



Speaker Characterization Using Long-Term and Temporal Information

Chien-Lin Huang, Hanwu Sun, Bin Ma, Haizhou Li

Human Language Technology Department,
 Institute for Infocomm Research, A*STAR, Singapore 138632
 {clhuang, hwsun, mabin, hli}@i2r.a-star.edu.sg

Abstract

This paper presents new techniques for front-end analysis using long-term and temporal information for speaker recognition. We propose a long-term feature analysis strategy that averages short-time spectral features over a period of time in an effort to capture the speaker traits that are manifested over a speech segment longer than a spectral frame. We found that the moving averages of temporal information are effective in speaker recognition as well. The experiments on the 2008 NIST Speaker Recognition Evaluation dataset show the long-term and temporal information contribute to substantial EER reductions.

Index Terms: speaker recognition, long-term feature, temporal information

1. Introduction

We have seen an increase of demand of speaker recognition technology in telephony applications [1], such as speaker identification and speaker verification. The state-of-the-art text-independent speaker verification used signal processing and statistical modeling techniques to characterize speakers. Speaker verification is typically formulated as a hypothesis test problem [2]. There are three major components: feature analysis, statistical modeling and verification decision.

The conventional short-time spectral features, such as Mel-frequency cepstral coefficients (MFCC), are useful acoustic features for speaker verification. Many efforts have been devoted to improving the effectiveness of MFCC, such as reducing the dimensionality, enhancing discriminative ability [3], and characterizing speakers with temporal features [4]. In this paper, we continue the study on feature analysis for effective and efficiency speaker characterization.

The moving average of temporal information is evidently good for the application of speech recognition [5]. Motivated by the findings in temporal features, we further the studies by proposing feature extraction using the mean of long-term feature approach. The method of averaging short-time spectral features in a long-time window captures the spectral statistics over a long period of time. In this way, it also reduces the amount of data involved in the modeling, testing, and thus the computational cost. We conduct the experiments on NIST 2008 Speaker Recognition Evaluation (SRE) dataset in the GMM-SVM framework [6].

The rest of this paper is organized as follows. Section 2 presents the proposed approach. The experimental results and analysis are presented in Section 3. Finally, Section 4 concludes this work.

2. Long-Term and Temporal Features

Feature extraction is an important process to estimate a numerical representation from speech samples and to characterize the speakers. The proposed feature extraction using long-term and temporal information is based on Mel-

cepstral analysis as shown in Fig. 1. MFCC is an effective acoustic feature for speaker recognition, where perceptually motivated Mel-frequency scaling is adopted. The standard Mel-filter bank $M(f)$ is used to reduce the correlations between the frequency sub-bands.

$$M(f) = 2595 \log_{10} \left(1 + \frac{f}{700} \right) \quad (1)$$

The Mel-frequency cepstral coefficient c^m is estimated with

$$c^m = \sum_{n=1}^N e[n] \times R_m[n], \quad m = 1, 2, \dots, M \quad (2)$$

where $e[n]$ denotes log filter-bank energies in the spectral domain. M is the number of cepstral coefficients. N is the number of filters in the Mel-filter bank. N is chosen according to the range of frequency band of applications. In this study, the speech signal is divided into 18 sub-bands between 250 Hz to 3500 Hz using the Mel-filter bank to make spectral contents similar to that of the telephone channel. In order to decorrelate the resulting feature vectors thus reducing redundancy, a discrete cosine transformation (DCT) is performed.

$$R_m[n] = \sqrt{\frac{2}{N}} \cos \left(\frac{\pi m}{N} (n - 0.5) \right) \quad (3)$$

where $R_m[n]$ is a DCT matrix used for the spectral-cepstral transformation.

