BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **10**, No 1

Sofia • 2010

Cepstral Features and Text-Dependent Speaker Identification – A Comparative Study

Atanas Ouzounov

Institute of Information Technologies, 1113 Sofia
E-mail: atanas@iinf.bas.bg

Abstract: *In the study, the effectiveness of combinations of cepstral features, channel compensation techniques, and different local distances in the Dynamic Time Warping (DTW) algorithm is experimentally evaluated in the text-dependent speaker identification task. The training and the testing has been done with noisy telephone speech (short phrases in Bulgarian with length of about 2 seconds) selected from the BG-SRDat corpus. The employed cepstral features are – Linear Predictive Coding derived Cepstrum (LPCC), Mel-Frequency Cepstral Coefficients (MFCC), Adaptive Component Weighted Cepstrum (ACWC), Post-Filtered Cepstrum (PFC) and Perceptually Linear Predictive coding derived Cepstrum (PLPC). Two unsupervised techniques for channel compensation are applied – Cepstral Mean Subtraction (CMS) and Relative Spectral (RASTA) technique. In the DTW algorithm two cepstral distances are utilized – the Euclidean and the Root Power Sum (RPS) distance. The experiments have shown that the best recognition rate for available noisy speech data was obtained by using the combination of the MFCC, CMS and the DTW-RPS distance.*

Keywords: *Cepstral analysis, speaker identification, dynamic time warping.*

1. Introduction

Automatic speaker recognition methods can be divided into two groups: text-independent and text-dependent methods. In text-independent methods, the speaker identity must be recognized without any information about the lexical content of analyzed speech. On the other hand, in text-dependent methods, the speaker has to utter only known to the recognition system words or phrases. In the case of the pure text-dependent recognition (also known as the fixed-text recognition), the speaker has to say the same word or phrase in the training and recognition modes [3].

The speaker recognition is a general term which embraces two tasks – speaker identification and speaker verification. The closed-set speaker identification (task chosen in this study) is a classical recognition task – the unknown speaker is associated with the speaker from the known speakers set whose model is the best-matched model for the unknown speaker test utterance. In the speaker verification task, the claimed speaker identity must be accepted or rejected based on a threshold comparison criterion [3].

Three speaker-modelling techniques: Dynamic Time Warping (DTW), Hidden Markov's Models (HMM) and VQ (Vector Quantization) are used in the text-dependent speaker recognition task and their performances are comparatively analyzed in [18]. The authors in [18] claim that for text-dependent case and limited amount of training data the DTW (with Linear Predictive Coding derived Cepstrum (LPCC) feature) outperforms the rest two approaches. For more data, the performances of the used algorithms are comparable to each other.

The effectiveness of different combinations of three cepstral features: LPCC, Mel-Frequency Cepstral Coefficients (MFCC) and Post-Filtered Cepstrum (PFC) and 3 speaker-modelling techniques: DTW, Gaussian Mixture Models (GMM) and Multi Layers Perceptron (MLP) is analyzed experimentally in [16]. In this study, the text-dependent speaker identification experiments are carried out with speech database that contains the recordings of the isolated digits 0-9 collected over 50 speakers. The experiments in [16] have shown that the highest identification rate is achieved by the GMM with MFCC feature.

It is known that the typical scenario in the area of the text-dependent speaker recognition is the speaker verification [3, 18]. However, the speaker verification paradigm includes techniques for thresholds settings and these techniques complicate the analysis of the recognition performance as a result of the feature extraction approaches and classification parameters settings. It seems reasonable to analyze in advance the recognition performance as a function of different parameters in a pure recognition task (for instance, the closed-set speaker identification) and subsequently to use this information in a forthcoming speaker verification research. This is possible because the performance trends in the speaker identification are usually applicable to the verification task [18].

The focus of this study is to evaluate experimentally with noisy telephone speech the effectiveness of different combinations of cepstral features, channel compensation methods, and different local distances in the DTW-based text-dependent speaker identification with fixed phrase. The speech data used in the experiments are selected from the BG-SRDat corpus [12]. These data are short phrase in Bulgarian with length of about 2 seconds recorded over noisy telephone channels and collected over 12 speakers.

