Audio Engineering Society

Convention Paper

Presented at the 120th Convention
2006 May 20–23 Paris, France

Multichannel-Noise Reduction-Systems for Speaker Identification in an Automotive Environment

Volker Mildner, Stefan Goetze and Karl-Dirk Kammeyer

University of Bremen, Dept. of Communications Engineering, P.O. Box 330 440, D-28334 Bremen, Germany

Correspondence should be addressed to Volker Mildner (mildner@ant.uni-bremen.de)

ABSTRACT

Devices for communication and information utilised by car drivers are facing two essential requirements: hands-free operation via distant microphones but also robustness against different noises depending on car speed etc. Automatic speaker identification can be utilized within such devices to either supply speech recognition systems with so called apriori information to achieve higher recognition rates or even to enable applications such as heating systems to adjust to the preferences of the driver. Thus identifying the driver from a predefined group of possible system users may be a task for future applications. The aim in this work is to investigate to which extent multi-channel noise reduction systems are suitable for improving the performance of speaker identification algorithms under different acoustic conditions in an automotive environment.

1. INTRODUCTION

Different devices such as mobile phones or navigation systems may be installed permanently or temporarily in the cockpit of a car. Instead of operating them manually it is the task to allow total hands-free control in order to minimize the distraction of the driver.

Applications of these devices include the aim of speech recognition e.g. in order to make a certain request to the navigation system. The information retrieved from automatic speaker identification can be exploited to support speaker dependent solutions for speech recognition. Moreover, other systems such as heating, air conditioning or radio may be adapted to the preferences of the driver in the case that is known who is conducting the vehicle. Thus, automatic speaker recognition in the sense of identifying an operator from a closed set of users may be considered a problem of the ambition to realize the task of voice controled systems.

Text-independent speaker identification via Gaussian Mixture Models (GMM) has been introduced by Reynolds et al. [1, 2]. During training, a statistical model is computed for each speaker based on features extracted from a speech sequence. During testing the speaker shall be identified by a shorter test-sequence that was not included in the training sequence. The performance can be investigated based on clean speech signals provided by a database [3].

In a car cabin two kinds of acoustic degradation occur during the potential test scenario: Firstly, reverberation caused by reflections inside the cabin and furthermore noises emerging from tire-friction, airstream or the engine. This results in a reduced recognition rate. Therefore, we investigate in this work to which extent multi-channel noise reduction systems are capable of improving the recognition rate.

1.1. Outline

This paper is organized as follows: In Section 2 we explain the idea of text-independent speaker identification via gaussian mixture models including the aspect of feature extraction. Performance limits due to chosen parameters are shown in terms of recognition rates based on clean speech signals only. Section 3 specifies the set-up for the acoustic scenario during the test case in a car cabin and defines the different test-cases. The approaches of multi-channel systems for noise reduction considered in this work are explained in Section 4. Finally, the performance of the different systems for the test-cases is outlined in Section 5. Conclusions are drawn in Section 6.

2. SPEAKER IDENTIFICATION

Gaussian mixture models (GMM) are used to build stochastic models of the features extracted from a speech sequence of a specific speaker. This has been introduced by Reynolds [2]. Here, we review the approach and point out aspects of parameter choice.

2.1. Feature Extraction

The discrete-time signal $s(k)$, sampled at a sampling frequeny $f_s = 8\text{kHz}$, is segmented into frames of $R = 256$ samples with a frame index $\tau = 1..T$, where T is the number of all frames. Adjacent frames overlap by $R/2 = 128$. For each frame a feature-vector $\mathbf{f}_\tau = (f_{\tau,1}...f_{\tau,D})^{\mathrm{T}}$ of D dimensions is extracted. All feature-vectors of one speech sequence form the set $\mathbf{F} = (\mathbf{f}_1, .., \mathbf{f}_T)^{\mathrm{T}}$.

The extracted features are the Mel Frequency Cepstral Coefficients [2, 4], widely used for recognition tasks in speech processing. The exact implementation was done in HTK format [4] leading at $f_s = 8\text{kHz}$ to $D = 12$ coefficients per vector.

2.2. Gaussian Mixture Models

The probability density function of an observed feature-vector $\mathbf{f}_\tau$ given by a single gaussian mixture with index i is

$$b_i(\mathbf{f}_\tau) = \frac{1}{(2\pi)^{D/2} \cdot |\mathbf{C}_{ff}|^{1/2}} \cdot \exp\left[-\frac{1}{2}\left(\mathbf{f}_\tau - \boldsymbol{\mu}_i\right)^{\mathrm{T}} \mathbf{C}_{ff,i}^{-1} \left(\mathbf{f}_\tau - \boldsymbol{\mu}_i\right)\right] \quad (1)$$

where $\boldsymbol{\mu}_i$ is the D-dimensional mean-vector and $\mathbf{C}_{ff,i}$ the $D \times D$ covariance matrix. Here, we restrict the model to have only diagonal covariance matrices in order to reduce computational complexity. For each mixture $i = 1..M$ a weighting factor p_i is given satisfying the condition $\sum^M p_i = 1$. The set of parameters for all M mixtures are summarized as $\lambda = \{p_i, \boldsymbol{\mu}_i, \mathbf{C}_{ff,i}\}$. Finally, the probability of an observation $\mathbf{f}_\tau$ given by the model λ is

$$p\left(\mathbf{f}_\tau|\lambda\right) = \sum_{i=1}^{M} p_i. \cdot b_i(\mathbf{f}_\tau) \quad (2)$$

