Model-Combination-Based Acoustic Mapping

Martin Westphal [*], *Alex Waibel*

Interactive Systems Laboratories – University of Karlsruhe (Germany) / Carnegie Mellon University (USA) – waibel@cs.cmu.edu

*) now with: European Speech Research – IBM Deutschland Entwicklung GmbH – westphal@de.ibm.com

ABSTRACT

We propose a new method for compensating distortions in the speech signal caused by environment changes. The basic method concentrates on additive noise, but can be extended to address also channel and to some extend speaker changes. By combining compensation with adaptation techniques it leads to high error rate reductions for mobile speech applications. Thereby, it is more efficient than adapting the acoustic model of the recognizer and more powerful than simple noise reduction techniques.

1. INTRODUCTION

In [7] we presented a speech based system that allows spontaneous queries to a navigation and information data-base. The first prototype did a good job for relatively noise free environments but was not very robust. However, this kind of application would be extremely useful especially for mobile environments like in a car or a portable device. Due to the changing environment the recognizer has to deal with distortions such as additive noise at different levels and channel changes. Also sudden speaker changes make recognition harder, but we would like to allow fast and multi-speaker access to the information provided by such a system.

Therefore, increasing the acoustic robustness of speech recognition is an important issue that we like to address in this paper. Several methods have been proposed in the past, and we want to follow a quite common categorization into three major approaches: robust features, compensation and model adaptation. The gain by using more sophisticated feature extraction is somehow limited, so we stayed with Linear Discriminant Analysis (LDA) based on widely used cepstral features. Utterance based adaptation using MLLR [8], MAP [6] or PMC [5] is computationally costly since the acoustic model of the recognizer must be modified (80,000 Gaussians in our case) and for MLLR and MAP, also a first classification run is required. Compensation methods are more efficient but often address only one kind of distortion (using an environment assumption like stationary, additive noise, constant channel or linear frequency warp due to different vocal tract lengths). After investigating some promising compensation methods, we developed a new approach that is also efficient and performed significantly better.

2. MOTIVATION AND TECHNIQUES

2.1 Review Of 2DCMS

Before we propose the new method that we called **MAM** (model-combination-based acoustic mapping) we review **2DCMS** (2-level delta cepstral mean subtraction) [11] which is quite simple but nevertheless effective in removing channel and noise distortions. Similar to the new method, it uses an acoustic model as well as an environment model. The latter one is mathematically described in the spectral domain (superscript s) as

$$\tilde{x}_i^{\,\mathrm{s}} = \mathrm{h}_i^{\,\mathrm{s}} \cdot x_i^{\,\mathrm{s}} + \mathrm{n}_i^{\,\mathrm{s}} \qquad (2.1)$$

The undisturbed spectral coefficient $x_i^{\,\mathrm{s}}$ is distorted by the channel $\mathrm{h}_i^{\,\mathrm{s}}$ and additive noise $\mathrm{n}_i^{\,\mathrm{s}}$. Only the distorted result $\tilde{x}_i^{\,\mathrm{s}}$ is available to the front-end of a recognizer. Considering only two classes, namely dominating speech ($\mathrm{h}_i^{\,\mathrm{s}} \cdot x_i^{\,\mathrm{s}} >> \mathrm{n}_i^{\,\mathrm{s}}$) and speech pauses ($\mathrm{h}_i^{\,\mathrm{s}} \cdot x_i^{\,\mathrm{s}} = 0$) we find for the log-spectral domain (superscript l)

$$\text{Speech:} \qquad \tilde{x}_i^{\,\mathrm{l}} \approx \mathrm{h}_i^{\,\mathrm{l}} + x_i^{\,\mathrm{l}} \qquad (2.2)$$

$$\text{Pause:} \qquad \tilde{x}_i^{\,\mathrm{l}} = \mathrm{n}_i^{\,\mathrm{l}} \qquad (2.3)$$

