Matching the Acoustic Model to Front-End Signal Processing for ASR in Noisy and Reverberant Environments

Roland Maas, Andreas Schwarz, Klaus Reindl, Yuanhang Zheng,
Stefan Meier, Armin Sehr, Walter Kellermann

Multimedia Communications and Signal Processing, University of Erlangen-Nuremberg, Erlangen, Germany

{*maas,schwarz,reindl,zheng,smeier,sehr,wk*}*@LNT.de*

Introduction

Distant-talking automatic speech recognition (ASR) represents an extremely challenging task. The major reason is that unwanted additive interference and reverberation are picked up by the microphones besides the desired signal. A hands-free human-machine interface should therefore comprise a powerful acoustic preprocessing unit in line with a robust ASR back-end. However, since perfect speech enhancement cannot be achieved in practice, the output of the front-end will always contain some residual interference and some distortion of the desired signal. It is hence of decisive importance to carefully adjust the hidden Markov models (HMMs) of the ASR system to the front-end. In this contribution, we present a two-channel acoustic front-end based on blind source separation along with Wiener filtering. For the front-end integration into the ASR system, different types of multi-style as well as adaptive training and HMM adaptation are investigated.

Acoustic Front-End

The acoustic front-end consists of two units. At first, the coefficients of a blocking matrix are determined by minimizing the TRINICON cost function for blind source separation [1]. Assuming the position of the desired speaker to be known, the cost function can be extended by a directional constraint such that it aims at separating the desired speech from the noise and interference components. The latter ones are then exploited to perform single-channel Wiener filtering in both microphone channels. The estimate of desired speech signal is finally fed into the ASR back-end. We refer to [2] for a detailed description of the acoustic front-end as well as its particular configuration for the given scenario.

Acoustic Model Matching

The integration of the front-end into the ASR back-end represents a crucial element in the system design. Besides the choice of the training data, the training procedure itself has a significant impact on the overall system performance. In the following, we focus on three different techniques of deriving the recognizer's acoustic model: multi-style training, mean adaptation as well as adaptive training.

A straightforward approach to train the acoustic model is to assemble a training set comprising different kinds of speakers, reverberation, or noise conditions. This so-called "multi-style" or "multi-condition" training is an efficient way to increase the robustness of a recognizer to different scenarios. A potential drawback of this method is, however, that the discrimination capability of the acoustic model might suffer from the wide range of conditions covered by the training set.

To cope with this problem, one could think of adapting the means of a multi-style HMM network to a particular speaker, reverberation, or noise condition. One of the most widely used adaptation techniques is to perform MLLR (Maximum Likelihood Linear Regression) and MAP (Maximum A-Posteriori estimation) on the mean values of an HMM network [3]. For each mean value μ, the multi-affine transform $\tilde{\mu} = A\mu + b$ with parameters A, b is applied, followed by a Bayesian estimation of the a-posteriori mean $\widehat{\mu} = \frac{N}{N+\tau}\bar{\mu} + \frac{\tau}{N+\tau}\tilde{\mu}$, where $\bar{\mu}$ is the maximum likelihood mean estimate on the adaptation data, N the occupation likelihood of the considered Gaussian component, and τ a hyperparameter controlling the adaptation speed [3]. The performance of such a mean adaptation might nevertheless be limited since all other parameters of the HMM network arose from the multi-style training procedure.

The idea of "adaptive training" aims at deriving a canonical acoustic model. This can be realized by, e.g., estimating a CMLLR (Constrained Maximum Likelihood Linear Regression) transform for each speaker, reverberation, or noise condition: $\widehat{x} = Ax + b$, where $\widehat{x}$ and x denote a transformed and the corresponding original feature vector, respectively, while matrix A and vector b capture the mapping parameters. These specific feature-domain transforms are then incorporated into the standard Baum-Welch training procedure in order to neutralize the influence of the particular condition.

Experimental Setup

We consider a scenario generalizing the PASCAL CHiME challenge [4] in order to investigate the effect of changing reverberation conditions. To this end, binaural room impulse responses are not only measured at a distance of 2 but also of 1 and 3 meters. The impulse responses are recorded with a binaural manikin in broadside direction in a room with a reverberation time T_{60} of 300 ms and convolved with the utterances of 10 different speakers out of the GRID corpus [5]. Each GRID utterance is of the form <command-color-preposition-letter-number-adverb>, where only the *keyword accuracy* is of interest for the ASR task, which is based on the number of correctly recognized <letter> and <number> tokens. We

distinguish three types of data sets:

1.) A "noise-free" set contains only artificially reverberated utterances.