2.1. Moving averages of temporal information

The moving averages of temporal information have been shown useful in speech recognition [5]. In this study, the auto-regression moving average filter (ARMA) is used in the cepstral domain. The ARMA process is a low-pass filter, that smoothes spikes in the time sequence thereby reducing noisy and defined by

$$\tilde{c}_t = \frac{1}{2\alpha + 1} \left(\sum_{i=1}^{\alpha} \tilde{c}_{(t-i)} + \sum_{j=0}^{\alpha} c_{(t+j)} \right) \quad (4)$$

where α is the order of the ARMA process. In theory, with a small value of α , we retain the short-time cepstral information which is however more vulnerable to noise; on the other hand, with a large value of α , the features tend to be more robust against noise, while at the cost of losing some short-time details. We believe that ARMA process allows us to retain temporal information that is unique to a speaker. There is an inherent trade-off to decide the order of α in the

10.21437/Interspeech.2010-133

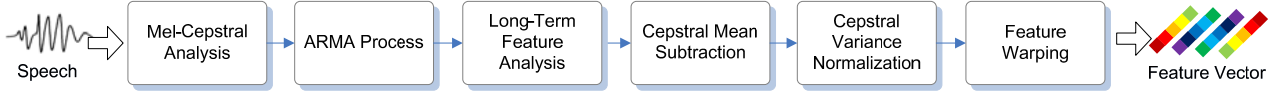


Figure 1: A diagram of feature extraction using long-term and temporal information.

ARMA process [5]. We will study the selection of α in the experiments.

2.2. Long-term feature analysis

In feature analysis, speech signal is represented as a sequence of frames for short-time analysis. These frames are small enough to ensure the frequency characteristics of the magnitude spectrum are relatively stable. However, speech timbre and prosody are manifested over a speech segment of multiple short-time spectrums through phonetic units, such as vowel and consonant [7]. To capture the spectral statistics over a long period of time, this paper proposes a way to analyze the short-time spectral features over a long-time window, that we call LTF. The mean of long-term feature (LTF) is estimated with the mean of multiple short-time spectral features. Moreover, this transformation results in a more compact feature vector for statistical modeling as shown in Fig. 2.

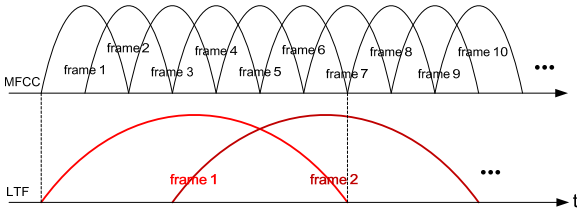


Figure 2: Illustration of feature analysis using the mean of short-time spectral features in a long-time window.

The overlapping long-time windows are applied on the short-term features, reducing short-term MFCC frames J to LTF frames K , with $K = (J - L) / Z + 1$. L denotes the size of the long-time window and Z is the step of the long-time window shift. With long-term feature analysis, a sequence of short-time feature vector \tilde{c}_i is represented as long-term feature vector \hat{c}_i as following

$$\hat{c}_i = \frac{1}{L} \sum_{i=(i-1) \times Z}^{(i-1) \times Z + L} \tilde{c}_i \quad (5)$$

The LTF is used to obtain a new representation of MFCC feature analysis which is more compact and suitable for statistical modeling. The purpose of LTF is to take account of short-time frequency characteristics and long-time resolution at the same time.

2.3. Feature normalization and visualization

The feature normalization is applied to reduce the noise and channel effects after the long-term feature analysis. The first step of feature normalization is the cepstral mean subtraction (CMS) defined as

$$c'_i = c_i - \mu, \quad \mu = \sum_{i=1}^T c_i / T \quad (6)$$

where μ denotes a mean vector. This normalization is used to remove the global shift of the cepstral vectors. The process of CMS compensates for the main effect of channel distortion and some of the side effects of additive noise [8]. The second step of feature normalization is the cepstral variance normalization (CVN) defined as follows

$$\hat{c}_i = c'_i / \sigma, \quad \sigma = \sqrt{\sum_{i=1}^T (c_i - \mu)^2 / T} \quad (7)$$

where σ is an estimate of the standard deviation. The cepstral mean subtraction and cepstral variance normalization are applied for slowly varying convolutional noises [9]. Finally, feature warping is used to reduce the additive noise and channel effects and to map a feature stream to a standard normal distribution. A lookup table is devised so as to map a rank order determined from the sorted cepstral feature elements to a warped feature using the desired warping target distribution [10].

