UNIVERSITY OF WEST BOHEMIA IN PILSEN
DEPARTMENT OF CYBERNETIC

Optimization of Features for Robust Speaker Recognition

Jan Vaněk, Aleš Padrta

E-mail: vanekyj@kky.zcu.cz , apadrta@kky.zcu.cz

Abstract

Currently, the old feature extraction method, which was used early for speech recognition, is used in speaker recognition in our speaker recognition group. Standard Mell Frequency Cepstral Coefficients (MFCC) features are used. They can be extended by delta and acceleration coefficients eventually. Whereas features for speech recognition has been evolved and optimized until now, features for speaker recognition remains same. These outdated features suffer from various deficiencies, regarding low robustness in particular. It can be said that these features are unsuitable in practical application. This study is aimed to examine possibilities of improving of the features. In conclusion then came up with suggestion of appropriate features extraction technique, which have been combined from examined method on the basis of the before explored methods. Main emphasis is placed on the robustness, i.e. noisy test data and/or channel disturbances (e.g. microphone mismatch). The study can be divided into several parts. At first, standard MFCC and Perceptual Linear Prediction (PLP) feature sets were optimized, i.e. the optimal numbers of the band filters and of the cepstral coefficients were examined. Next, the influence of delta and acceleration coefficients was discussed. Then, the channel normalization techniques were employed. Next, the possibilities of the linear transformations Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) were investigated. Then, the smoothing of spectrum or cepstrum in time was examined. Finally, several proposed combinations of above described approaches were tested. The new proposed features allow us to decrease the recognition error rate by 35-50%.

1. INTRODUCTION

All kinds of recognition work with some features. The quality and robustness of the features considerable affect the recognition. Speaker recognition task used the acoustic features extracted from the speech signal. Commonly used acoustic features are the same as the acoustic features used for speech recognition, but the characteristic of speech is different from the characteristic of speaker. Thus it is necessary to find acoustic features which are suitable for speaker recognition. This task is examined in this paper.

Next important thing is the robustness of acoustic features to the signal distortion. The performance of the state-of-art systems for speaker verification is acceptable for high quality signal, but the performance of the system drops much when any distortion is present in the signal. Thus the examination of the robustness of acoustic features is the second goal of this paper.

The paper is organized as follows: The performed experiments are described in Section 2. Next, in Section 3, the experimental results are presented. Finally, a conclusion is given.

2. DESCRIPTION OF THE EXPERIMENT

2.1. Speech data

Utterances from 100 speakers (64 male and 36 female) were used in our experiments. They were recorded in the same way as in the [1]. Each speaker read 24 sentences that were divided into three parts: 21 sentences of each speaker were used for training of the GMM of the speaker, 2 sentences were used for the construction of the background model, and 1 sentence was used for the tests.

Six test sets were prepared for testing of the robustness of the features. They were denoted as **A** to **F**. Each test set represents one typical distortion of the signal. These distortions were as follows:

A – Original data from the close talk microphone were used.
B – The noise with SNR from 15 to 20dB was added to the original data.
C – Channel distortion is applied on the original data.
D – Both noise and channel distortion like **B** and **C** were added.
E – Original data from the desktop microphone.
F – Out-of-database telephone data were used

2.2. Acoustic modeling

Several types of acoustic features were tested. All of them are based on MFCC and PLP. All utterances were resampled to 8~kHz and parametrized using a 32 ms-long Hamming window with a 10 ms overlap.

The models of the speakers and the background model were represented by Gaussian mixture models created using the HTK toolkit [2]. The model of each speaker consists of 5 Gaussian and the background model consists of 9 Gaussian. All models are trained from original data (data set A).

Six test sets were created, each represents one type of distortion and are denoted in the same way as the appropriate distortion (i.e. **A**, **B**, **C**, **D**, **E**, **F**).

2.3. Description of test

In order to find the best setting of MFCC and PLP according to the speaker verification task, series of the tests was performed.

Each test consisted of a set of verification trials. In each trial, a test utterance was verified against each speaker model. Since we had 100 test utterances and 100 models

of speakers, there were 100 × 100 = 10,000 verification trials in one test. 100 of the trials were the trials of the true speaker, the remaining 9,900 trials were impostor trials.