Assuming independence of the observations the overall probability of a set of observed feature vectors from one speech sequence fitting a model q can be computed as

$$p\left(\mathbf{F}|\lambda_q\right) = \prod_{\tau=1}^{T} p\left(\mathbf{f}_\tau|\lambda_q\right) \quad (3)$$

or in the logarithmic scale as

$$\log p\left(\mathbf{F}|\lambda_q\right) = \sum_{\tau=1}^{T} \log p\left(\mathbf{f}_\tau.|\lambda_q\right) \quad (4)$$

2.2.1. Model Estimation from Training Data

Having the set of observations $\mathbf{F}_{Train} = (\mathbf{f}_1, .., \mathbf{f}_T)^{\mathrm{T}}$ from a training sequence of speaker q at hand the parameter set λ_q has to be determined such that the log-likelihood $\log\left[p\left(\mathbf{F}_{Train}|\lambda_q\right)\right]$ is maximized. For this purpose the expectation-maximization-algorithm (EM) is applied [2, 5]. Starting from an initial set of parameters λ the algorithm performs an iteration to find a new set of parameters $\bar{\lambda}$ for which

$$\log p\left(\mathbf{F}|\bar{\lambda}\right) \geq \log p\left(\mathbf{F}|\lambda\right) \tag{5}$$

is known to be guaranteed.

Repeating the iteration of the EM-algorithm will lead to an improved or equivalent model. The algorithm is known to converge to a local maximum of $\log p\left(\mathbf{F}|\bar{\lambda}\right)$. Since the number of iterations W, needed for convergence of the model, strongly depends on the training data, we refrain from fixing W to a certain value but use a different criterion instead. Let

$$\Theta = \left|\log p\left(\mathbf{F}|\bar{\lambda}^{w+1}\right)/\log p\left(\mathbf{F}|\bar{\lambda}^{w}\right) \; - 1\right| \tag{6}$$

indicate the relative increase of the logarithmic probability gained from iteration step w to $w+1$, then we terminate the algorithm if $\Theta < 1e-6$ [6]. The initialization of the model parameters is known to be an uncrucial aspect [1]. For the M mean-vectors $\boldsymbol{\mu}_i$ we draw M different observations $\mathbf{f}_\tau$ from the training data at random. The diagonal entries of all coariance matrices $\mathbf{C}_{ff,i}$ are initialized by the variance values $\sigma^2_{1..D}$ of the training data. The mixture weights are set to $p_i = 1/M$. The remaining question of choosing the order of the model M shall be addressed in Section 2.3.

2.2.2. Classification of Test-Data

From a test-sequence the set of feature vectors $\mathbf{F}_{Test} = (\mathbf{f}_1, .., \mathbf{f}_T)^{\mathrm{T}}$ is extracted as described in 2.1. To perform a closed-set classification of the test-data that model of speaker $\hat{q}$ maximizing the probability of the observations $\mathbf{F}_{Test}$ fitting the model $\lambda_{\hat{q}}$ is chosen [1]

$$\hat{q} = \arg\max_{1\leq q\leq Q} \sum_{\tau=1}^{T} \log p\left(\mathbf{f}_\tau|\bar{\lambda}_q\right) \tag{7}$$

identifying speaker number $\hat{q}$ as the one to have spoken the test sequence.

2.3. Performance on Clean Speech Data

The performance of the MFCC-features in conjunction with GMMs is first evaluated on clean speech data from the KING database [3]. The database provides speech samples from 51 male speakers out of which we choose those 26 that were recorded during 10 sessions in San Diego. The speech samples from the speakers recorded in New Jersey are known to suffer from an insufficient SNR level and are therefore not considered.

The material is divided into 90 seconds of training sequences for each speaker and the rest for test sequences of 10 seconds length. In total there are 410 test sequences giving an approximate average of 16 test sequences per speaker.

For feature extraction from the speech sequences only frames of speech acitivity are considered and pause frames are neglected. The necessary voice activity detection (VAD) [7] has been performed in a so called batch-mode: the decision of the VAD was performed based on the knowledge of the whole utterance.

During the simulations the model order of the GMMs was varied from $M = 5$ up to $M = 50$. The result can be seen in Figure 1 as the solid black line on which we want to focus at first.

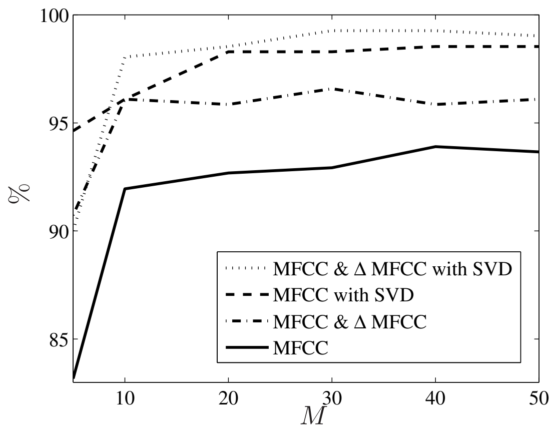

Figure 1: Recognition rates for different feature types and varied model order M

Starting from $M = 5$ the results clearly show with increasing model order higher recognition rates and therefore also indicate more accurate models. But the results also show that it is not possible to attain perfect recognition results by just increasing the model order beyond a certain value. At first sight it seems to be sufficient to assume $M = 40$ mixtures for the GMMs.

2.4. Variants of MFCC Features

In order to yield higher recognition rates further variants of MFCC features that can be found in [1] are considered.