Assuming that $x_i^{\,\mathrm{l}}$ has a zero mean (because any offset can be seen as a kind of channel caused by the speaker), we can calculate the channel and noise mean from the signal using a speech-pause decision. 2DCMS then tries to restore the channel and noise level to the values found in the training environment. We describe this simple acoustic model with the two mean vectors (prototypes) for channel h and noise n

$$\lambda = \{\mu_\mathrm{h}, \mu_\mathrm{n}\} \qquad (2.4)$$

The current situation, regardless of whether we consider an utterance for training or recognition, is described with

$$\tilde{\lambda} = \{\tilde{\mu}_\mathrm{h}, \tilde{\mu}_\mathrm{n}\} \qquad (2.5)$$

Using the correspondence between the prototypes of the two classes, an estimate $\hat{\mathbf{x}}^\mathrm{l}[k]$ for clean speech at time frame k is derived using the delta means weighted by an (estimated) probability for the class.

$$\begin{aligned}\hat{\mathbf{x}}^\mathrm{l}[k] = \tilde{\mathbf{x}}^\mathrm{l}[k] &- \mathrm{P}\left(\text{Speech}\left|\tilde{\mathbf{x}}^\mathrm{l}[k]\right.\right) \cdot \left(\tilde{\mu}_\mathrm{h} - \mu_\mathrm{h}\right) \\ &- \mathrm{P}\left(\text{Pause}\left|\tilde{\mathbf{x}}^\mathrm{l}[k]\right.\right) \cdot \left(\tilde{\mu}_\mathrm{n} - \mu_\mathrm{n}\right)\end{aligned} \qquad (2.6)$$

2.2 Basic Concept of MAM

Using delta means in the log-spectral domain is most often identified with channel normalization. [10] uses Gaussian mixtures simultaneously trained on stereo data to perform an acoustic mapping between two different channels. In 2DCMS the means are also used to compensate additive noise using very coarse environment assumptions and thus a very simple model. The new method MAM concentrates on compensating the additive noise but – as we will see later – can also cope with a channel mismatch. It uses a more precise model than 2DCMS, that is a Gaussian mixture for clean speech with mean vectors $\boldsymbol{\mu}_m$, covariance matrices $\mathbf{C}_m$ and priors $\hat{\mathrm{P}}(m)$:

$$\lambda = \left\{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_M, \mathbf{C}_1,\ldots,\mathbf{C}_M, \hat{\mathrm{P}}(1),\ldots,\hat{\mathrm{P}}(M)\right\} \tag{2.7}$$

In a first step, the distortion caused by additive noise is simulated using model combination. Then, in the second step, an acoustic mapping is performed that compensates the distortion in the signal (feature vectors).

2.3 Model Combination

In order to obtain more complex acoustic models of generic speech with corresponding prototypes, we make use of the well known technique of model combination. This technique was developed to adapt the acoustic models of the speech recognizer. However, applying it to complex features used by most recognizers makes it very difficult and expensive or restricts the kind of preprocessing that can be used to compute the input features. At this point, we aim only at simulating the effect caused by additive noise by means of a secondary, generic acoustic model. We call it secondary model since it is independent of the recognizer's model that often distinguishes between thousands of phonetic classes.

Therefore, given a Gaussian mixture model representing clean speech we have to provide an additional model representing noise. Both models can be combined under the assumption that speech and noise are statistically independent. As we are free to choose the model space (domain) for the secondary model we decided to use log-spectral features (to be more precise: *Mel*-frequency spectral coefficients MFSC). As with cepstral features we can then apply utterance or speaker based mean subtraction in order to remove channel dependencies and thus keep the number of model parameters low.

Since the actual combination is done in the spectral domain (due to the additive relation between speech and noise in this domain), the clean speech model and the noise model have to be transferred into this domain. After the combination, the resulting model for noisy speech has to be transferred back into the log-spectral domain to obtain a model that corresponds with the original model for clean speech. This can be done using the log-normal approximation given for example in [4]. The prototypes of the models (the mean vectors of the Gaussians) will correspond as we use a single Gaussian density to represent the noise characteristic (mean and variance of the noise).

