INTERSPEECH 2009 BRIGHTON

Speaker Identification using Warped MVDR Cepstral Features

Matthias Wölfel[1], *Qian Yang*[2], *Qin Jin*[3], *Tanja Schultz*[2,3]

[1]ZKM|Center for Art and Media, Germany
[2]Fakultät für Informatik, Universität Karlsruhe (TH), Germany
[3]Language Technologies Institute, Carnegie Mellon University, USA
woelfel@zkm.de, {qian, tanja}@ira.uka.de, qjin@cs.cmu.edu

Abstract

It is common practice to use similar or even the same feature extraction methods for automatic speech recognition and speaker identification. While the front-end for the former requires to preserve phoneme discrimination and to compensate for speaker differences to some extend, the front-end for the latter has to preserve the unique characteristics of individual speakers. It seems, therefore, contradictory to use the same feature extraction methods for both tasks. Starting out from the common practice we propose to use *warped minimum variance distortionless response* (MVDR) cepstral coefficients, which have already been demonstrated to perform superior for automatic speech recognition in particular under adverse conditions. Replacing the widely used mel-frequency cepstral coefficients by WMVDR cepstral coefficients improves the speaker identification accuracy by up to 24% relative. We found that the optimal choice of the model order within the WMVDR framework differs between speech recognition and speaker recognition, confirming our intuition that the two different tasks indeed require different feature extraction strategies.

1. Introduction

The performance of speaker identification and recognition systems has improved dramatically over the past years. The traditional approaches, such as Gaussian mixture models [1] [2], have achieved high accuracies on clean speech under well-matched conditions. Unfortunately, when facing real world conditions their performance degrades significantly [3] [4]. To reduce this performance drop we propose to increase the robustness of the speaker recognition system by extracting more robust features. The standard feature extraction techniques used for speaker recognition tasks are based on either Fourier transformation – *mel-frequency cepstral coefficients* (MFCC) [5] – or linear prediction – e.g. *perceptual linear predictive* [6]. However, they are either

- *not robust* and therefore perform poorly under adverse conditions which is the case for MFCC, or are
- *ill-suited* for the reliable estimation of the spectra of speech signals, especially for voiced speech, which is true for all methods using linear prediction envelopes. They tend to overestimate and overemphasize sparsely spaced harmonic peaks.

Therefore, we propose to replace the traditionally used front-ends by a front-end which is based on *minimum variance distortionless response* (MVDR) spectral estimation [7]. While MVDR spectral estimation overcomes the problems apparent in linear prediction spectral estimation, *warped minimum variance distortionless response* (WMVDR) [8] better suits the spectral characteristics of human hearing. For automatic speech recognition WMVDR has already been demonstrated to outperform other features, including their unwarped counterpart, particularly under adverse conditions. In the following speaker recognition experiments we investigate WMVDR spectral estimation method to extract relevant speaker information.

A detailed overview of the different spectral estimation methods (such as the Fourier transform, LP, MVDR, and warping) including their application in various front-ends is presented in [9]. A brief review and comparison between mel-frequency and WMVDR cepstral coefficients is given in the next section. Section 3 presents speaker recognition experiments on clean and noise speech data. Section 4 presents our conclusion and outlook.

This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

2. Comparison between Mel-Frequency and Warped MVDR Cepstral Coefficients

In this section we briefly review the properties of mel-frequency and warped MVDR cepstral coefficients, abbreviated as MFCC and WMVDRCC respectively. The flowchart of both front-ends is illustrated in Figure 1. It is easy to see that, to derive WMVDRCC, the Fourier transformation including the mel-scale filterbank is replaced by wapred MVDR spectral estimation [8], while all other steps are identical.

