e)(\mathbf{y}^e_{t_e} - \mathbf{x}^e_{t_e})}{\sum_{t_e} p(s_l|\mathbf{l}^e_{t_e}) p(s^e_y|\mathbf{y}^e_{t_e}, e)}, \quad (8)$$

several kinds of degradation, like rotations or variance deformations. To compensate these effects, in this work we propose to use an on-line unsupervised acoustic model adaptation method based on rotation transformations over an expanded HMM-state space (augMented stAte space acousTic dEcoder, MATE [8]).

Thus, normalized and close talk microphone feature vectors ($\hat{\mathbf{x}}$ and $\mathbf{x}$, respectively) are modelled as both GMMs, where the corresponding components are defined by $s_{\hat{x}}$ and s_x, respectively. Furthermore, a rotation matrix $\mathbf{A}_{s_{\hat{x}},s_x}$ is defined to represent the mismatch between clean and normalized feature vectors for each pair of Gaussians $s_{\hat{x}}$ and s_x as $\hat{\mathbf{x}}_t \approx \mathbf{A}_{s_{\hat{x}},s_x}\mathbf{x}_t$, where $\mathbf{A}_{s_{\hat{x}},s_x}$ is estimated by regression in an unsupervised training process. In order to determine the corresponding rotation matrix for each normalized feature vector in decoding, ML maximization criterion is used in the Viterbi decoding process by expanding the corresponding acoustic models using all the rotation matrices. To know the details of the technique (estimation of the rotation matrices, expansion of the acoustic models...), the authors recommend reading [7].

5. Experimental setup

In order to carry out the experiments, a Spanish database was collected in clean environment with 20 speakers (10 male and 10 female). It contains records of sequences of rich phonetically balanced sentences (50 utterances per speaker) and connected and isolated digits (63 utterances per speaker), with a close talk, a far field and a throat microphones. The three signals were recorded simultaneously in a synchronized way. In order to build the different acoustic environments, real car noise was added in an artificially way to the far field microphone signals with different SNR: 20dB, 15dB, 10dB, 5dB and 0dB.

As feature set, the standard ETSI front-end [9] features plus energy and the corresponding delta and delta delta coefficients are used. Cepstral mean normalization is applied to testing and training data. The different feature vector normalization techniques are applied to the 12 MFCCs and energy, whereas the derivatives are computed over the normalized static coefficients. The acoustic models are composed of 16 state HMM for each digit, a 3 state begin-end silence HMM and a 1 state inter-word silence HMM. In all cases, each pdf state is composed by a mixture of three Gaussians. Due to the small recorded database, the acoustic models are built with the close talk microphone (clk) signals of Spanish SpeechDat Car database [10]. Two different kinds of experiments were carried out if transcriptions for training data are used (supervised) or not (unsupervised).

6. Results

6.1. Unsupervised experiments

In order to define the testing partitions, leave-one-out technique is applied. So, 5 different testing partitions are created by the digits utterances of 4 speakers (two males and two females), while the corresponding training partitions are composed by the rich phonetically balanced utterances of the other different 16 speakers. Also, the training partitions do not include 0dB SNR signals. Observe that the training task is not the same as the testing one, and 0dB SNR condition remains as an unseen training condition.

Table 1: Average WER, AWER, baseline results, in %, from the different acoustic conditions and microphones.

Train	Test	AWER (%)
clkm	ffm clean	2.19
clkm	ffm 20dB	18.25
clkm	ffm 15dB	33.86
clkm	ffm 10dB	55.96
clkm	ffm 5dB	79.66
clkm	ffm 0dB	89.40
clkm	thm	35.77

The average WER, AWER, baseline results for the 5 testing partitions are included in Table 1, where ffm corresponds to the far field microphone, while thm is the throat microphone. It can be appreciated how the throat microphone, although does not capture the background noise, produces a poor speech recognition performance due to the frequency limitations of the recorded signal, confirming the results presented in [4].