Below, we give the equations needed to derive the model for noisy speech $\tilde{\lambda}^{\mathrm{l}}$ given the secondary model λ^{l} for clean speech and a noise model $\lambda_{\mathbf{n}}{}^{\mathrm{l}}$. The superscript "l" indicates that we are in the log-spectral domain, whereas "s" stands for the spectral domain. $\boldsymbol{\mu}$ is the mean vector of a Gaussian, $\mathbf{C}$ the covariance matrix. μ_{i} and σ_{ij} are their components.

From log-spectral to spectral domain:

$$\lambda_{\mathbf{n}}{}^{\mathrm{l}} \rightarrow \lambda_{\mathbf{n}}{}^{\mathrm{s}};\ \lambda^{\mathrm{l}} \rightarrow \lambda^{\mathrm{s}} \tag{2.8}$$

$$\mu_{\mathrm{i}}{}^{\mathrm{s}} = e^{\mu_{\mathrm{i}}{}^{\mathrm{l}} + \frac{1}{2}\sigma_{\mathrm{ii}}{}^{\mathrm{l}}} \qquad \sigma_{\mathrm{ij}}{}^{\mathrm{s}} = \mu_{\mathrm{i}}{}^{\mathrm{s}} \cdot \mu_{\mathrm{j}}{}^{\mathrm{s}} \cdot \left(e^{\sigma_{\mathrm{ij}}{}^{\mathrm{l}}} - 1\right)$$

Combination in the spectral domain:

$$\lambda^{\mathrm{s}}, \lambda_{\mathbf{n}}{}^{\mathrm{s}} \rightarrow \tilde{\lambda}^{\mathrm{s}} \tag{2.9}$$

$$\tilde{\boldsymbol{\mu}}^{\mathrm{s}} = \boldsymbol{\mu}^{\mathrm{s}} + \boldsymbol{\mu}_{\mathbf{n}}{}^{\mathrm{s}} \qquad \tilde{\mathbf{C}}^{\mathrm{s}} = \mathbf{C}^{\mathrm{s}} + \mathbf{C}_{\mathbf{n}}{}^{\mathrm{s}}$$

From spectral to log-spectral domain:

$$\tilde{\lambda}^{\mathrm{s}} \rightarrow \tilde{\lambda}^{\mathrm{l}} \tag{2.10}$$

$$\mu_{\mathrm{i}}{}^{\mathrm{l}} = \ln \mu_{\mathrm{i}}{}^{\mathrm{s}} - \frac{1}{2}\ln\left(\frac{\sigma_{\mathrm{ii}}{}^{\mathrm{s}}}{\left(\mu_{\mathrm{i}}{}^{\mathrm{s}}\right)^2} + 1\right) = \ln\left(\frac{\left(\mu_{\mathrm{i}}{}^{\mathrm{s}}\right)^2}{\sqrt{\left(\mu_{\mathrm{i}}{}^{\mathrm{s}}\right)^2 + \sigma_{\mathrm{ii}}{}^{\mathrm{s}}}}\right)$$

$$\sigma_{\mathrm{ij}}{}^{\mathrm{l}} = \ln\left(\frac{\sigma_{\mathrm{ij}}{}^{\mathrm{s}}}{\mu_{\mathrm{i}}{}^{\mathrm{s}} \cdot \mu_{\mathrm{j}}{}^{\mathrm{s}}} + 1\right)$$