The first one is the attempt of including temporal information in the features by the so called Δ-coefficients. For the frame τ the vector of Δ-coefficients is

$$\Delta \mathbf{f}_\tau = \mathbf{f}_\tau - \mathbf{f}_{\tau-1}. \quad (8)$$

This vector is concatenated with the initial feature-vector $\mathbf{f}_\tau$ to form the new feature-vector

$$\boldsymbol{\Delta}\mathbf{f}_\tau = [\ \mathbf{f}_\tau \ \ \Delta\mathbf{f}_\tau\ ]\,. \quad (9)$$

It shall be noted that the new feature vectors now have $2D$ dimensions, demanding higher computational complexity. The results for such feature-vectors are plotted as the dash-dotted line in Figure 1. Apparently appending the temporal information in form of the Δ-coefficients leads to improved statistical models indicated by the higher recognition rates.

The next variant is that of removing the bias from the feature-vectors caused by the acoustic channel between the speaker and the microphone. As it has been shown by Reynolds and Rose [1], the convolutive component in the time domain caused by an acoustic channel results in an additive component in the cepstral domain. It is assumed that this additive component is constant for one recorded speech sample. The method now is to remove the bias $\mathbf{b}$ of a set of feature-vectors

$$\mathbf{b} = \frac{1}{T}\sum_{\tau=1}^{T} \mathbf{f}_\tau \quad (10)$$

from all feature-vectors

$$\hat{\mathbf{f}}_\tau = \mathbf{f}_\tau - \mathbf{b}. \quad (11)$$

As an additional step of processing we perform a singular value decomposition (SVD) of the matrix formed by the feature-vectors $\hat{\mathbf{F}} \,\epsilon\, \mathbb{R}^{T\times D}$ followed by a projection onto the principal axes. The singular value decomposition is given by

$$\hat{\mathbf{F}} = \mathbf{U}\ \mathbf{S}\ \mathbf{V}^{\mathrm{T}} \quad (12)$$

where $\mathbf{V} \,\epsilon\, \mathbb{R}^{D\times D}$ contains the principal axes of the features onto which they are projected by

$$\hat{\mathbf{F}}_{proj} = \hat{\mathbf{F}}\ \cdot\ \mathbf{V}\ . \quad (13)$$

The projection onto the principal axes reduces the cross-correlation between different feature-dimensions and therefore makes diagonal covariance-matrices more suitable for modeling the true covariance matrices of the true probability density function.

The principal axes $\mathbf{V}_q$ obtained from the training sequence of a speaker have to be kept along with the model parameters λ_q for the step of classification. There, the matrix of the test-features has to be projected onto the principal axes of the speaker model it is tested against:

$$\hat{\mathbf{F}}^{test}_{proj} = \hat{\mathbf{F}}^{test}\ \cdot\ \mathbf{V}_q \quad (14)$$

The recognition rates (plotted as the dashed line in Figure 1) obtained by such features are even better than for just appending $\Delta-$coefficients .

Appending $\Delta-$coefficients after the bias removal and including them before the singular value decomposition yields features which here give a further increase of the recognition rate plotted as the dotted line in Figure 1.

It shall be a main aspect in Section 5 to evaluate the robustness of the presented feature variants in the case of noisy test sequences.

3. ACOUSTIC SCENARIO

The four test cases compared in this work are:

1. No reverberation - no noise
   (ideal acoustic case)
2. Reverberation by car-cabin - no noise
   (speaker in the car, car halted, engine off)
3. Reverberation by car-cabin - idle engine
   (speaker in the car, car halted, engine on)
4. Reverberation by car-cabin - noise at 50 km/h
   (speaker in the car, car driving)

For the multi-channel system a linear array of four microphones placed 6 cm apart from each other is considered.

3.1. Reverberation

The acoustic effect of reverberation is emulated by computing impulse responses $h_r(k)$, $r = 1..4$ between the speaker and the four microphones in a small room after [8]. The setup is illustrated in Figure 2.

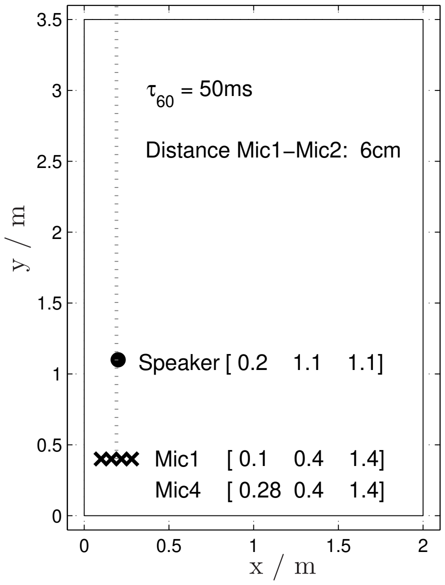

Figure 2: Room for simulating reverberation; the coordinates [x y z] indicate the position of the speaker and the microphones 1 and 4 (in meters)

The specifications are:

- The room is of 2 meter width (x) , 3 meter depth (y) and 1.5 meters height (z).
- The microphone array is placed at the same width as the speaker, but 30 cm above. The distance of the speaker to the array is 70 cm.
- The microphone array is pre-steered directly towards the speaker.
- The reverberation time was chosen as τ_{60} =50ms which is sufficient as an assumption for car-cabins.