In the study the employed cepstral features are – LPCC, MFCC, Adaptive Component Weighted Cepstrum (ACWC), PFC and Perceptually Linear Predictive coding derived Cepstrum (PLPC). Two unsupervised techniques for channel compensation are applied – Cepstral Mean Subtraction (CMS) and Relative Spectral (RASTA) technique. In the DTW algorithm two cepstral distances are utilized – the Euclidean and the Root Power Sum (RPS) distance.

2. Cepstral features

2.1. LPCC

To calculate the LPCC, the Linear Predictive Coding (LPC) coefficients must be first calculated. Then the cepstral coefficients $c_{\mathrm{LPC}}(m)$ can be computed by the following recursion [14]:

$$c_{\mathrm{LPC}}(m) = \begin{cases} -a(m) - \sum_{i=1}^{m-1} (1 - \frac{i}{m}) a(i) c_{\mathrm{LPC}}(m-i), & 1 \le m \le P, \\ -\sum_{i=1}^{m-1} (1 - \frac{i}{m}) a(i) c_{\mathrm{LPC}}(m-i), & m > P, \end{cases} \tag{1}$$

where $a(m),\ m = 1, ..., P,$ are LPC coefficients and P is the model order.

2.2. MFCC

The often-used Mel-cepstrum is based on the Fourier power spectrum and its coefficients are the Mel-frequency cepstral coefficients. In this case, the Fourier power spectrum of speech signal is filtered by band pass filters (with triangular form) placed along the Mel-frequency scale (linear up to 1000 Hz and logarithmic above). To obtain the Mel-cepstrum, the cosine transform is applied to the logarithmic output of the filters [14].

If the logarithmic energy on the output of the *k*-th filter is $\log(E_k)$ then the Mel-cepstrum $c_{\mathrm{MEL}}(m)$ is

$$c_{\mathrm{MEL}}(m) = \sum_{k=1}^{K} \log(E_k) \cos(m(k - 0.5)\pi / K), \tag{2}$$

where K is the number of the band-pass filters, m=1,…, M is the cepstral coefficients index [14].

2.3. ACWC

The all-pole model for a given speech frame can be expressed in parallel form by means of partial expansion [2]

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 + \sum_{i=1}^{P} a_i z^{-i}} = \sum_{i=1}^{P} \frac{r_i}{(1 - z_i z^{-1})}, \tag{3}$$

where $a_i,\ i = 0,..., P,$ are LPC coefficients and P is the model order, r_i are residues of the poles and z_i represents the center frequency and the bandwidth of the *i*-th component of the LP model. It is observed in [2] that the residues are highly sensitive to the channel effects. In the ACW cepstrum proposed in [2] the variations caused by channel variability are removed by residues normalization.

The ACW spectrum (all residues are set to be equal to unity) is proposed in [2] and it is in the form

(4) $$\hat{H}(z) = \sum_{i=1}^{P} \frac{1}{(1 - z_i z^{-1})} = \frac{N(z)}{1 + \sum_{i=1}^{P} a_i z^{-1}},$$

where

(5) $$N(z) = P\left(1 + \sum_{i=1}^{P-1} b_i z^{-i}\right).$$

According to [2] the ACWC $c_{\mathrm{ACW}}(m)$ is

(6) $$c_{\mathrm{ACW}}(m) = c_{\mathrm{LPC}}(m) - c_n^b(m),$$

where $c_{\mathrm{LPC}}(m)$ is the LPCC, $m > 0$ and $c_n^b(m)$ can be computed by a recursion using the coefficients $\{b_i\}$ [2, 15].

The LP power spectrums of the two voiced frames (vowel *e*) are shown in Fig. 1. These voiced frames are selected from two utterances of the same phrase. The utterances are obtained from the same speaker but they are recorded from different telephone calls. In Fig. 1, there is a large mismatch between spectrums regardless that some spectral peaks are nearly at the same frequencies.

The ACW spectrums of the same signals are shown in Fig. 2. It is evident that the processing proposed in [2] reduces significantly the mismatch between two spectrums. It is interesting to note that in Fig. 2 there is no apparent spectral tilt.