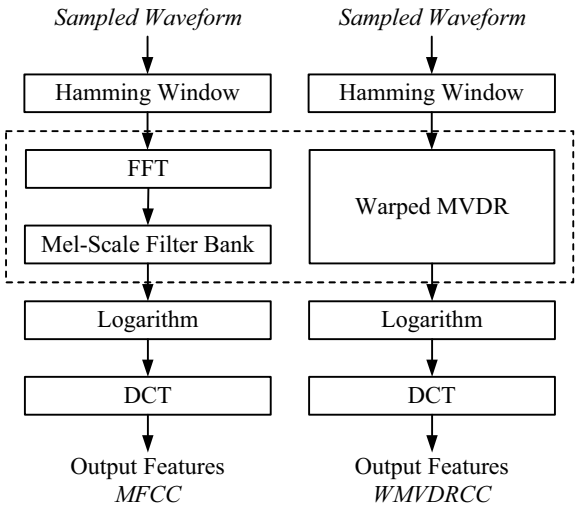

Figure 1: Flowchart of MFCC and WMVDRCC front-end.

While the Fourier transformation faithfully represents the spectral estimate, the MVDR spectral envelope provides an accurate description only for spectral peaks, which are in general less distorted by noise in the logarithmic magnitude domain than spectral valleys. No detailed spectral structure is extracted at the frequency regions with low signal energy. The influence of the free parameter, namely the model order which influences the provided spectral resolution, will be investigated in Section 3.1. To account for the non-linear pitch perception characteristics of the human ear the Fourier transformation is followed by a mel-scale filterbank while in the case of WMVDR, a bilinear-transformation in the time-domain is applied prior to spectral estimation.

3. Speaker Identification Experiments

The speaker identification experiments described below were conducted with the *Janus Recognition Toolkit* (JRTk), which is developed and maintained jointly at the Universität Karlsruhe (TH), in Karlsruhe, Germany and at the Carnegie Mellon University in Pittsburgh, Pennsylvania, USA.

Our baseline system is a mixture model-based classifier with 64 Gaussians, which is trained using K-means clustering followed by expectation maximization. Each of the two investigated front-ends provided 13 cepstral coefficients every 10 ms. The cepstral coefficients have been derived from spectral estimates with a frame size of 32 ms which have been obtained by either the Fourier transformation, or WMVDR spectral estimation. While the Fourier transformation is followed by a mel-filterbank which reduces the number of spectral bins from 129 to 30, in the case of WMVDR no filterbank is used and thus the number of spectral bins is not altered. To obtain the cepstral sequence a 129x13 cosine transformation matrix is applied to the spectrum.

To evaluate the performance of the speaker recognition system we use *accuracy*, which is defined as the percentage of correctly identified test trials.

3.1. Adjusting the Model Order

In our first set of experiments we want to optimize the provided resolution of the WMVDR spectral estimation by adjusting the model order to allow for the highest possible speaker identification accuracy. The influence of the model order on the spectral estimate of voiced speech is represented in Figure 2:

- A *higher model order* shows more detail of the fine structure of the spectrum and represents the fundamental frequency.
- A *low model order* reduces the influence of the excitation and is more or less a representation of the transfer function of the vocal track.

In order to find a good choice for the model order in a broad number of possible acoustic environments we decided to evaluate on the *far-field speaker identification database* (FarSID), which provides diverse acoustic environments. FarSID has been collected at Carnegie Mellon University to investigate the performance of speaker identification algorithms under adverse conditions [10] [11]. It consists of conversational speech recorded in face-to-face dialog sessions under two different reverberation conditions (small vs. large-sized room). For each reverberation condition, six noise conditions were applied by playing interfering signals (music, white noise and speech) at