3.2. Noise

For multichannel noise-reduction systems certain assumptions are made about the noise-signals perceived at the different microphone-channels. The assumptions and their consequences will be discussed in detail in Section 4. To account for a realistic environment the noise signals sensed at the microphones were not generated by artificial noise but recorded. An array of four microphones with the analogue spacing of d = 6cm was mounted at the sun visor of a medium-sized vehicle. Two kind of noises were recorded:

- Halted car, engine running idle
- Car driving at 50 km/h on asphalt

The noise generated by air stream, the engine and tire friction at 50 km/h is shown in Figure 3. It has obviously strong low-pass characteristics which emerge from vibrations of the car body and the inner lining. This shall be addressed again in Section 4.

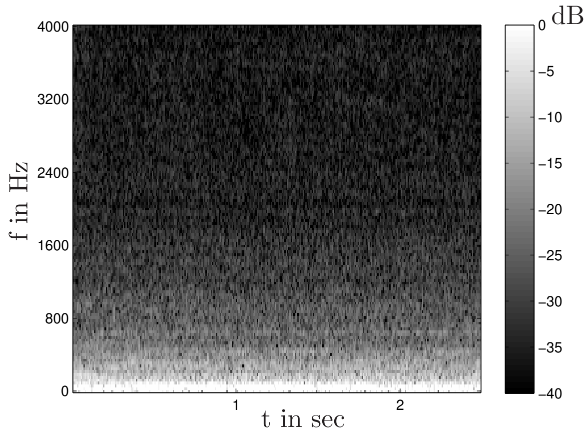

Figure 3: Car noise at 50 km/h

3.3. Generating Test-Sequences

For the task of generating proper test sequences under the different acoustic conditions the following steps were performed:

- For consideration of reverberation the clean speech signal $\tilde{s}(k)$ was convolved with the impulse responses $h_r(k)$ to yield the signal at the specific microphone r leading to $s_r(k) = \tilde{s}(k) * h_r(k)$.

- If noise had to be considered the noise signal $n_r(k)$ was added to the reverberated speech signal giving $x_r(k) = s_r(k) + n_r(k)$.

The noise level for the test cases 3 and 4 were mixed at signal-to-noise ratios equivalent to recorded samples taken during the recording session. Exemplarily we depict four variants of speech samples in Figure 4.

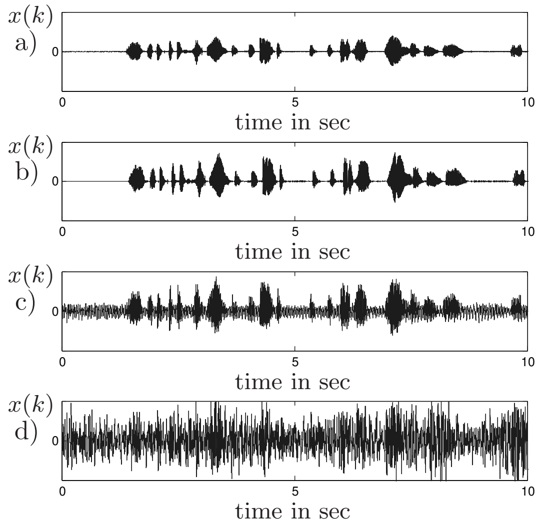

Figure 4: Examples of speech samples:
a) clean speech b) reverberation only
c) noise at idle engine d) noise at 50km/h

4. MULTI-CHANNEL NOISE-REDUCTION

4.1. Preprocessing

The noise perceived at a microphone while driving at 50 km/h was presented in Figure 3. The strong lowpass characteristics of the noise with the noise power being strongest below 200Hz motivate highpass filtering of the microphone signals prior to subsequent processing. Applying a highpass filter with a cutoff frequency of 200Hz would on the one hand remove the greatest amount of noise power, but at the same time also parts of the speech signal. The main formants of male speakers may lie as low as 100 Hz. Therefore, we use a highpass filter with a cutoff-frequency of 50Hz, allowing frequencies around 100Hz to remain unattenuated. Harsher attenuation of low frequencies might remove substantial information of the speech signal which shall be extracted by the MFCC-features.

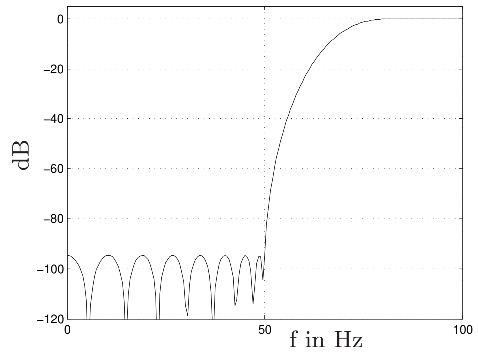

Figure 5: Highpass-Filter for Preprocessing

The fact that all test-sequences are consequently high-pass filtered demands an adaption of the GMMs due to the fact that low frequency-components are now totally missing for extracting the MFCC-features. Thus, we decided to also apply highpass-filtering to the training sequences prior to feature extraction and model estimtation. Testing the new models with features extracted from undisturbed and high-pass filtered test-sequences resulted in almost the same recognition rates as those presented in Section 2.3. Exact results will be presented in Section 5.

4.2. Considered Systems

Multi-channel noise reduction systems are known to be superior to single-channel solutions in terms of noise reduction and speech signal quality. The best-known single-channel solution for noise reduction is that of Ephraim and Malah [9].

The multi-channel systems considered in this contribution are [10, 11]:

- Delay&Sum-Beamformer
- Delay&Sum-Beamformer followed by a Post-Filter
- Superdirective-Beamformer
- Superdirective-Beamformer followed by a Post-Filter

The different systems are now explained more detailed.