Figure 3 shows the visualization of cepstral domain plots of the time sequence of speech feature. The time sequence of the first order of cepstral coefficient c^1 and the related delta coefficient are plotted (corresponding to the first row and second row). The x-axis is time sequence and the y-axis means the log magnitude. Within each box, the first column shows original MFCC feature. Next, the process of moving averages of temporal information was applied and shown in the second column. After ARMA process, spikes are obviously smoothed in the time sequence thereby reducing noisy. The third column shows the contour of the mean of long-term feature analysis. The smoothing and compact feature contour can be found after LTF processing.

The visualization of spectral plots with CMS and CVN processing was shown in the fourth column. We can find the y-axis shown that the value of cepstral coefficients is re-estimated into the scale of zero mean and unit variance after CMS and CVN. Finally, the fifth column denotes the log magnitude with feature warping. In the y-axis of c^1 and its delta coefficient, the value of each dimension of parameters is mapped into the same maximum and minimum values after feature warping.

3. Experiments

The NIST SRE data are collected from different channel types such as cellular, cordless and land-line. There are also several other concerning factors, such as language, various type of recording (telephone transmission type and microphone type). Most of the data was recorded over telephone lines. All verification data are sampled at 8 kHz frequency.

We evaluated the system on the conversational telephone English speech of NIST Speaker Recognition Evaluation (SRE) 2008. The NIST SRE-2004 one-side data was used to train the gender-independent universal background models. All results were obtained using the version 3 of the NIST SRE-2008 answer key in this study.

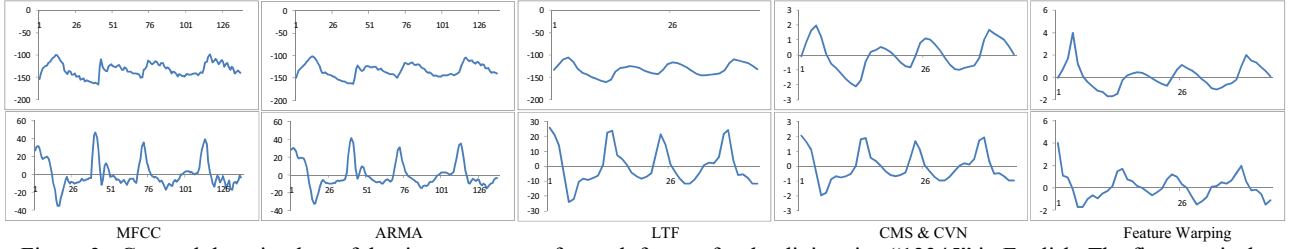


Figure 3: Cepstral domain plots of the time sequence of speech feature for the digit string “12345” in English. The first row is the first order of cepstral coefficient c^1 and the second row is the delta coefficient of c^1 .

3.1. Speaker recognition system

The accuracy of speech activity detection is important for reliable and robust speaker recognition. This study applied a hybrid endpoint detector [11]. The strategy was to find endpoints using a three-pass approach in which energy pulses were located and edited, and the endpoint pairs were scored in the order of most likely candidates.

GMM supervectors were used to construct kernels of support vector machines (SVM) in this study [6]. The iterative EM algorithm was adopted to estimate the parameters of Gaussian components. The gender-independent UBM was trained with 512 mixture numbers. Given a speaker’s speech data, a GMM was estimated by using MAP adaptation of the means of the UBM. The means of mixture components in the GMM were concatenated to a GMM supervector, which was used as SVM kernels. We used the NIST SRE-2004 data as the background training data. It was used for the training of UBM as well as composing the set of background speakers in SVM training [12]. At the same time, the NIST SRE-2004 data were used to derive the Nuisance Attribute Projection (NAP) matrix. For all the SVM-based classifiers, NAP was performed to project out the nuisance subspace from the original supervector space. NAP and SVM were performed to compensate for the nuisance effects.