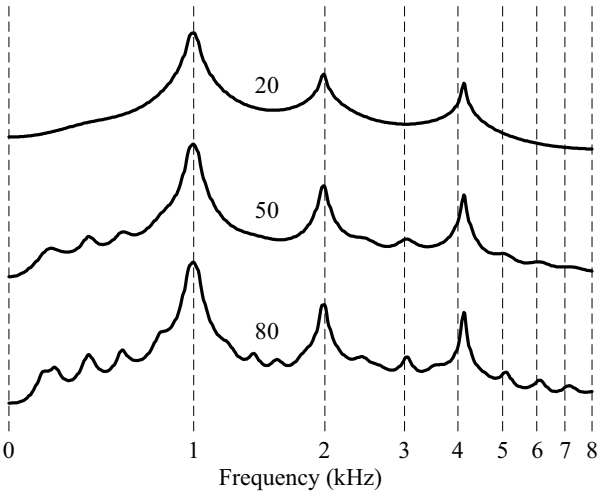

Figure 2: Influence of the model order on voiced speech (in the example the fundamental frequency is at 200 Hz).

different *signal-to-noise ratios*. The final set includes 10 native speakers of American English from both genders.

A duration of 60 seconds of speech per speaker is selected from the FarSID database for training while 30 seconds per speaker have been chosen for testing. The speaker models are trained and tested under all noisy conditions and channels separately. By comparing the overall performance of the speaker identification system for model orders between 20 and 120 with a step size of 10, Figure 3, we observe that the accuracy of the speaker identification system improves with an increase in model order until a model order of 70 is reached. For higher model orders, between 70 and 120, no clear trend can be observed.

It is interesting to note that for lower model orders the performance difference is relative high while for higher model orders the performance difference is relative low. This can be explained by the fact that the influence of the overall spectral resolution is higher for the same step size in the case of lower model orders as compared to higher model orders.

Comparing the optimal choice of model order for speech recognition which is around 30[1] with the optimal choice for speaker identification it becomes apparent that for speaker identification a significant higher model order is required.

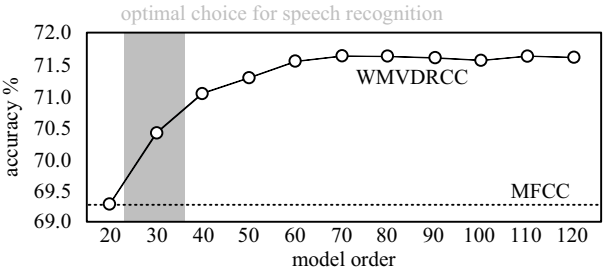

Figure 3: The overall identification accuracy for MFCC and WMVDRCC using different model orders.

[1]Note that in those experiments where a high model order is used for speech recognition, the spectral estimate is followed by a filterbank which adds additional smoothing.

3.2. Detailed Analysis for Different Distortion Types

In order to gain more insights of the influence of the various distortions, the speaker models are trained and tested under each noisy condition and channel on the FarSID dataset using the same experimental set-up as described in the previous section. We use the model order which gave the highest accuracy in the previous experiment. The graphs in Figure 4 show the identification accuracy using MFCC and WMVDRCC respectively. To provide a better overview of the results each sub-figure shows the results from training under one noise condition and testing under different noise conditions.

From this figure, we observe that WMVDRCC based front-ends improve speaker identification accuracy by up to 24% relative to MFCC. Its improvement differs for different background noises. Except for one tested condition, trained on speech interferred by music at 10 dB and tested on an interfering speaker, in all other conditions (also in clean training and testing) the WMVDRCC leads to improved speaker recognition performance. When training on distorted speech data the WMVDR front-end shows higher improvements in comparison to models trained on clean speech. In general it can be said that higher gains can be observed in more difficult conditions (with low accuracy rates) which underlines once more the robust estimation of speech spectra by MVDR spectral envelopes.