The performance of the tests can be measured by the detection error trade-off (DET) curve, which shows the value of false acceptance and the value of false rejection for various operating points of the verification system. At the point of the DET curve where the false rejection rate and the false acceptance rate are equal so-called equal error rate (EER) is defined. The EER values are used for evaluation of our tests, because EER is more suitable for the comparison of high amount of tests. In order to suitable comparison of used acoustic features, the overall criterion **J** was defined according to the formula

$$\mathbf{J} = \frac{\left(\frac{\mathbf{A}}{2}+\frac{\mathbf{B}}{30}+\frac{\mathbf{C}}{25}+\frac{\mathbf{D}}{30}+\frac{\mathbf{E}}{15}+\frac{\mathbf{F}}{40}\right)}{\left(\frac{1}{2}+\frac{1}{30}+\frac{1}{25}+\frac{1}{30}+\frac{1}{15}+\frac{1}{40}\right)} \quad [\%],$$

where **A, B, C, D, E, F** are the values of EER for appropriate test data set, the weights of individual EERs are derived from the baseline test results. The base acoustic features were the MFCC with 15 band-filters and 13 cepstral coefficients (include 0^{th} coefficient). Results of the baseline test are shown in Table 1.

Method	**A**	**B**	**C**	**D**	**E**	**F**
Baseline	2.00%	28.26%	24.26%	32.00%	17.96%	46.12%

Table1: Results of the baseline test

In Table 1, you can see that any mismatch of the testing data highly increases errors. These features have very low robustness and are practically unusable.

3. EXPERIMENTAL RESULTS

3.1. Optimizing standard MFCC and PLP coefficients

Standard acoustic features for speech recognition used 9-20 band filters and 7-15 cepstral coefficients. The differences between speakers are rather in detail of spectrum instead of differences between phonemes. Thus we need greater numbers of filters and coefficients to capture all this details [3]. In following experiments, we used number of filters 15, 17, 20, 25, 35 and number of coefficients 13, 15, 17, 20, 25 and 35 .

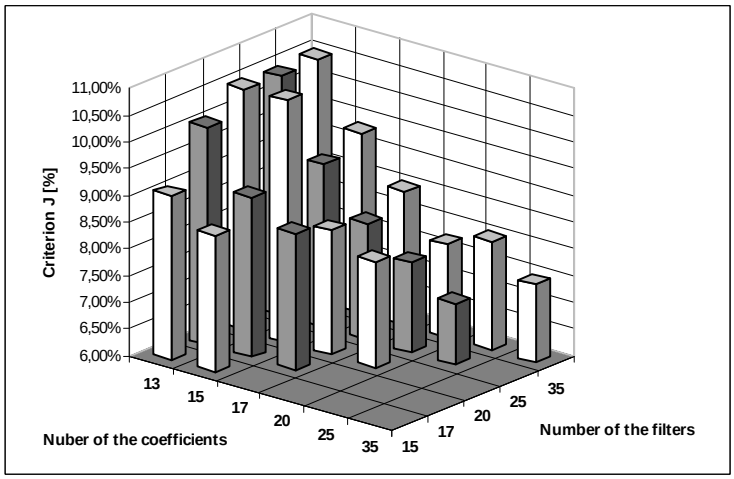

Figure 1**:** The dependence of the criterion **J** on the numbers of the filters and the coefficients for MFCC features.

The results for specified numbers of the band filters and the numbers of cepstral coefficients are depicted in Figure 1. You can see that higher number of cepstral coefficients reduces the error rate. The trends are similar for all number of filters. If we increase only numbers of filters then errors doesn't reduce, rather increase. Best result is observed for 25 filters and 25 coefficients. For PLP were the results very similar.

Method	**A**	**B**	**C**	**D**	**E**	**F**	**J**
MFCC 15f 13c	2.00%	28.26%	24.26%	32.00%	17.96%	46.12%	9.06%
MFCC 25f 25c	2.00%	22.22%	17.55%	24.55%	11.00%	39.73%	7.14%
MFCC 35f 35c	3.00%	17.92%	15.68%	20.00%	11.78%	41.32%	7.46%
PLP 15f 13c	3.00%	36.00%	29.36%	38.48%	14.38%	34.25%	9.98%
PLP 25f 25c	1.55%	26.00%	20.00%	29.02%	9.95%	40.70%	7.29%
PLP 35f 35c	2.09%	21.78%	15.91%	23.39%	10.00%	43.84%	7.09%

Table 2**:** The EERs for all test sets and overall criterion **J** for several numbers of band filters and cepstral coefficients for MFCC and PLP features.

In Table 2, there are depicted the representative results for MFCC and PLP. The mark "25f" means 25 band filters and mark "25c" means 25 cepstral coefficients. "MFCC 15f 13c" are the baseline acoustic features. It is clear, that it is necessary to use higher numbers of cepstral coefficients and therefore higher number of band filters.