2.) A "noisy" set consists of reverberated utterances mixed with binaural background noise recordings from the CHiME domestic audio corpus [4] at Signal-to-Noise-Ratio (SNR) levels of $-6, -3, 0, 3, 6$, and 9 dB.

3.) A "processed" set is obtained by applying the proposed acoustic front-end to the corresponding noisy set.

For recognition, we employed the ASR toolkit HTK [3] with word-level HMMs using 7 Gaussian densities per state. From the input signals, features consisting of 13 mel-frequency cepstral coefficients, including the 0'th, as well as 13 delta and 13 acceleration coefficients are derived. Furthermore, cepstral mean normalization is applied. The training set consists of 470 utterances per desired speaker, microphone distance, and SNR level.

Experimental Results

Table 1 shows the keyword accuracies for different test and training conditions.

In the case of an acoustic model trained on noise-free data, we observe a severe degradation in ASR performance for noisy test data. Employing the proposed acoustic front-end reduces the word error rate (WER) by almost 20%.

Training the acoustic model on noisy data is obviously counterproductive if the front-end processing is performed. This observation is consistent since the Wiener filters have intentionally been configured to significantly reduce the noise components while allowing the introduction of some artifacts. As to be expected, the noisy training procedure seems, however, to be extremely beneficial for the case of unprocessed noisy test data leading to a WER improvement of more than 66% compared to the noise-free acoustic model.

A remarkable WER improvement of another 35 % is achieved by passing both the test and the training data through the front-end while training the acoustic model in multi-style mode. Moreover, front-end processing along with specifically adapted or adaptively trained models allow by far for the highest recognition rates. It is, however, interesting to note that this benefit is due to the speaker adaptation. The additional distinction w.r.t. the speaker distance and the SNR level seems to be ineffective for the given scenario.

We conclude from these results that the back-end training procedure employed in [2] consisting of multi-style training with processed data along with MLLR+MAP speaker adaptation is well justified. The adaptation of other HMM parameters than the mean values as well as the explicit consideration of the speaker distance only slightly impact the ASR performance. Similarly, SNR-specific acoustic models do not further improve the recognition results, which might be due to the fact that different noise sources coinciding in SNR can exhibit extremely different feature-domain representations. For future work, it could therefore be promising to investigate noise-specific acoustic models based on a feature domain measure in combination with automatic model selection.

test data	training data	training mode	WA
noise-free	noise-free	multi-style	98.3
noisy	noise-free	multi-style	59.5
processed	noise-free	multi-style	67.2
processed	noisy	multi-style	65.8
noisy	noisy	multi-style	86.8
processed	processed	multi-style	91.4
processed	processed	MLLR+MAP[1]	95.2
processed	processed	MLLR+MAP[2]	95.2
processed	processed	MLLR+MAP[3]	94.8
processed	processed	adaptive[1]	95.1
processed	processed	adaptive[2]	95.2
processed	processed	adaptive[3]	95.3

[1] for each speaker across all distances and SNR levels.
[2] for each speaker and distance across all SNR levels.
[3] for each speaker, distance, and SNR level.

Table 1: Comparison of keyword accuracies (WA) in %.

Acknowledgement

The authors would like to thank the Deutsche Forschungsgemeinschaft (DFG) for supporting this work (contract number KE 890/4-1).

References

[1] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of a class of blind source separation algorithms for convolutive mixtures," in *Proc. IEEE Int. Symposium on Independent Component Analysis and Blind Signal Separation (ICA)*, 2003, pp. 945–950.

[2] R. Maas, A. Schwarz, Y. Zheng, K. Reindl, S. Meier, A. Sehr, and W. Kellermann, "A two-channel acoustic front-end for robust automatic speech recognition in noisy and reverberant environments," in *Proc. Int. Workshop on Machine Listening in Multisource Environments (CHiME)*, 2011, pp. 41–46.

[3] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*. University of Cambridge, 2009.

[4] H. Christensen, J. Barker, and P. Green, "The CHiME corpus: a resource and a challenge for Computational Hearing in Multisource Environments," in *Proc. Interspeech*, 2010, pp. 1918–1921.

[5] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.