3.3. Clean Speech Experiments

The former set of experiments included speech signals which have been undergone various distortions. In this section we investigate the performance of the two front-end for clean speech signals. We decided to evaluate on the *naked scientist shows* [12] which have been originally chosen to investigate the speaker diarization task within the Quaero project. The shows are recorded in the studio, therefore the recorded speech is quite clean, approximately 30 dB SNR with low variation for different recordings. To perform speaker identification the total amount of 48 speakers appearing in the shows had to be reduced because each individual speaker time had to be at least 10.5 seconds. The final 31 speakers are from both genders. In order to obtain more test trials and to get more reliable results, we divide the whole dataset into 7 subsets, each part contained almost equal speaking time for each of the 31 speakers. There are at least 1.5 seconds for each speaker per subset. We then run the experiments by testing on each subset and training speaker models on the other 6 subsets. The training data for each speaker includes speech ranging from 9.0 seconds to 3,400 seconds. About 11,900 test trials in total are used for testing. Each test trial lasts about 1.5 seconds. The result of the experiments is shown in Table 1.

Cepstral Coefficients	Mel-Frequency	Warped MVDR
Accuracy	91.01%	91.28%
Relative Improvement	–	2.97%

Table 1: Speaker identification accuracy on clean speech by mel-frequency and warped MVDR cepstral coefficients.

In this experiment on clean speech data, the WMVDR-based front-end provides 3% relative improvement on identification accuracy. The proposed use of the WMVDR spectrum instead of mel-frequency, therefore, does not only provide increased robustness against various kinds of distortions, but also improves speaker recognition performance in the case of clean speech.

4. Conclusions and Outlook

In this paper it has been proposed to replace the widely used mel-frequency cepstral coefficients by warped minimum variance distortionless response cepstral coefficients for speaker identification. We found that the latter is not only superior in automatic speech recognition (as for example reported in [8]), but also superior in speaker recognition. In particular the WMVDR is capable to improve the performance on speaker recognition under adverse conditions due to its more robust extraction of relevant acoustic information. While a lower model order (around 30) is better for automatic speech recognition, a higher model order (above 70) is leading to an improved speaker identification accuracy. This result suggests that different feature extraction strategies should be considered for speech and speaker recognition. In the case of speech recognition the fundamental frequency should be suppressed (for non-tonal languages) while it seems helpful to preserve at least some information of the fundamental frequency for speaker recognition.

Further work will be separated into two research areas: speech feature extraction and speech feature enhancement. While for the former we try to further improve the extracted features by front-ends particularly designed for speaker identification, for the latter we want to adapt feature enhancement techniques which have already been developed for automatic speech recognition, e.g. the joint compensation of noise and reverberation as proposed in [13].

5. References

[1] D.A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91–108, 1995.

[2] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, Magrin I. Chagnolleau, S. Meignier, T. Merlin, Ortega J. Garcia, Petrovska Delacretaz, and Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 430–451, 2004.

[3] C.H. Lee, F.K. Soong, and K.K. Paliwal, *Automatic Speech and Speaker Recognition: Advanced Topics*, Kluwer Academic Publishers, Norwell, MA, USA, 1999.

[4] S. Furui, "Towards robust speech recognition under adverse conditions," in *ESCA Workshop on Speech Processing in Adverse Conditions*, 1992, pp. 31–42.

[5] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 28, pp. 357–366, Aug. 1980.

[6] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Jour. of Acoustic Society of America*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

[7] M.N. Murthi and B.D. Rao, "All–pole modeling of speech based on the minimum variance distortionless response spectrum," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 3, pp. 221–239, May 2000.

[8] M. Wölfel and J.W. McDonough, "Minimum variance distortionless response spectral estimation, review and refinements," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 117–126, Sept. 2005.

[9] M. Wölfel and J.W. McDonough, *Distant Speech Recognition*, John Wiley & Sons, 2009.