2.4 Acoustic Mapping

The assumption underlying Cepstral Mean Subtraction (CMS) is that the channel is constant over time. Therefore it is sufficient to find a constant correction vector and to subtract it from the distorted input samples in the log-spectral domain. As with the 2DCMS, if we consider additive noise in this domain, even if it is stationary, the correction vector will change over time k and depends on the clean speech sample $\mathbf{x}[k]$. For the distorted sample we can write:

$$\tilde{\mathbf{x}}[k] = \mathbf{x}[k] + \boldsymbol{\Delta}(\mathbf{x}[k]) \tag{2.11}$$

As an extension of the 2DCMS we can think of two corresponding models for clean and noisy speech, each consisting of M prototypes. For the prototypes we assume the same relationship as for the speech samples:

$$\tilde{\boldsymbol{\mu}}_m = \boldsymbol{\mu}_m + \boldsymbol{\Delta}(\boldsymbol{\mu}_m) \tag{2.12}$$

We can now make the following approximation to obtain a MMSE estimate (minimum mean square error) for the correction vector:

$$\begin{aligned}
\hat{\mathbf{x}} &= E\{\mathbf{x} \mid \tilde{\mathbf{x}}\} = \int_{\mathbf{x}} \mathbf{x} \cdot \mathrm{p}(\mathbf{x} \mid \tilde{\mathbf{x}})\,\mathbf{dx} \\
&= \tilde{\mathbf{x}} - \int_{\mathbf{x}} \boldsymbol{\Delta}(\mathbf{x}) \cdot \mathrm{p}(\mathbf{x} \mid \tilde{\mathbf{x}})\,\mathbf{dx} \\
&\approx \tilde{\mathbf{x}} - \sum_{m=1}^{M} \boldsymbol{\Delta}(\boldsymbol{\mu}_m) \cdot \mathrm{P}(m \mid \tilde{\mathbf{x}})
\end{aligned} \tag{2.13}$$

The last line is very similar to the estimation formula of 2DCMS (equation 2.6), except now, there are M instead of 2 classes and the classes are not given explicitly. The 2 explicit classes speech and pause can be identified by means of a speech activity detector or by two simple acoustic models which would give a probability estimation $\mathrm{P}(class \mid \tilde{\mathbf{x}})$. However, if the M prototypes are identified with the components of a mixture density with M Gaussians

$$\tilde{\lambda} = \left\{\tilde{\boldsymbol{\mu}}_1,\ldots,\tilde{\boldsymbol{\mu}}_M, \tilde{\mathbf{C}}_1,\ldots,\tilde{\mathbf{C}}_M, \hat{\mathrm{P}}(1),\ldots,\hat{\mathrm{P}}(M)\right\} \tag{2.14}$$

we can calculate an estimate for clean speech with the normal distribution $\mathrm{N}(\tilde{\mathbf{x}}[k];\tilde{\boldsymbol{\mu}}_m,\tilde{\mathbf{C}}_m)$ as:

$$\begin{aligned}
\hat{\mathbf{x}}[k] &= AM\left(\tilde{\mathbf{x}}[k];\lambda,\tilde{\lambda}\right) \\
&= \tilde{\mathbf{x}}[k] + \frac{\sum_{m=1}^{\mathrm{M}} \hat{\mathrm{P}}(m) \cdot \mathrm{N}(\tilde{\mathbf{x}}[k];\tilde{\boldsymbol{\mu}}_m,\tilde{\mathbf{C}}_m) \cdot (\boldsymbol{\mu}_m - \tilde{\boldsymbol{\mu}}_m)}{\sum_{m=1}^{\mathrm{M}} \hat{\mathrm{P}}(m) \cdot \mathrm{N}(\tilde{\mathbf{x}}[k];\tilde{\boldsymbol{\mu}}_m,\tilde{\mathbf{C}}_m)}
\end{aligned} \tag{2.15}$$

3. APPLYING MODEL-COMBINATION-BASED ACOUSTIC MAPPING

Before we could use the proposed method we had to train a generic speech model using clean training data. We used 30 mean subtracted MFSC as feature vectors to train this secondary model. We experimented with different numbers of Gaussians *M*. Seeing a clear performance gain when increasing this number from 10 to 100, there was only a very small difference between 100 and 1000 Gaussians. The results presented in this paper where obtained with 100 Gaussians. As this secondary model has much fewer parameters compared to the recognizer model we reduced the amount of training data to one tenth of the available clean speech data.