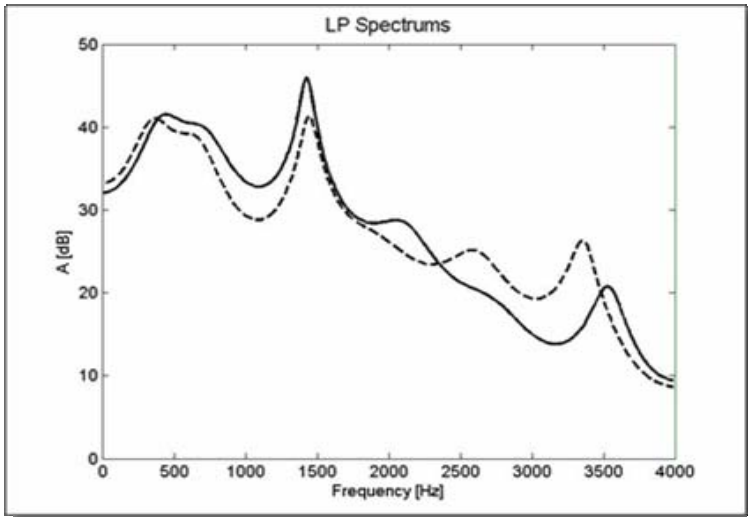

Fig. 1. LP spectrums for vowel *e*

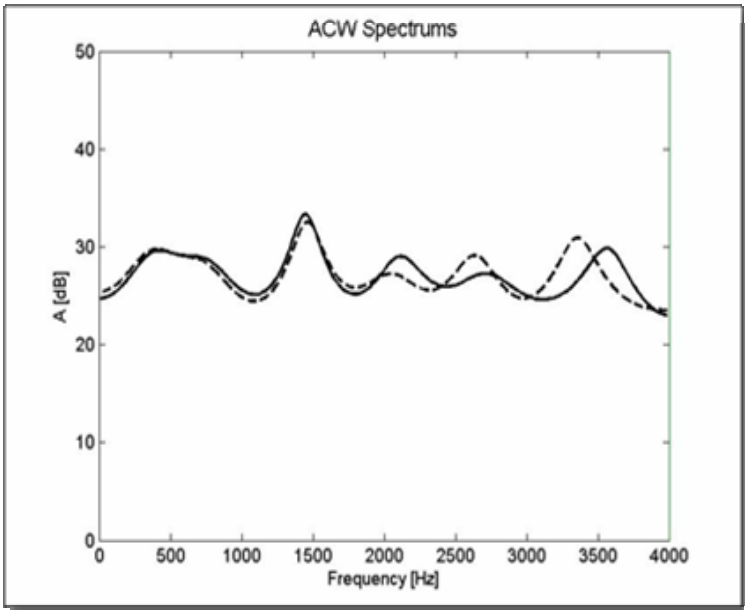

Fig. 2. ACW spectrums for vowel *e*

2.4. PFC

The post-filtered cepstrum $c_{\mathrm{PFC}}(m)$ is obtained in the form $c_{\mathrm{PFC}}(m) = c_{\mathrm{LPC}}(m)[\alpha^m - \beta^m]$, where $\alpha = 1$, $\beta = 0.9$, $m > 0$ and $c_{\mathrm{LPC}}(m)$ is the LPCC. The post-filtered cepstrum includes a subtractive component that can be thought as a channel cepstral estimate and which is adaptive on a frame basis [10].

2.5. PLPC

In this approach, some processing steps based on characteristics of human ear are applied on the speech power spectrum to produce the auditory spectrum. Then this spectrum is approximated by the spectrum of the linear prediction all-pole model.

To obtain the auditory spectrum from the power spectrum the following processing steps are executed:

- critical-band filtering in the frequency domain;
- frequency correction (preemphasis) with equal-loudness curve;
- amplitude compression by intensity-loudness power law.

The inverse Fourier transform on the auditory spectrum yield the autocorrelation function. Then the values of autocorrelation function are used by the Levinson algorithm to estimate the linear prediction coefficients. The cepstral coefficients of the PLPC are obtained later from them [7].