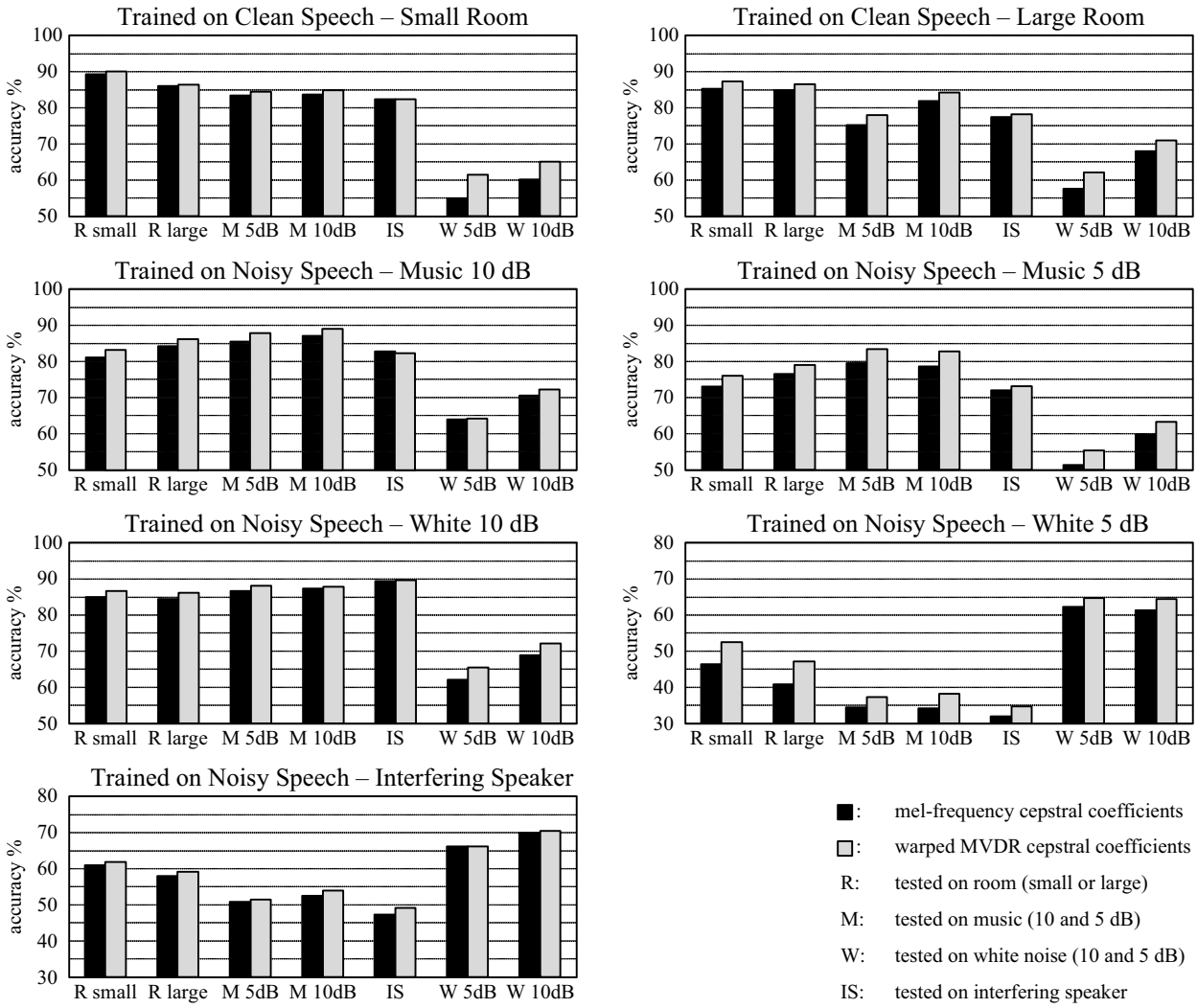

Figure 4: Identification accuracy for MFCC and WMVDRCC front-ends in various acoustic environments.

[10] Q. Jin and T. Schultz, "Robust far-field speaker recognition under mismatched conditions," in *Proc. of Interspeech*, 2008.

[11] Q. Jin, K. Kumar, T. Schultz, and R. Stern, "Compensation approaches for far-field speaker identification," in *NIST SRE Workshop*, 2008.

[12] "The naked scientists online," http://www.thenakedscientists.com/.

[13] M. Wölfel, "Enhanced speech features by single channel joint compensation of noise and reverberation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 312–323, 2009.