For the decoding of an utterance we only used the data from this individual utterance. A speech pause detector identified the noise frames. The detected feature vectors in the secondary model domain were collected and used to generate a single Gaussian noise model. This noise model was combined with the secondary model for clean speech to derive a corresponding model for noisy speech.

Based on the two corresponding models we could then use Acoustic Mapping *AM(..)* to estimate a clean feature vector in this domain. Therefore, each feature vector $\tilde{\mathbf{x}}[k]$ was evaluated by the noisy speech model. It was assigned with $\mathrm{P}(m \mid \tilde{\mathbf{x}}[k])$ to the M Gaussians (see equation 2.15). The transformation performed by *AM(..)* is a shift by the sum of all mean vector differences $(\mu_m - \tilde{\mu}_m)$ weighted by $\mathrm{P}(m \mid \tilde{\mathbf{x}}[k])$.

Since the secondary model had used mean subtracted (channel compensated) features, we also made sure that the noise frames used to build the noise model are also channel compensated. This has to be a similar channel estimate like during the training of the secondary clean speech model. In order to decrease the dependency on the current noise level we used a scheme similar to speech based mean subtraction, that is, we took only frames with high energy to estimate the mean. Using a channel compensated noise and speech model is an advantage since no weighting factor is required for the combination. The combined model for noisy speech will be channel normalized as well and should therefore be used to compensate a channel normalized input feature.

After noise and channel compensation is done using MAM, the estimated clean feature vector can be further processed. Here we used our standard scheme: transformation into the cepstral domain, liftering, adding first and second order derivatives and LDA transformation.

4. EXPERIMENTS

4.1 Recognition Task

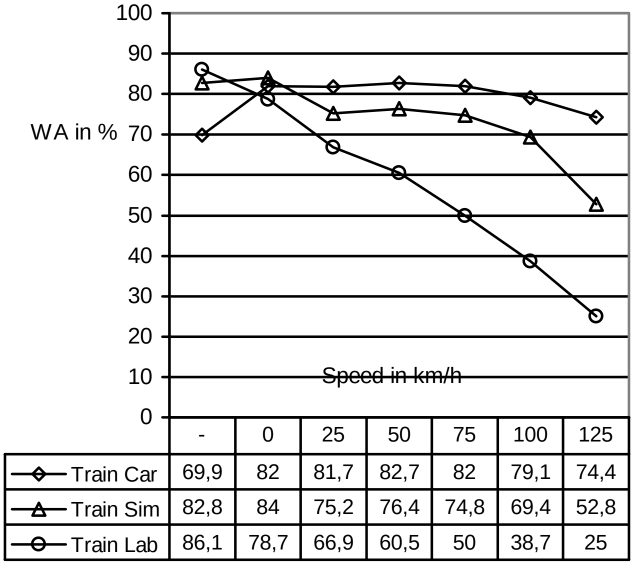
To evaluate the proposed algorithm MAM, we used a test set described in [12]. Each of the 12 subsets consists of 300 utterances from 10 speakers and covers the same navigation queries uttered in the car but was recorded under a different environmental condition. In this paper, we report results on the quiet category (engine and fan off) and six categories covering different speed values (between 0 and 125 km/h).

We used a class based language model for spontaneous speech. The size of the vocabulary was about 3000 words including 1800 street names for the city of Karlsruhe. Similar classes exists for neighborhoods, numbers, points of interest, and so forth.