3. Channel compensation techniques

Two often-used unsupervised techniques for channel compensation – the CMS and the log-RASTA are included in this study. These techniques are distinguished by the fact that they do not explicitly use any channel information.

3.1. Cepstral mean subtraction

In many cases, the communication channel can be approximated by a linear system. Therefore, the channel influence on the speech can be represented in the cepstral domain through an additive component to the cepstrum of the clean speech. It is supposed that the cepstral mean of the clean speech is zero. In this case, to compensate the channel effect, the channel cepstrum can be removed by subtraction of the cepstral mean. This temporal mean is a rough estimate of the channel response. Despite all, this approach is widely used in the speaker recognition system now [4].

3.2. RASTA

The relative spectral analysis technique (RASTA) is based on the idea that the rate of changing of the short-term spectrum for linguistic and non-linguistic components in speech is different [8]. This means that the spectral components of the communication channel vary more quickly or more slowly than the spectral components of the speech and they could be separated (filtered). The core part of RASTA processing is a band-pass filtering of the spectral parameters trajectories by an IIR filter. The convolutive (in the time domain) distortions in the communication channel can be reduced by using the RASTA filtering in the logarithmic domain (spectral or cepstral). The RASTA approach can be combined with the perceptually linear prediction method (so called PLP-RASTA approach) or can directly be applied to the cepstral trajectories [8].

4. Dynamic time warping

In the study, the DTW algorithm called the normalize-wrap method is applied [11]. In this algorithm, the length normalization on both the reference and test pattern is used before performing the actual DTW algorithm. In the DTW, the relaxed endpoints constraints, Itakura's form of local constraints and Euclidean and Root Power Sum cepstral distances as local distances are implemented [9, 11, 19].

5. Experiments, results and discussions

The speech data are selected from the BG-SRDat corpus [12]. This corpus is in Bulgarian language and it is recorded over noisy telephone channels and intended for speaker recognition. The speech data is collected from different types of

telephone calls and various acoustical environments. The data are sampled with frequency of 8 kHz at 16 bits, PCM format, and mono mode.

The speech data used in the study include 261 records of a short phrase (with length of about 2 seconds) collected from 12 speakers (male). Each speaker utters the phrase at least 15 times. For training are used 120 utterances or 10 utterances per speaker for his reference model (template) creation. The rest of data – 141 utterances (some speakers possess more than 10 utterances for test) are used in testing mode. In the study, the 5-fold cross-validation algorithm for data selection is implemented.

In the pre-processing step, the preemphasis is not applied. Hamming windowing frames of 32 ms are utilized, with frame rate of 10 ms. The autocorrelation method for linear prediction is used and for each frame is computed the LPCC with 14 coefficients. The number of the PLP and the MFC cepstral coefficients is 14, too. These cepstral coefficients are calculated by using of 17 critical-band filters and 24 Mel-frequency spaced filters, respectively. The zeroth cepstral coefficients are not used.

To avoid processing of non-speech parts in the signal (located before and after the recorded phrase) the endpoint detection is applied. The endpoint detection algorithm is based on the previous research of the author [13]. It is worth to note that in the experiments are used all frames between the phrase endpoints. No additional frames selection is applied.

In the DTW algorithm the reference is placed along the Y-axis and the path width is set at 300 ms. The speaker's reference is obtained by averaging (after dynamic time warping alignment) of his training utterances [11, 19].

In text below the combinations of different processing algorithms are noted with the acronyms. For instance, the acronym LPCC-CMS-ACWC placed in the Table 1 means the sequential calculations, first the LPCC for each frame, next the CMS over the whole phrase and finally the ACW cepstrum for each frame [1].

In the study are experimentally evaluated three RASTA filtered cepstrums – PLP-RASTA, MFCC-RASTA and LPCC-RASTA. The last two cepstrums are obtained by applying of the RASTA filtering directly on their trajectories. The temporal filtering of the cepstral trajectories is often used approach for channel effects reducing in the cepstral features-based speech and speaker recognition systems. The RASTA filtering is performed typically on the MFCC trajectories. That way of direct filtering is explained in [5, 6] and is based on the linear relationship between Mel-frequency log spectrum and the MFCC. This relationship allows the applying of the RASTA filter directly on the MFCC trajectories. The MFCC, filtered in this way, are named in the text as MFCC-RASTA. The LPC cepstral trajectories are also filtered by the RASTA filter and the obtained cepstrum is named as LPCC-RASTA.