4.3. Delay&Sum-Beamformer

The Delay&Sum-Beamformer is the optimal solution of a system with *Minimum Variance Distortionless Response* (MVDR) for the assumption of an uncorrelated noise field [10]. The design criterion is a minimization of the signal power under the constraint of distortionless transfer function for the desired source signal in look direction, which is the speech signal.

The general structure can be seen in Figure 6: For each microphone channel the signal is delay-comensated to pre-steer the beamformer towards the desired source. In our case no delay compensation is necessary since the beamformer is placed in line with the speaker. The signals $X_r(\Omega)$ from the $r = 1..4$ microphones are summed and normalized by the number of microphones R. This yields the signal $Y(\Omega)$, while Ω denotes the frequency index in the Fourier domain. As an extension a post filter

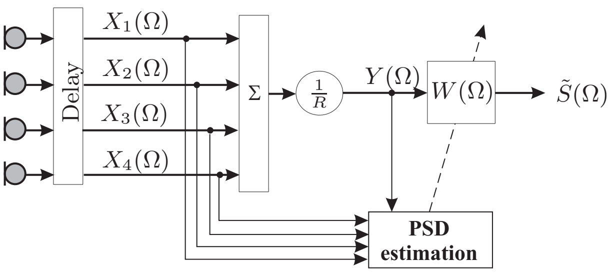

Figure 6: Delay&Sum-Beamformer with Postfilter

$W(\Omega)$ can be applied, leading to the output signal $\tilde{S}(\Omega) = Y(\Omega) \cdot W(\Omega)$. This shall yield a higher noise reduction. Under the assumption that the speech signal $S(\Omega)$ and the noise signal $N(\Omega)$ are uncorrelated for each channel r and that the noise signals $N_r(\Omega)$ between different channels are also uncorrelated, Zelinski [12] designed a postfilter

$$W_Z(\Omega) = \frac{\frac{2}{R(R-2)}\Re\{\sum_{i=1}^{R-1}\sum_{j=i+1}^{R} X_i(\Omega)^* X_j(\Omega)\}}{\frac{1}{R}\sum_{i=1}^{R} X_i^*(\Omega) X_i(\Omega)} \tag{15}$$

where $\Re$ denotes the real part of a complex variable, and $()^*$ the conjugate complex. It can be looked at as an implementation of a Wiener filter [13, 14]

$$W_W(\Omega) = \frac{\Phi_{SS}(\Omega)}{\Phi_{SS}(\Omega) + \Phi_{NN}(\Omega)}. \tag{16}$$

for which $\Phi_{SS}(\Omega)$ is the power spectral density of the speech signal and $\Phi_{NN}(\Omega)$ that of the noise signal for an uncorrelated noise field.

Please note that for deriving (15) the influence of the beamformer causing noise attenuation has been neglected.

As it has been investigated previously [15, 16, 13, 17] the assumption for uncorrelated noise signals between different microphone channels is usually not fulfilled for all frequencies. A measure for this is the *Magnitude Squared Coherence* (MSC)

$$MSC = \Gamma^2_{X_iX_j}(\Omega) = \frac{\left|\Phi_{X_iX_j}(\Omega)\right|^2}{\Phi_{X_iX_i}(\Omega)\Phi_{X_jX_j}(\Omega)}, \tag{17}$$

between the noise signals of different channels $\{i, j\}$. For the design of the Zelinski rule the MSC was assumed to be zero for all frequencies.

Plotted below in Figure 7 is the MSC between the recorded noise signals of channels $\{1, 2\}$ and $\{1, 4\}$ respectively.

The MSC clearly follows the function being characteristic for a diffuse noise-field

$$\Gamma^2_{X_iX_j}(\Omega) = \mathrm{si}^2(2\pi \cdot \Omega \cdot d_{ij}/c) \tag{18}$$

depending on the distance between the microphones d_{ij} only (c is the speed of sound). Obviously, there is no microphone pair $\{i, j\}$ at hand to provide reliable cross-correlation terms $X_i^*(\Omega)X_j(\Omega)$ for frequencies below 1000 Hz as we see from Figure 7.

To encounter this problem, the computation of the Zelinski-Postfilter was implemented for seperate

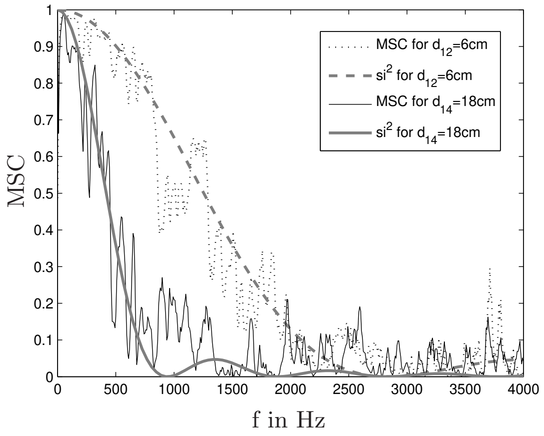

Figure 7: MSC for microphones $\{1, 2\}$ and $\{1, 4\}$

subbands [17] as

$$W_t(\Omega) = \frac{\frac{2}{t(t-2)} \Re \left\{ \sum_{i=1}^{t-1} \sum_{j=i+R-t+1}^{R} X_i^*(\Omega) X_j(\Omega) \right\}}{\frac{1}{R} \sum_{i=1}^{R} X_i^*(\Omega) X_i(\Omega)} \tag{19}$$

where t the index of the subband B_t with $t = 1..R$.

For the lowest frequencies below 1000 Hz it is refrained from computing a post-filter via (19) since no microphone pairs will provide reliable estimates. Instead, the filter coefficients $W(\Omega)$ are computed as for the single-channel solution of Ephraim-Malah [9].

