Demonstrator for Automatic Text-independent Speaker Identification

Kristian Kroschel, Dirk Bechler

Institut für Nachrichtentechnik, Universität Karlsruhe, Kaiserstr. 12, 76128 Karlsruhe, Email: kroschel@int.uka.de

Introduction

For over three decades automatic speaker recognition is in the focus of research. An overview is found in [1, 2]. Besides forensic applications, mainly man-machine interaction is of interest. Typical applications are the identification of speakers by humanoid robots or the identification of the driver to adjust driver-specific parameters for the seat, mirrors etc. in a car.

Automatic text-independent speaker identification consists of two steps: first appropriate features are extracted from the speech signal, then these fearures are applied to a speaker identification system.

Feature Extraction

The *Mel Frequency Cepstral Coefficients* (MFCC) have proven to be the most appropriate parameters for speaker identification [3] which are also used as basic features for speech recognition. Fig. 1 shows a block diagram for the extraction of the MFCC feature vector $\mathbf{x}$ from the sampled speech signal $s(k)$. The sampled instationary speech signal $s(k)$ requires a short time spectral analysis based on segments of 16 ms each within which the signal is assumed to be stationary. These segments overlapping by the factor 0.5 and weighted with a Hamming window are transformed into the frequency domain by FFT of length $N = 256$. Using the Mel filter bank [4] which is similar to the spectral selectivity of the human ear, a reduced spectral representation is found by 40 filters with a triangular spectral shape. Below 1 kHz, 13 filters are spaced equally whereas the other 27 filters are spaced logarithmically along the frequency axis. The logarithm of the output of the 40 filters is applied to the Discrete Cosine Transform (DCT) which decorrelates the parameters. The 13 largest of these parameters form the MFCC vector $\mathbf{x}$ of the analyzed speech segment.

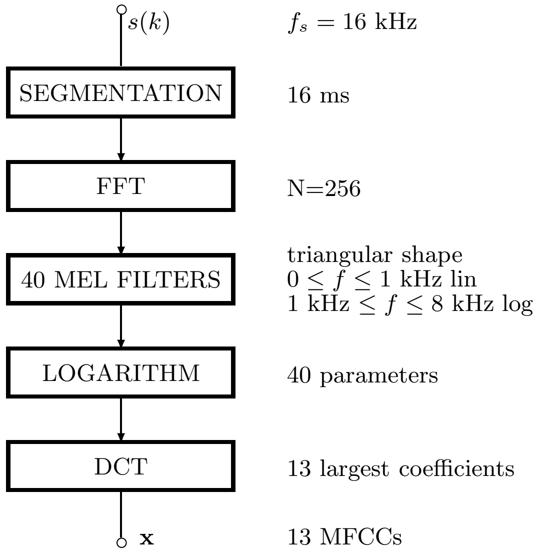

Figure 1: Block diagram for the generation of the MFCC parameters

Speaker Identification

The individual speakers are discriminated on the basis of their specific vocal tract configurations. Therefore speaker identification requires models which represent the characteristic vocal tract configuration of the individuals. A statistic source model can be applied with the speaker as the source of a random process. Within this random source, so-called hidden states can be identified. These represent the individual vocal tract configurations and generate the MFCC feature vectors.

Statistical Speaker Model

Speech generation is not deterministic and thus the spectra generated by the specific vocal tract configurations can vary significantly. This is why it is assumed that each state generates spectral feature vectors with a Gaussian density function $b_i(\mathbf{x})$ of dimension $D = 13$

$$\begin{aligned} b_i(\mathbf{x}) &= \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}_i|^{1/2}} \\ &\quad \cdot \exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T(\mathbf{\Sigma}_i)^{-1}\mathbf{x}-\boldsymbol{\mu}_i)\}, \end{aligned} \tag{1}$$

with the state-dependent mean vector $\boldsymbol{\mu}_i$ and the covariance matrix $\mathbf{\Sigma}_i$.

Gaussian Mixture Model (GMM)

The probability density function of the observed spectral feature vector which has been derived from the speaker model defined in Eqn. (1) is equivalent to the so-called *Gaussian Mixture Model* (GMM) [5]. For a statistical speaker model with M states, a GMM probability density function can be defined as

$$f(\mathbf{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\mathbf{x}), \tag{2}$$

with p_i the probability being in state i and

$$\lambda = (p_i, \boldsymbol{\mu}_i, \mathbf{\Sigma}_i)\,,\; i = 1, \ldots, M \tag{3}$$

representing the parameters of the speaker model. By $f(\mathbf{x}|\lambda)$ the probability is given that an unknown feature vector $\mathbf{x}$ is generated by a specific GMM speaker model.