In the study, the log-RASTA filtering with non-zero initial conditions is used. In this case, an additional speech snatch with length of about 200 ms is located before the utterance-starting sample provided by the endpoints detector. The parameters of the RASTA filter are selected according to the recommendations in [8].

In Table 1 the identification results are shown – the overall accuracy and the average half width of the 95% confidence interval for each feature combination and for each DTW local distance. In the study, the overall accuracy of the classification is calculated by the trace of the confusion matrix, divided by the sum of the elements in the matrix.

Table 1. The identification results

No	Features combinations	DTW local distance			
		Euclidean		RPS	
		Accuracy	95% CI	Accuracy	95%CI
1	LPCC-CMS-ACWC	0.8865	0.0234	0.8822	0.0238
2	LPCC-CMS	0.8028	0.0293	0.8907	0.0230
3	MFCC-CMS	0.8127	0.0288	0.9106	0.0211
4	LPCC-CMS-PFC	0.8751	0.0244	0.8921	0.0229
5	PLPC-CMS	0.7134	0.0333	0.8156	0.0286
6	LPCC-RASTA	0.8014	0.0278	0.8609	0.0236
7	MFCC-RASTA	0.7971	0.0280	0.8751	0.0224
8	PLP-RASTA	0.7007	0.0326	0.8595	0.0237

The goal of this research is to examine experimentally different combinations of cepstral features, channel compensation techniques, and DTW local distances in recognition tasks with short and noisy speech data. The short length of the phrase and the fact that it is phonetically unbalanced (more exactly, the consonants predominate over the vowels [12]) complicate the recognition process.

These experiments revealed that the log-RASTA filtering did not outperform the CMS as channel compensation technique in the DTW-based fixed-text speaker identification task. In comparison with the CMS, as can be seen in the Table 1, the RASTA filtering always provides lower accuracy for the Euclidean distance cases. However, for the RPS distance this filtering produces ambiguous results (higher accuracy for the PLP-RASTA and lower one for the MFCC-RASTA and LPCC-RASTA).

It is necessary to note that the cepstral mean in the CMS technique is computed over short phrase with length of about 2 seconds. It is known that the accuracy of the channel cepstrum (i.e. the cepstral mean) estimation depends on the amount of speech data. For short length data, it is most likely that the cepstral mean is an inaccurate estimate of the channel cepstrum [4]. Nevertheless, the highest accuracy in the study is obtained with the feature combination MFCC-CMS (for the DTW-RPS distance).

The RPS distance is the cepstral distance between two index-weighted cepstral vectors. It is known that index weighting in cepstral domain reduces the influence of low-order cepstral coefficients that convey the information for spectral slope [17]. The use of the DTW-RPS distance increases the accuracy for all feature combination with the exception of the LPCC-CMS-ACWC. Probably the twofold effect on the spectral slope, firstly by the ACW algorithm, and secondly by the cepstral index weighting in the DTW-RPS distance results in undesirable

enhancement of the high-order cepstral coefficients and causes the error rate increasing for this feature combination.

In the study, it is found that the log-RASTA filtering is inefficient for available speech data. Its performance level was worse than the CMS one. It is evident that the RPS distance has substantial effect on the recognition rate. In other words, the suppression of the speech spectrum tilt by the RPS distance has more influence on the final recognition rate than the used cepstrum-based channel compensation techniques.

6. Conclusions

In the study, the effectiveness of combinations of cepstral features, channel compensation techniques, and different local distances in the DTW algorithm is experimentally evaluated in the fixed-text speaker identification task with short phrases of telephone speech. Based on the experimental results the following conclusions are made:

– the highest accuracy of 0.9106 is achieved for the feature combination MFCC-CMS and the DTW-RPS distance;

– when RASTA filtering is applied the highest accuracy of 0.8751 is achieved for the MFCC-RASTA and the DTW-RPS distance;

– when the DTW-Euclidean distance is used, the highest accuracy of 0.8865 is obtained for the feature combination LPCC-CMS-ACWC;

– the DTW-RPS distance has more substantial effect on the recognition rate than the channel compensation techniques.