3.2. Delta and acceleration coefficients

Delta and acceleration coefficients are often used to acquire the dynamics of the speech. Thus our next experiment was focused on utilization of delta (Δ) and acceleration (ΔΔ) coefficients for MFCC and PLP.

Method	A	B	C	D	E	F	J
MFCC	2.00%	28.26%	24.26%	32.00%	17.96%	46.12%	9.06%
MFCC +Δ	3.00%	24.07%	18.58%	28.64%	13.32%	45.21%	8.62%
MFCC +Δ+ΔΔ	2.54%	25.00%	15.53%	25.00%	14.00%	45.21%	8.05%
PLP	3.00%	36.00%	29.36%	38.48%	14.38%	34.25%	9.98%
PLP +Δ	2.98%	32.00%	23.00%	36.00%	14.00%	32.07%	9.18%
PLP +Δ +ΔΔ	4.00%	28.06%	18.52%	32.00%	16.29%	29.91%	9.42%

Table 3**:** The EERs for using delta and acceleration coefficients

The results are depicted in Table 3. You can see, that the delta coefficients lower ERR in most test sets, thus using of them is recommended. On the other side, the contribution of acceleration coefficients is various for individual test.

3.3. Channel normalization

Now, the channel normalization techniques were examined. At first, we focused on energy normalization. We tried three energy normalization techniques. At first, normalization of the input wave to maximum amplitude (denoted as NORM). Second, is well known Cepstral Mean Subtraction (CMS) [4], which was applied only on the 0^{th} cepstral coefficient, which represents the logarithm of the energy. Third is very simply - the zero-th cepstral coefficient is ignored (denoted as W.0c).

Method	A	B	C	D	E	F	J
Without norm	2.00%	28.26%	24.26%	32.00%	17.96%	46.12%	9.06%
NORM	2.74%	30.00%	16.00%	31.00%	17.30%	43.84%	9.05%
CMS on 0.coef.	2.00%	27.00%	14.82%	34.18%	14.10%	44.21%	8.13%
W.0c	2.20%	25.00%	14.00%	29.83%	12.45%	41.43%	7.67%

Table 4: The EERs for several energy normalization methods

As you can see in Table 4, the third method gives the best results. Thus, it is better to ignore the zero-th coefficient.

Method	A	B	C	D	E	F	J
MFCC W.0c	2.20%	25.00%	14.00%	29.83%	12.45%	41.43%	7.67%
MFCC W.0c CMS 1c	2.32%	26.00%	10.91%	30.00%	12.46%	41.96%	7.65%
MFCC W.0c CMS 1-2c	3.06%	22.84%	9.00%	27.54%	12.00%	39.73%	7.68%
MFCC W.0c CMS 1-3c	4.09%	23.49%	5.61%	26.45%	13.01%	38.36%	8.25%
MFCC W.0c CMS All	4.00%	29.87%	4.82%	31.00%	14.00%	30.17%	8.46%

Table 5**:** The EERs for CMS applied to different numbers of first cepstral coefficients.

Next, we applied CMS on next cepstral coefficients. The results are depicted in Table 5. (Note: CMS 1-2c means that cepstral mean is subtracted from the first to the second coefficient).

If the low mismatch between channels is expected then normalizing of the first several coefficients is preferable. If the high mismatch between channels is expected normalization to all coefficients has to be used.

3.4. Linear transformation

The linear transformation is used often in the speech or speaker recognition for greater discrimination of the recognized classes [5]. We used linear discriminant analysis (LDA) and principal component analysis (PCA). Both assume normal distribution of each class and equal covariance matrix for all classes. The assumptions are not completely satisfied, regardless this methods can improve the recognition rates. In our case, the transformation matrix was computed with using training data and no dimension reduction has been used.

Method	**A**	**B**	**C**	**D**	**E**	**F**	**J**
Without transformation	4.07%	14.06%	13.12%	23.00%	7.5%	42.62%	7.68%
LDA	3.00%	14.00%	10.00%	23.55%	8.10%	42.47%	6.81%
PCA	3.49%	17.00%	16.98%	28.80%	11.43%	43.47%	8.30%

Table 6**:** The EERs for LDA and PCA

Note that experiment in Table 6 has been set up differently and these results cannot be compared with other results out of this table.The training data match to set **A** and therefore the transformations have the greatest effect in the test set **A**. The results in Table 6 shows that suitable linear transformation can lower the error rate of recognition. Using of the PCA is unsuitable for our task.