4.2 Baseline Results

Our goal was to increase robustness by compensation or adaptation methods that will allow us to use training data that was collected in a quiet office or lab environment since this kind of data has been collected in larger amount in the past. Of course, it is also very common to collect data for new environments of interest (for example noisy car) because this practice will most often improve performance for this conditions. However, this procedure is very expensive, covers only the new, very specific environment and task, and will most often degrade the performance for other environments (office data, to pursue the example above). As another reference for the investigated methods, we not only trained a system on clean training data but also performed a continuous speech data collection in the car (see [12] for details).

Figure 4.1 shows results for three reference systems. Using 30 hours of clean spontaneous speech training data (Train Lab) gives the best word accuracy for the clean condition (-) but degrades very drastically for the other, mismatched conditions. The system trained with about 10 hours real car speech (Train Car) performs much better overall but shows losses for the clean condition. The third system (Train Sim) is also specialized for the car environment as recorded car noise data was added to the clean lab data. It shows significant improvements for most distorted categories compared to 'Train Lab' but can not reach the accuracy of 'Train Car' here.



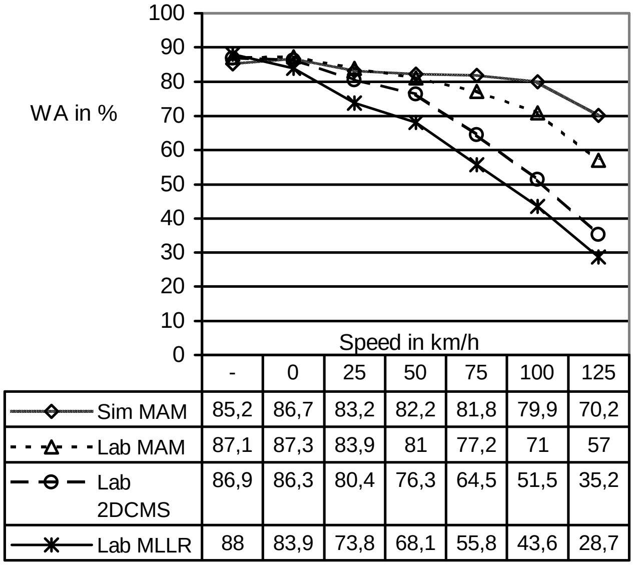
	-	0	25	50	75	100	125
◇ Train Car	69,9	82	81,7	82,7	82	79,1	74,4
△ Train Sim	82,8	84	75,2	76,4	74,8	69,4	52,8
○ Train Lab	86,1	78,7	66,9	60,5	50	38,7	25

Figure 4.1: Results for 3 recognition systems based on different training data.

4.3 Results Using Compensation Methods

Noise reduction methods such as Spectral Subtraction [1] or after Ephraim and Malah [3] gave us large improvements for a single word task but not when using them for spontaneous speech input. Therefore we tried different approaches like unsupervised adaptation and the methods discussed above. The test setup requires that they all work on an utterance base. As depicted in **Figure 4.2**, adaptation with MLLR showed some improvements but is also costly in terms of computation. 2DCMS, although very simple and computationally cheap, did somewhat better but was outperformed by the new method MAM. MLLR and MAM were both put on top of the 'Train Lab' system mentioned above. 2DCMS required a new training on the same clean data.

The performance of MAM could be further increased by using the noise mixed data as used for 'Train Sim'. Except for the last condition (125 km/h), it was able to reach better or similar results than the system trained on real car speech data.



	-	0	25	50	75	100	125
◇ Sim MAM	85,2	86,7	83,2	82,2	81,8	79,9	70,2
△ Lab MAM	87,1	87,3	83,9	81	77,2	71	57
○ Lab 2DCMS	86,9	86,3	80,4	76,3	64,5	51,5	35,2
✱ Lab MLLR	88	83,9	73,8	68,1	55,8	43,6	28,7

Figure 4.2: Results of the MAM compared with 2DCMS and unsupervised model adaptation (MLLR).