Future work will focus on two main objectives – the evaluation of different algorithms, (e.g., the HMM and the MLP) in the same experimental framework and the examination of some modifications of the channel compensation techniques to obtain better results with short and noisy speech data.

References

1. Alonso-Martinez, C., M. Faundez-Zanuy. Speaker Identification in Mismatch Training and Testing Conditions. – In: Proc. of EEE ICASSP, Vol. **II**, 2000, 1181-1184.
2. Assaleh, K., R. Mammone. New LP-Derived Features for Speaker Identification. – IEEE Transactions on Speech and Audio Processing, Vol. **2**, 1994, No 4, 630-638.
3. Campbell, J. P. Speaker Recognition. Biometrics: Personal Identification in Networked Society. Anil Jain Ed., Boston, Kluwer Academic Publishers, 1999, 165-189.
4. De Veth, J., L. Boves. Channel Normalization Techniques for Automatic Speech Recognition over the Telephone. – Speech Communication, **25**, 1998, 149-164
5. Han, J., M. Han, G. Park, J. Park, W. Gao. Relative Mel-Frequency Cepstral Coefficients Compensation for Robust Telephone Speech Recognition. – EUROSPEECH'97, 1531-1534.
6. Han, J., W. Gao. Robust Telephone Speech Recognition Based on Channel Compensation. – Pattern Recognition, **32**, 1999, 1061-1067.
7. Hermansky, H., B. Hanson, H. Wakita. Perceptually Based Linear Predictive Analysis of Speech. – In: Proc. of the IEEE ICASSP, 1985, 13.10.1-13.10.4.
8. Hermansky, H. RASTA Processing of Speech. – IEEE Transactions on Speech and Audio Processing, Vol. **2**, 1994, No 4, 578-589.

9. Itakura, F. Minimum Prediction Residual Principle Applied to Speech Recognition. – IEEE Transactions on ASSP, Vol. **23**, 1975, No 1, 67-72.
10. Mammone, R., X. Zhang, R. Ramachandran. Robust Speaker Recognition. – IEEE Signal Processing Magazine, September 1996, 58-71.
11. Myers, C. et all. Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Word Recognition. – IEEE Transactions on ASSP, Vol. **28**, 1980, No 6, 623-635.
12. Ouzounov, A. BG-SRDat: A Corpus in Bulgarian Language for Speaker Recognition over Telephone Channels. – Cybernetics and Information Technologies, Vol. **3**, 2003, No 2, 101-108.
13. Ouzounov, A. Further Results on Speaker Identification using Robust Speech Detection and Neural Network. – Cybernetics and Information Technologies, Vol. **9**, 2009, No 1, 37-45.
14. Picone, J. Signal Modelling Techniques in Speech Recognition. – In: Proc. of IEEE, Vol. **81**, 1993, No 9, 1215-1247.
15. Swanson, A., R. Ramachandran, S. Chin. Fast Adaptive Component Weighted Cepstrum Pole Filtering for Speaker Identification. – In: Proc. of IEEE ISCAS'2004, V.612-V.615.
16. Tanprasert, C., V. Achariyakulporn. Comparative Study of GMM, DTW, and ANN on Thai Speaker Identification System. – In: Proc. of the ICSLP'2000, 234-237.
17. Tohkura Y. A Weighted Cepstral Distance Measure for Speech Recognition. – In: IEEE Transactions on ASSP, Vol. **ASSP-35**, 1987, No 10, 1414-1422.
18. Yu, K., J. Mason, J. Oglesby. Speaker Recognition Using Hidden Markov Models, Dynamic Time Warping and Vector Quantization. – In: IEE Proc. on Vision, Image and Signal Processing, Vol. **142**, 1995, No 5, 313-318.
19. Zelinski, R., F. Class. A Learning Procedure for Speaker-Dependent Word Recognition System Based on Sequential Processing of Input Tokens. – ICASSP'83, 1053-1056.