4.3.1. Modified Post-Filter

Another variant of a post-filter deals with the problem, that for the derivation of (15) the noise attenuation introduced by the beamformer has been neglected. This might lead to a too strong attenuation in general by the Post-Filter. Thus, Simmer [14]defined a modfied post-filter, for which the denominator of (15) is replaced by the autocorrelation term behind the beamformer $Y_i^*(\Omega)Y_i(\Omega)$. Exploiting the subband approach, Simmer's weighting rule can be implented as [13]

$$W_{SW}(\Omega) = \frac{\frac{2}{R(R-2)} \Re \left\{ \sum_{i=1}^{R-1} \sum_{j=i+1}^{R} X_i^*(\Omega) X_j(\Omega) \right\}}{Y^*(\Omega) Y(\Omega)} \tag{20}$$

4.4. Superdirective-Beamformer

As outlined in the previous section, the noise field in a car cabin can not be considered uncorrelated for all frequencies but rather as a diffuse one, for which the MSC has si^2-characteristics.

The optimal solution of a MVDR-Beamformer for a diffuse noise field is the *Superdirective-Beamformer* depicted in Figure 8.

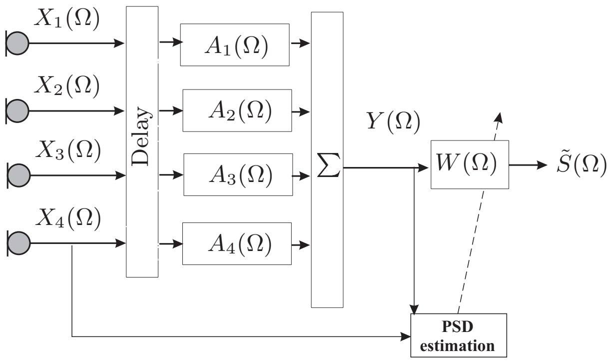

Figure 8: Superdirective-Beamformer with Postfilter

The filter coefficients $A_{1..4}(\Omega)$ are designed such that for a diffuse noise field signals, coming from the direction the beamformer is pre-steered to, are kept unattenuated [10].

The post-filter proposed in [18] to gain higher noise reduction is defined by the autocorrelation terms of the beamformer's output $Y(\Omega)$ and *one* channel signal $X(\Omega)$ as

$$W_{SD} = \frac{Y^*(\Omega)Y(\Omega)}{X^*(\Omega)X(\Omega)} \tag{21}$$

5. RESULTS AND DISCUSSION

The systems for multi-channel noise reduction presented in the previous section were applied to the test-sequences of the different test cases 1.to 4. before voice activity detection, feature extraction and classification of the signals. This Section shall discuss the question which algorithms of noise reduction are suitable in order to improve recognition rates for text-independent speaker identification.

5.1. Test Case 1

For the case of no reverberation and no noise, single-channel sequences are classified after high-pass filtering. The recognition rates show only marginal deviations from those of Section 2.3 without high-pass filtering. The results are depicted in Figure 9.

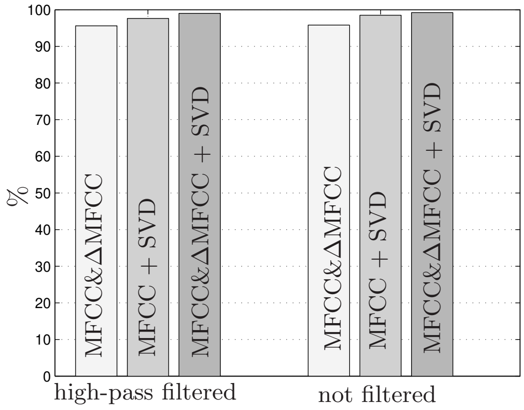

Figure 9: Recognition rates of clean sequences with and without high-pass filtering

Obviously the highpass-filtering had no significant impact on the features and the models.

As for non-filtered test- and training-sequences the features giving the best performance include the delta coefficients and a singular value decomposition after bias removal.

The model order chosen here was $M = 40$. For all following results shown below this will be the case. A variation of the model order was investigated but did not lead to superior results.

5.2. Test Case 2

For reverberation of the test sequences the speech signals were convolved with the impulse responses from section 3.3. We consider three types of signal processing for the test-sequences after high-pass filtering:

- Delay&Sum-Beamformer (D&S-BF)
- Superdirective-Beamformer (SD-BF)
- Single microphone, no further processing

Post-filters were not considered here since a high noise reduction is not demanded in this test case.

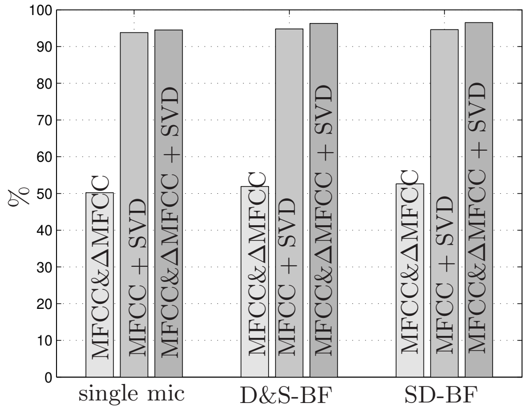

Figure 10: Recognition rates of test-sequences with reverberation

Apparently the application of bias-removal and projection of the features onto their principal axes obtained from the svd is the key to obtaining robust features for the case of reverberation. Channel-compensation performed by the bias-removal is essential. The gain in recognition rate achieved by multi-channel systems is minor compared to a single microphone. Nevertheless, the results show that speaker verification under the effect of reverberation in a car-cabin is feasible.