Now, four robustness approaches are compared.

- Conventional MEMLIN approach [6], where ffm feature vectors are mapped into the close talk microphone feature space. Note that no throat microphone is used.
- Proposed S-MEMLINT extension, where the thm feature vectors are mapped directly into the close talk microphone feature space without considering ffm feature vectors.
- Proposed MEMLINT, where thm and ffm feature vectors are used jointly to obtain an estimation of the corresponding close talk feature vector.
- Proposed MEMLINT with unsupervised acoustic model adaptation technique based on rotation matrices. It is identified as MEMLINT-HMMadapt.

The average WER results are depicted in Fig. 2. In all of them, 128 Gaussians were used to model the different environments per microphone. Also, 17 rotation matrices are estimated for MEMLINT-HMMadapt technique ($\#s_x = \#s_{\hat{x}} = 4$, plus the identity matrix). It can be verified the important improvement that the presented approaches obtain. Conventional MEMLIN approach produces a consistent improvement, although it is not as satisfactory in unseen conditions (0dB). However, in moderate degraded acoustic environments (SNR>5dB), the performance is superior with respect to the one obtained when just the throat microphone signals are used (Train clk, Test thm in Table 1). On the other hand, S-MEMLINT reaches an important improvement (24.33% AWER) with respect to the results obtained just with throat microphone signals. Note that the results are SNR-invariant since the throat microphone does not capture background noise. However, better results for

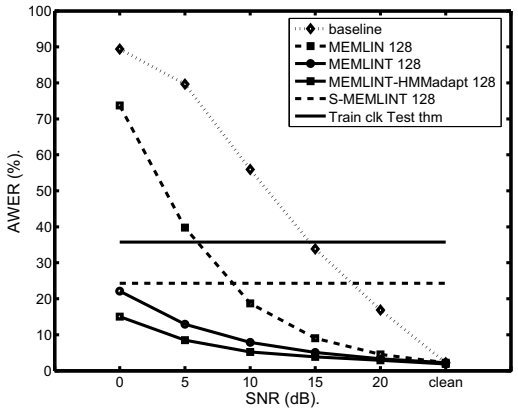

Figure 2: Average WER, AWER, in % for the four MEMLIN approaches considered: MEMLIN, S-MEMLINT, MEMLINT and MEMLINT with acoustic model adaptation based on rotation matrices.

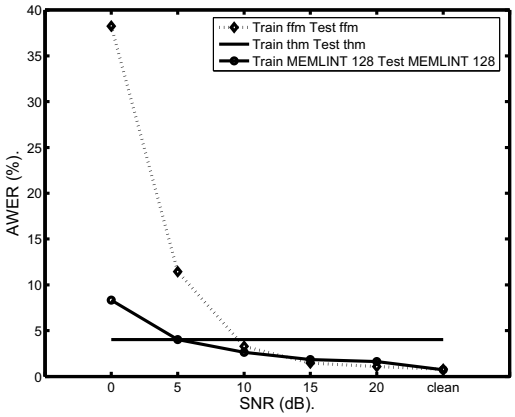

Figure 3: Average WER, AWER, in % for the supervised approaches.

all SNR based environments are obtained if throat and standard noisy microphones are combined (MEMLINT). Finally, if MEMLINT approach is combined with acoustic model adaptation based on rotation matrices, impressed AWER are obtained for all conditions. Thus we can conclude that the combination of throat and noisy microphones feature vectors provides a more satisfactory feature vector mapping than if just throat microphone (S-MEMLINT) or standard noisy microphones (MEMLIN) are used, reducing dramatically the AWER. Note that even feature vectors of unseen acoustic environments are compensated in a proper way, providing a very good performance.

6.2. Supervised experiments

In this case the transcription of the training data is used to train new acoustic models. Thus, while the testing partitions are the same as the ones considered in unsupervised experiments, the training ones differ. Since the used acoustic model units in this work are words, the new training partitions for this experiment are composed by digit task utterances. Observe that the training task is the same as the testing one, although 0dB SNR condition remains as an unseen training condition.