5. DISCUSSION

The proposed method MAM combines compensation with model adaptation. However, it is more efficient than adapting the model of the recognizer and very effective in compensating for the noise mismatch between signal and acoustic model. The knowledge about the structure of clean speech given by the secondary model, together with the environment model of channel and additive noise, made it superior over other methods we investigated. It not only considers an estimated noise level but also the noise variance.

MAM is able to reconstruct spectral coefficients that might be heavily distorted by noise using other spectral parts that help to identify the "right" prototypes. Therefore it is also interesting to compare it with methods used in missing feature theory [2][9]. They are based on an explicit identification of missing data, whereas MAM does this in an implicit way. Components of the prototypes that are missing (because they are below a noise dependent level) will be similar in the secondary model for noisy speech (after the combination with the noise model). Thus, these components do not discriminate between different prototypes, but the undistorted components do.

6. SUMMARY

Applying MAM for the recognition of spontaneous navigation queries in the car on top of a recognizer trained on clean speech data (never seen any noisy data) resulted in a 53 % relative error reduction for 100 km/h. When this method was also used during the training with clean speech data mixed with noise recordings the relative error reduction increased to 67 %. This compares with a system especially trained on real car speech data for this condition but is also appropriate for other conditions like quiet office or with different noise types.

REFERENCES

[1] **Steven F. Boll**: "*Suppression of Acoustic Noise in Speech Using Spectral Subtraction*", IEEE Transactions on Signal Processing, Vol. ASSP-27, No. 2, pp. 113-120, Apr 1979

[2] **Martin Cooke, Andrew Morris, and Phil Green**: "*Missing Data Techniques for Robust Speech Recognition*", ICASSP, IEEE, pp. 863-866, Munich, 1997

[3] **Y. Ephraim and D. Malah**: "*Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator*", IEEE Transactions on Acoustic, Speech, and Signal Processing, Vol. 33, No. 2, pp. 443-445, 1985

[4] **Mark John Francis Gales**: "*MODEL-BASED TECHNIQUES FOR NOISE ROBUST SPEECH RECOGNITION*", Dissertation, Gonville and Caius College, University of Cambridge, Sep 1995

[5] **M.J.F. Gales and S. J. Young**: "*HMM RECOGNITION IN NOISE USING PARALLEL MODEL COMBINATION*", EUROSPEECH, pp. 837-840, Berlin, Sep 1993

[6] **Jean-Luc Gauvain and Chin-Hui Lee**: "*Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains*", TransSAP, Vol. 2, No. 2, IEEE, pp. 291-298, Apr 1994

[7] **P. Geutner, M. Denecke, U. Meier, M. Westphal and A. Waibel**: "*Conversational Speech Systems For On-Board Car Navigation And Assistance*", ICSLP '98, Adelaide, Australia, 1998

[8] **C.J. Legetter and P.C. Woodland:** "*Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models*", Academic Press, Computer Speech and Language, Vol. 9, pp. 171-185, 1995

[9] **Richard P. Lippmann and Beth A. Carlson**: "*Using Missing Feature Theory to actively select Features for Robust Speech Recognition with Interruptions, Filtering, and Noise*", EUROSPEECH, KN-37, Rhodes, Sep 1997

[10] **Pedro J. Moreno, Bhiksha Raj, Evandro Gouvea and Richard M. Stern**: "*Multivariate-Gaussian-Based Cepstral Normalization for Robust Speech Recognition*", ICASSP, IEEE, pp. 137-140, Detroit, 1995

[11] **Martin Westphal**: "*THE USE OF CEPSTRAL MEANS IN CONVERSATIONAL SPEECH RECOGNITION*", EURO-SPEECH, Rhodes, Sep 1997

[12] **Martin Westphal and Alex Waibel:** "*TOWARDS SPONTANEOUS SPEECH RECOGNITION FOR ON-BOARD CAR NAVIGATION AND INFORMATION SYSTEMS*", EUROSPEECH, ESCA, Budapest, Sep 1999