Performance of Text-Independent Speaker Identification considering In-Car Acoustics

Volker Mildner, Stefan Goetze, and Karl-Dirk Kammeyer
University of Bremen, Dept. of Communications Engineering, P.O. Box 330 440, D-28334 Bremen, Germany,
Email: mildner@ant.uni-bremen.de.de

Abstract

Hands free operation of communication devices (mobile phones, navigation systems etc.) in cars is becoming obliged to greater extents. A major task for such systems is that of speech recognition, for instance handling a navigation system via voice commands only. Algorithms for speaker identification may be used to provide speaker dependent speech recognition systems with the necessary a-priori information. Furthermore, the retrieved information of who is speaking (and operating the car) may be exploited to enable other systems (radio etc.) adapting to the preferences of the driver.

The performance of text-independent speaker identification under the acoustic circumstances occuring in a car is evaluated by the recognition rate of the system comparing different possibilities of multi-channel signal processing.

Introduction

Text-independent speaker identification using Gaussian Mixture Models was introduced by Reynolds et al. [1, 2]. In this work we want to evaluate the performance of speaker identification under the acoustic degradations occuring in a car environment: Reverberation and aditive noise. For different test-cases we try to improve the recognition rate of the system by applying multi-channel speech processing [3]. The results will give an indication to which extent speaker identification is possible under the given circumstances.

Speaker Identification

Feature Extraction

To extract representative features from one speech sequence, the discrete-time signal $s(k)$ is segmented into frames of R samples with frame index $\tau = 1..T$. The overlapping of the frames is R/2 samples. Only frames of speech activity were taken into account by decision from a voice activity detection [4]. For each frame the Mel-Frequency Coefficients (MFCC) are extracted giving a feature-vector $\mathbf{f}_\tau = (f_{\tau,1}...f_{\tau,D})^{\mathrm{T}}$ with D dimensions. For the need of channel-compensation [1] the overall bias of the feature-vectors $\mathbf{b} = \frac{1}{T}\sum_{\tau=1}^{T}\mathbf{f}_\tau$ is removed from all vectors yielding the compensated features $\hat{\mathbf{f}}_\tau = \mathbf{f}_\tau - \mathbf{b}$. In order to consider temporal information the Delta-coefficients $\Delta\hat{\mathbf{f}}_\tau = \hat{\mathbf{f}}_\tau - \hat{\mathbf{f}}_{\tau-1}$ are included in the feature vectors by

$$\mathbf{\Delta}\hat{\mathbf{f}}_\tau = \left[\ \hat{\mathbf{f}}_\tau \ \ \Delta\hat{\mathbf{f}}_\tau\ \right]. \tag{1}$$

The matrix formed by all feature-vectors

$$\mathbf{\Delta}\hat{\mathbf{F}}_{Train} = \left(\mathbf{\Delta}\hat{\mathbf{f}}_1, .., \mathbf{\Delta}\hat{\mathbf{f}}_T\right)^{\mathrm{T}} \epsilon\mathbb{R}^{T\times 2D} \tag{2}$$

is decomposed by the singular value decomposition (SVD) $\mathbf{\Delta}\hat{\mathbf{F}} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^{\mathrm{T}}$ giving $\mathbf{V}\,\epsilon\,\mathbb{R}^{D\times D}$ as the principal axes of the features which they are projected on

$$\mathbf{\Delta}\mathbf{F}^p = \mathbf{\Delta}\hat{\mathbf{F}}\ \cdot\ \mathbf{V}\,. \tag{3}$$

Gaussian Mixture Models

For each speaker with index $q = 1..Q$, a gaussian mixture model defined by its parameter set $\lambda_q = \{p_i, \boldsymbol{\mu}_i, \mathbf{C}_{ff,i}\}$ is computed via the EM-algorithm [1]. Here, $\boldsymbol{\mu}_i$ is the $2D$-dimensional mean-vector, $\mathbf{C}_{ff,i}$ the $2D \times 2D$ covariance matrix and p_i the weighting factors of the $M = 40$ mixtures ($i = 1..M$) satisfying the condition $\sum^{M} p_i = 1$.