Training the GMMs

To determine the model parameters of the GMM of a speaker, the GMM has to be trained. For this purpose the *Expectation-Maximization* (EM) algorithm [6] has proven to be most efficient. Iterative application of the EM algorithm is used to determine the parameters of the GMM speaker model on the basis of the MFCC feature vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ extracted from the *training set*. By this operation, the probability that the extracted GMM is equivalent to the training data is maximized. The *expectation* and the *maximization* step according to Eqns. (4)-(7) is repeated so that the parameter set $\lambda = (p_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i),\ i = 1, \ldots, M$ of the GMM converges:

expectation:

$$f(i|\mathbf{x}_n, \lambda) = \frac{p_i b_i(\mathbf{x}_n)}{\sum_{k=1}^{M} p_k b_k(\mathbf{x}_n)} \tag{4}$$

maximization:

$$p_i = \frac{1}{N} \sum_{n=1}^{N} f(i|\mathbf{x}_n, \lambda) \tag{5}$$

$$\boldsymbol{\mu}_i = \frac{\sum_{n=1}^{N} f(i|\mathbf{x}_n, \lambda)\mathbf{x}_n}{\sum_{n=1}^{N} f(i|\mathbf{x}_n, \lambda)} \tag{6}$$

$$\boldsymbol{\Sigma}_i = \frac{\sum_{n=1}^{N} f(i|\mathbf{x}_n, \lambda)(\mathbf{x}_n - \boldsymbol{\mu}_i)(\mathbf{x}_n - \mu_i)^T}{\sum_{n=1}^{N} f(i|\mathbf{x}_n, \lambda)}. \tag{7}$$

This training with the EM algorithm can be executed without supervision and does not require additional parameters. Furthermore, the algorithm to determine the parameters of the GMM converges after typically 40 iterations.

Application of the GMM

If S speaker models $\{\lambda_1, \ldots, \lambda_S\}$ are available after the training, speaker identification can be executed based on a new speech data set. First, the sequence of MFCC feature vectors $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ is calculated. Then the speaker model $\hat{s}$ is determined which maximizes the a posteriori probability $P(\lambda_s|\mathbf{X})$. The mixed form of the Bayes rule yields the following result:

$$\hat{s} = \max_{1 \le s \le S} P(\lambda_s|\mathbf{X}) = \max_{1 \le s \le S} \frac{f(\mathbf{X}|\lambda_s)}{f(\mathbf{X})} P(\lambda_s)\,. \tag{8}$$

Assuming equal probability of all speakers and the statistical independence of the observations, the decision rule for the most probable speaker can be redefined:

$$\hat{s} = \max_{1 \le s \le S} \sum_{t=1}^{T} \log f(\mathbf{x}_t|\lambda_s), \tag{9}$$

with T the number of feature vectors of the speech data set under test and $f(\mathbf{x}_t|\lambda_s)$ given by Eqn. (2).

Real-time Demonstrator

The system for speaker identification presented in this paper has been implemented as a real-time demonstrator. For the GMM of a speaker $M = 32$ states have been assumed. Currently the test set includes 12 speakers. To make the training process for a new speaker comfortable, the length of the training data is set to just 50 s. Short test sequences of about 1.5 s are sufficient for a reliable identification of the speaker. Together with a speech synthesis system which articulates the name of the identified speaker, the system is implemented in an artificial head with two microphones mounted in the ears. First tests showed a recognition rate of over 0.9. A detailed statistical test will be executed in the future.

Summary and Conclusion

The system presented in this paper is a first solution of a real-time speaker identifier which will be improved in the future. By increasing the number of states in the GMM, more training data and longer test sequences the recognition rate can be further improved. The selection of the system parameters can be adapted to the requirements of the application.

With additional spectral parameters from *Linear Predictive Coding*, *Frequency Cepstral Coding* etc. or other classifiers like *Neural Networks*, *Hidden Markov Models*, *Support Vector Machines* etc. the recognition rate might be improved further but the computational demand will go up, too.

The system might be trained for other acoustical sources: Experiments have shown that different bell and buzzer signals can be reliably identified.

In this investigation background noise has been ignored, i.e. the signal-to-noise ratio was high. Further investigations will tackle the problem of robustness and will include a rejection class. Up to now the system decides always for one of the trained speakers even if the distance to the trained data is very high.

Acknowledegement

This work has been supported by the German Science Foundation DFG within the Sonderforschungsbereich 588 *Hunanoide Roboter.*

References

[1] J.P. Campbell. *Speaker Recognition - a Tutorial.* Proc. of the IEEE, 85(9), September 1997, 1437-1462

[2] D.A. Reynolds, T.F. Quatieri, R.B. Dunn. *Speaker Verification Using Adapted Gaussian Mixture Models.* Digital Signal Processing, 10, 2000, 19-41

[3] D. O'Shaughnessy. *Speech Communications - Human and Machine.* IEEE Press, New York, 2000

[4] B. Gold, N. Morgan. *Speech and Audio Signal Processing.* Wiley, New York, 2000

[5] D.A. Reynolds, R.C. Rose. *Robust Text-independent Speaker Identification Using Gaussian Mixture Models.* IEEE Trans. on Speech and Signal Processing, 3(1), Januar 1995, 72-83

[6] A Dempster, N. Laird, D. Rubin. *Maximum Likelihood from Incomplete Data via the EM Algorithm.* Journal of the Royal Statistical Society, 39(1), 1977, 1-38