5.3. Test Case 3

In this test case noise emerging from an engine running idle was added to the reverberated speech sequences.

First, recognition rates obtained from sequences recorded by a single microphone shall be considered. The recognition rate for no signal processing is plotted in Figure 11 along with recognition rates achieved by the single-channel speech enhancement of Ephraim and Malah (E&M) [9]. Their weighting rule is known to suppress noise present in the recorded signal, but also to degrade and attenuate the desired speech signal to some extent. This depends on the signal-to-noise ratio and has been investigated in [Goetze] by instrumental measures. The aspect addressed now is to which extent noise reduction and/or signal degradation play a substantial role for speaker identification.

In case of no signal processing the recognition rate

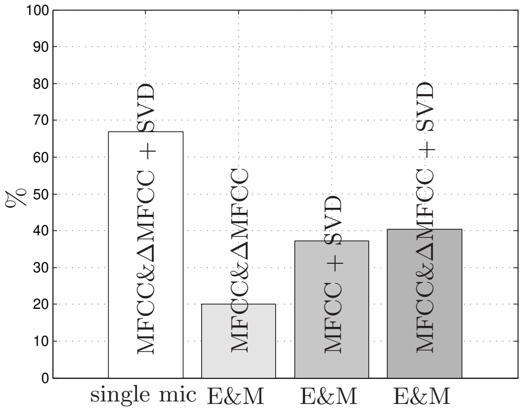

Figure 11: Recognition rates of test-case 3 for single-channel speech enhancement

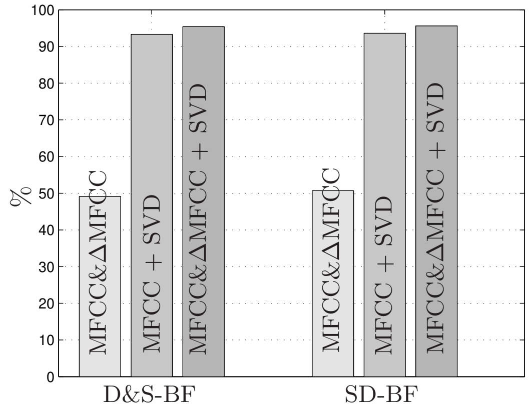

Figure 12: Recognition rates of test-case 3 for different beamformers

was highest for the same features as for the previous test cases, as denoted in the Figure, yielding 67% recognition rate. Improvement was not achieved by single-channel speech enhancement regardless of feature choice. This indicates that noise reduction does not seem to be the main goal of signal processing in our case. Instead, we suspect that signal degradation leads to distorted features and lower recognition rates.

As a next step, we examine the results when applying a Delay&Sum-Beamformer and a Superdirective-Beamformer.

The difference in recognition rate for the two beamformers are absolutely marginal. Just as for test-case 2 discussed in the previous Section, the features not having undergone bias removal and projection onto the principal axes are not robust enough for this test-case. They are therefore disregarded from here on.

We now want to focus on the techniques of post-filtering outlined in Section 4. We consider the following possibilities:

- Delay&Sum-Beamformer + Zelinski-Postfilter (D&S+Zel)
- Delay&Sum-Beamformer + Simmer-Postfilter (D&S+Sim)
- Superdirective-Beamformer + Postfilter W_{SD} (SD+Post)

The corresponding results are drawn in Figure 13.

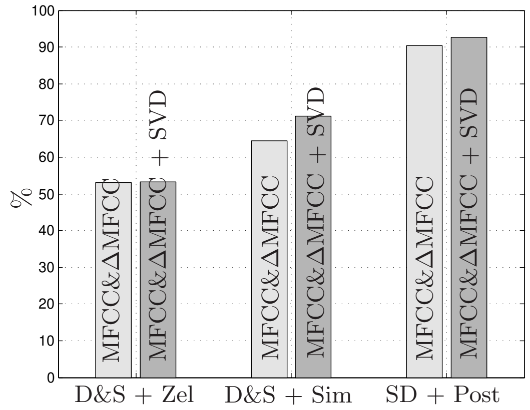

Figure 13: Recognition rates of test-case 3 for different postfilters

In general, no post-filter technique caused a better recognition rate than those when beamformers only. Apparently, a reduction of noise level achieved by these post-filters (as proven in [13]) is also for this test-case not the main goal. Much more, signal degradation is the major aspect. In [13] the signal-degradation caused by a Zelinski-post-filter was found to be more harsh than for the post-filter defined by Simmer, while their achievement of noise reduction is at equal level. This corresponds to the recognition results.

The postfilter W_{SD} of the Superdirective-Beamformer yields better results than other postfilters but not better than the results for using the beamformer only.

Also for this test-case speaker identification seems to be feasible.