For closed-set classification of a test-sequence, its features are compared to the Q different models. That model $\lambda_{\hat{q}}$ is chosen as the actual speaker model , for which the logarithm of the probability for the feature-vectors $\mathbf{\Delta}\mathbf{f}_\tau^p$ given the model takes its maximum [1].

$$\hat{q} = \arg\max_{1\le q\le Q}\sum_{\tau=1}^{T}\log p\left(\mathbf{\Delta}\mathbf{f}_\tau^p|\lambda_q\right). \tag{4}$$

Test Cases

In our investigation we considered a linear array of four microphones with a spacing of $d = 6$cm. The acoustic conditions considered were reveberation caused by the car-cabin and additive noise. The impulse responses for reverberating the clean test-sequences were generated by simulation [5], with a reverberation time $\tau_{60} = 50$ms. The noise signals were recorded using an array of equivalent spacing in a medium-sized vehicle. Altogether, four test cases were considered

1. Clean test signals
2. Reverberation only
3. Reverberation and noise from idle engine
4. Reverberation and noise at 50km/h

Signal Processing

All noisy test sequences were pre-processed by a highpass-filter with a cut-off frequency of 50kHz. This was done due to the strong lowpass-characteristic of the noise signals occuring in an automotive environment.

One alternative of multi-channel systems was the Delay&Sum-Beamformer (D&S). The beamformer was also extended by a different post-filters, including the weighting rule by Simmer W_{Sim} [6, 3] to achieve higher noise reduction. Another multi-channel system considered was the Superdirective-Beamformer (SD). Also for this beamformer a postfilter W_{SD} was used as an extension as described in [6, 3]. For reasons of comparison the single-channel weighting rule by Ephraim and Malah (E&M) [7] was applied to the noisy signals.

Results

The algorithms were tested on speech sequences from the KING database [8], using 90 seconds for training and 10 seconds for testing. There were 26 male speakers with an average of 15 test-sequences per speaker. Two kinds of tests were performed per test-sequence:

- Identfying a speaker out of all 26 speakers
- Identfying a speaker out of a group of 4 speakers

Using clean speech for testing we obtained a recognition rate of 99.2%. Under the influence of reverberation only we achieved 93% using only a single microphone (mic1), 96% for the Delay&Sum-Beamformer and 96.4% for a Superdirective-Beamformer, while no postfilters were applied. Depicted in Figure 1 are the recognition rates for the different of signal-processing under conditions of test-case 3 and 4. The results show that for moderate SNR conditions (case 3) a recognition rate of 90% or higher can be achieved by applying a beamformer while single channel solutions (mic1, E&M) perform worse. For test-case 4 we see that multi-channel solutions still achieve the best results, but these are at an insufficient level. The results when trying to identify a speaker from a group of 4 speakers are depicted in Figure 2. We draw the conclusion that speaker identification in a car from a large group of speakers under moderate noise-conditions is feasible, but not yet in a driving situation.

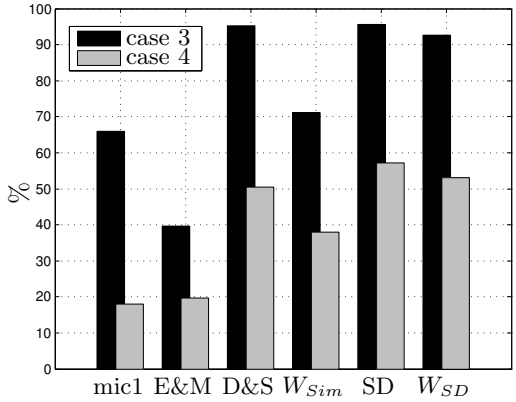

Abbildung 1: Recognition rates of test-cases 3 and 4 for identification out of a group of 26 speakers

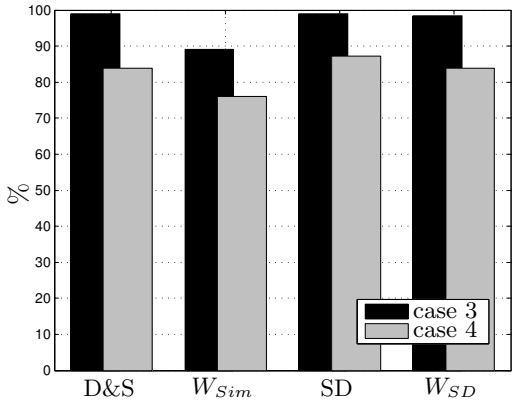

Abbildung 2: Recognition rates of test-cases 3 and 4 for identification out of a group of 4 speakers

Literatur

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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