3.5. Smoothing spectra or cepstra in time

It is well known that the vocal tract has a specific mass therefore it can move only with limited velocity. Environmental noise does not have any limitations therefore smoothing of the spectra, log-spectra or cepstra can suppress part of the noise [6]. We used the first-order Butterworth low-pass filter. The best results were achieved with the cut-off frequency equal 10Hz.

Method	**A**	**B**	**C**	**D**	**E**	**F**	**J**
Baseline	2.00%	28.26%	24.26%	32.00%	17.96%	46.12%	9.06%
Spectrum 10Hz	2.04%	27.80%	19.14%	29.00%	13.63%	43.84%	8.14%
Cepstrum 10Hz	1.98%	31.00%	19.28%	34.53%	15.00%	43.84%	8.65%

Table 7**:** The EERs for smoothing spectra or cepstra.

In Table 7, you can see that the smoothing spectra increase the robustness, has positive effect on all test sets and do not worse the results of set **A**. Smoothing log-spectra has identical results as smoothing cepstra, therefore is not depicted in Table 7.

3.6. New proposed features

And finally, we proposed the new features with using presented methods. We proposed four new features and these are denoted as **F1** to **F4**. All features have some identical attributes. These are same numbers of band filters (28) and cepstral coefficients (25). Spectral smoothing with 10Hz cut off frequency is applied and delta and acceleration coefficients are added to the all features. The **F1** and the **F2** were based on the MFCC, the **F1** included LDA. The **F3** and the **F4** was based on the PLP. In the **F4** was applied CMS to all static cepstral coefficients.

Method	**A**	**B**	**C**	**D**	**E**	**F**	**J**
Baseline	2.00%	28.26%	24.26%	32.00%	17.96%	46.12%	9.06%
F1	**1.24%**	11.28%	5.68%	13.13%	6.47%	39.85%	**4.42%**
F2	2.25%	**9.00%**	6.12%	**11.00%**	5.66%	38.96%	4.85%
F3	3.00%	10.00%	7.87%	16.36%	**4.51%**	34.25%	5.51%
F4	3.92%	12.98%	**5.32%**	13.88%	6.00%	**15.07%**	5.51%

Table 8**:** The EERs for new proposed features.

In Table 8, you can see that new proposed features can considerable lower EER. Each of proposed features is suitable for specific test set (i.e. type of distortion). Thus ideal features for all circumstances do not exist.

CONCLUSION

We tested several different acoustic features in order to increase its robustness. Main conclusions follow: Speaker recognition task requires higher numbers of band filters and higher number of cepstral coefficients. Zero-th cepstral coefficient is not suitable for speaker recognition. Static coefficient should be augmented by delta coefficients. Smoothing spectra in time is positive.

Regardless, the proposal of the universal acoustic features for all types of signal distortion is not recommended. Specialized acoustic features outperform the universal acoustic features for specific distortion. It means that it would be better to detect the type of distortion and then apply the appropriate specialized acoustic features.

ACKNOWLEDGEMENT

This work was supported by the Ministry of Education of the Czech Republic, project No. MSM235200004 and the Grant Agency of the Czech Republic, project No. 102/02/0124.

REFERENCES

[1] V. Radová, J. Psutka, „ UWB_S01 Corpus - A Czech Read-Speech Corpus“, Proc. of the ICSLP 2000, pp. 732--735, Beijing, China, 2000.

[2] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, P. Woodland, „ The HTK Book. Revised for HTK Version 3.1.1“, July 2002. http://htk.eng.cam.ac.uk

[3] H. A. Murthy,, F. Beaufays, L. Heck, M. Weintraub, „Robust Text-Independent Speaker Identification over Telephone Channels“,IEEE Transactions on Speech and Audio Processing, vol. 7, no. 5, pp. 554-568, September 1999.

[4] A. E. Rosenberg, C. H. Lee, F. K. Soong, "Cepstral Channel Normalization Techniques for HMM-Based Speaker Verification", Proc. of the ICSLP 1994.

[5] X. Wang, D. O'Shaughnessy, „Improving the Efficiency of Automatic Speech Recognition by Feature Transformation and Dimensionality Reduction“, EUROSPEECH 2003, Geneva, Switzerland 2003.

[6] S. Vuuren, H. Hermansky, „!Mess: A Modular, Efficient Speaker Verification System“, Speaker Recognition and its Commercial and Forensic Applications, pp. 189-201, France, April 1998.