The average WER results are shown in Fig. 3. It can be observed that very competitive results can be obtained if acoustic models are trained with the feature vectors (4.02% AWER, Train thm Test thm). On the other hand, a better behavior is obtained in moderate noisy environments if acoustic models for all available SNRs are trained with ffm feature vectors (multicondition training, Train ffm Test ffm). However, in most adverse and unseen environments, the performance is not as satisfactory. Finally, the most competitive results in average for all the SNRs environments are reached when MEMLINT approach is applied with matched acoustic models (Train MEMLINT 128 Test MEMLINT 128). To train the new acoustic models all available training ffm feature vectors are normalized using MEMLINT approach and ML technique is applied. In all cases 128 Gaussians per different environments and microphones are used.

7. Conclusions

In this paper we have presented Multi-Environment Model-based LInear Normalization with Throat microphone information, MEMLINT, an on-line unsupervised compensation approach which combines standard and throat microphone feature vectors. This solution, which can be seen as an extension of MEMLIN formulation, is combined with an acoustic model adaptation technique based on rotation transformations to compensate jointly the shift and rotation introduced by the acoustic environment. Some results with an own recorded database show the effective performance of the proposed technique with respect to single microphone robustness techniques, obtaining very competitive results even in unseen conditions. As future lines we propose to develop a non stereo data training process and to carry out new experiments with different kinds of noise.

8. References

[1] T. Chen and R. R. Rao, "Audio-visual integration in multimodal communication," *IEEE*, vol. 86, no. 5, pp. 837–852, May 1998.

[2] M. Graciarena, H. Franco, K. Sommez, and H. Bratt, "Combining standard and throat microphones for robust speech recognition," *IEEE Signal Processing Letters*, vol. 10, no. 3, pp. 72–74, March 2003.

[3] Z. Zhang, Z. Liu, M. Sinclair, A. Acero, L. Deng, J. Droppo, X. Huang, and Y. Zheng, "Multi-sensory microphones for robust speech detection, enhancement and recognition," in *ICASSP*, May 2004, vol. 3, pp. 781–784.

[4] S. Dupont, C. Ris, and D. Bachelart, "Combined use of close-talk and thoat microphones for improved speech recognition under non-stationary background noise," in *Workshop (ITRW) on robustness issues in conversational interaction*, August 2004.

[5] B. Acker-Mills, A. Houtsma, and W. Ahroon, "Speech intelligibility with acoustic and contact microphones," Tech. Rep., Army Aeromedical Research Lab Fort Rucker AL, April 2005.

[6] L. Buera, E. Lleida, A. Miguel, A. Ortega, and O. Saz, "Cepstral vector normalization based on stereo data for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 15, pp. 1098–1113, March 2007.

[7] L. Buera, A. Miguel, E. Lleida, O. Saz, and A. Ortega, "Robust speech recognition with on-line unsupervised acoustic feature compensation," in *in Proc. ASRU*, Dic. 2007.

[8] A. Miguel, E. Lleida, R. Rose, L. Buera, O. Saz, and A. Ortega, "Capturing local variability for speaker normalization in speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 16, no. 3, pp. 578–593, March 2008.

[9] ETSI, "Speech processing transmission and quality aspects (stq); distributed speech recognition; front-end feature extraction algorithm; compression algorithms," Tech. Rep., ETSI ES 201 108 version 1.1.2, April 2000.

[10] Henk van den Heuvel, Jerme Boudy, Robrecht Comeyne, Stephan Euler, Asuncion Moreno, and G. Richard, "The speechdat-car multilingual speech databases for in-car applications: some first validation results," in *Proceedings of Eurospeech*. Budapest, Hungary, Sept. 1999, vol. 5, pp. 2279–2282.