5.4. Test Case 4

The last case considered was car noise at 50 km/h. The signal-to-noise ratio in this case is quite extreme as can be seen from the speech sample depicted previously in Figure 4. For different signal processing techniques the results are presented in Figure14

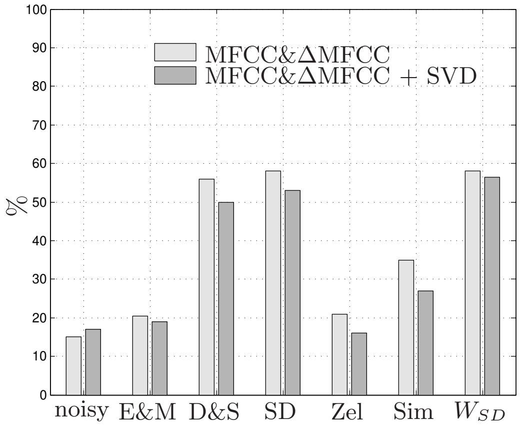

Figure 14: Recognition rates of test-case 4 for different filtering techniques

Let us at first not pay attention to the choice of feature but the general capabilities of the different filtering techniques. For classifying the noisy sequences ('noisy') directly the recognition rate drops to approximately 15%. The single-channel technique of Ephraim and Malah (E&M) gives only little improvement by a few percent. The Delay&Sum-Beamformer (D&S) and the Superdirective-Beamformer both yield recognition rates above 50% with the latter one being slightly superior. The Zelinski- (Zel) and the Simmer-postfilter (Sim) applied after the Delay&Sum-Beamformer once again cause detirioration of the recognition rates. The post-filter W_{SD} applied after the Superdirective-Beamformer neither reduces nor improves the recognition rate significantly.

While in test-case 3 the signal-to-noise ratio was at a rather moderate level, it was in test-case 4 at a rather extreme one. Under this condition even signal processing techniques still performing at a satisfactory level in case 3 fail to conserve the clean speech signal. As well as previously experienced, applying beamformers only to the signal seems to be the best approach in order to achieve good results since a minimum of signal degradration is the main issue for speaker recognition.

Focusing now on the feature choice, we see that the including the Δ-coefficients in the features did not give better results, but instead slightly worse. While under sufficient acoustic conditions in test-case 1-3 the temporal information extracted by the Δ-coefficients yielded more sensitive information and thus slightly higher recognition rates, this is not possible any more under adverse conditions.

6. CONCLUSION

In this paper, we combined gaussian mixture models for text-independent speaker identification with the acoustic circumstances of in-car communications. The general procedure of buiding stochastic models for speakers and comparing test sequences with the models was reviewed.

The acoustic conditions in a car-cabin were considered. Possible techniques of signal processing were outlined and discussed. Different test-cases were investigated in terms of recognition rates yielded by simulation results relying on a database.

The results showed that multi-channel techniques are superior to single-channel techniques for speaker identification under the considered circumstances. In general, beamformers applied to the signals alone are more favourable than any post-processing steps due to the fact that the signal degradation has to be kept at a minimum.

It has been shown that speaker identification is in some cases a feasible task in an automotive environment, but not under extreme conditions.

7. REFERENCES

[1] D. A. Reynolds, "An overview of automatic spekaer recognition technology," in *Proc. IEEE*

Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Orlando, Florida, May 2002.

[2] D. A. Reynolds and R. C. Rose, "Robust text-independend speaker identification using gaussion mixture speaker models," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[3] Graff D. Godfrey, J., "Public databases for speaker recognition and verification," *ECSA Workshop Automat. Speaker REcognition*, März 1994.

[4] S. Young, G. Evermann, D. Kershaw, J. Moore, G. Odell, D. Ollason, and P. Valtchev, V. Woodland, *The HTK Book*, Cambridge University Engineering Department, Cambridge, 2002, For more information visit http://htk.eng.cam.ac.uk/.

[5] Laird N. Dempster, A. and D. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *J. Royal Stat. Soc.*, vol. 39, pp. 1–38, 1977.

[6] N. Vlassis and A. Likas, "A greedy em alogrithm for gaussian mixture learning," in *Neural Processing Letters*, Netherlands, 2002.

[7] Q. Li, J. Zheng, A. Tsai, and Q. Zhou, "Robust endpoint detection and energy normalization for real-time speech and speaker recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 3, pp. 146–157, März 2002.

[8] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small–Room Acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, 1979.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean–square error log-spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[10] J. Bitzer and K. U. Simmer, "Superdirective microphone arrays," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. Ward, Eds., chapter 2, pp. 19–38. Springer-Verlag, 2001.

[11] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. Ward, Eds., chapter 3, pp. 39–60. Springer-Verlag, 2001.

[12] R. Zelinski, "A microphone array with adaptive post–filtering for noise reduction in reverberant rooms," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, New York City, New York, Apr. 1988, pp. 2578–2581.

[13] S. Goetze, V. Mildner, and K.-D. Kammeyer, "A Psychoacoustic Noise Reduction Approach for Stereo Hands-Free Systems," in *Audio Engineering Society (AES), 120th Convention*, Paris, France, 20.-23. May 2006.

[14] K. U. Simmer and A. Wasiljeff, "Adaptive microphone arrays for noise suppression in the frequency domain," in *Second Cost 229 Workshop on Adaptive Algorithms in Communications*, Bordeaux, Frankreich, Oct. 1992, pp. 185–194.

[15] J. Meyer (Bitzer) and K. U. Simmer, "Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, München, Deutschland, Apr. 1997, pp. 1167–1170.

[16] J. Li and M. Akagi, "A Hybrid Microphone Array Post-Filter in a Diffuse Noise Field," in *Proc. Eurospeech 2005*, Lisbon, Portugal, September 2005.

[17] V. Mildner, S. Goetze, and K.-D. Kammeyer, "Multi-Channel Speech Enhancement using a Psychoacoustic Approach for a Post-Filter," in *German ITG-Symposium on Speech Communication*, Kiel, Germany, 26.-28. April 2006.

[18] J. Bitzer, K. U. Simmer, and K. D. Kammeyer, "Multi-microphone noise reduction techniques for hands–free speech recognition –a comparative study–," in *Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finland, May 1999.