In evaluation, the various score normalizations were applied for score calibration [1]. With Tnorm, the input test speech utterance is evaluated against cohort models to obtain the normalization scores using mean and standard deviation [13]. With Znorm, a speaker model is tested against imposter speech utterances to obtain the mean and standard deviation scores of normalization [14]. For run-time efficiency, Znorm can be estimated in an offline mode. The outputs of our SVM classifier were normalized with ZTnorm, which further compensate for the nuisance effects. In this study, the NIST SRE-2005 one side training data was used for Tnorm and NIST SRE-2004 data for Znorm, respectively.

3.2. Performance evaluation

Two types of errors, false acceptance and false rejection, can occur in speaker verification. Equal error rate (EER) reports the system performance when the false acceptance and false rejection rates were equal. The normalized detection cost function (DCF) is a weighed sum of miss detection and false alarm rates as defined in NIST SRE 2008 [2], and shown as follows

$$DCF = C_{Miss} \times P_{Miss|Target} \times P_{Target} + C_{FalseAlarm} \times P_{FalseAlarm|NonTarget} \times (1 - P_{Target}) \quad (8)$$

where $C_{Miss} = 10$, $P_{Target} = 1$ and $C_{FalseAlarm} = 0.01$. In this

study, results of speaker verification are both reported in terms of EER and $100 \times DCF$.

3.3. Effect of the ARMA process

Each frame of the speech data was represented by a 48-dimensional feature vector, consisting of 16 MFCC c^1, \dots, c^{16} , along with their deltas, and double-deltas as the raw features in the speaker recognition. The short-time frequency analysis of 16 ms with the long-time window of 4 short-time cepstral frames was applied for the feature extraction.

We report the results of the NIST SRE-2008 core task. Here we are interested in the choice of ARMA order α as shown in Table 1.

Table 1: Results with different order of ARMA process on NIST SRE-2008 (all English trials).

ARMA α	Male		Female		All	
	EER	100xDCF	EER	100xDCF	EER	100xDCF
0	2.84	1.24	3.37	1.49	3.17	1.43
1	3.14	1.13	2.84	1.43	2.96	1.33
2	3.14	1.22	3.63	1.43	3.47	1.36
3	3.14	1.32	3.54	1.45	4.16	1.79

Note that $\alpha = 0$ means no ARMA is involved. The best result was achieved when $\alpha = 1$, with ARMA contributing 6.62% EER reduction from 3.17% to 2.96%. A similar reduction was also observed with $100 \times DCF$. The experiment confirms that, with appropriate setting, ARMA process consistently improves the performance by retaining temporal information carried by the feature frames. $\alpha = 1$ was used in the following experiments.

3.4. Effect of the long-term feature analysis

The number of FFT sample points is usually a power of 2. This study applies the short-time spectral analysis of 16 ms to obtain MFCC features (128 samples at 8k Hz sampling rate and 64-sample shift) denoted as MFCC128. We compared the conventional MFCC feature with the different frame size of 128, 256 and 512 (denoted as MFCC128, MFCC256 and MFCC512, respectively). This group of settings served as the baseline to compare with the group of long-term features.

We applied long-term feature analysis (see Figure 1) to MFCC128 with various long-term windows of $L = 4, 6, 8$ frames, namely LTF4, LTF6 and LTF8, respectively. In other words, LTF4 represents the mean of every 4 frames of MFCC128. The feature of MFCC128 can be viewed as LTF0. We evaluated the variety of the window length L to find a best setting in core test as shown in Table 2, with the frame size of 16 ms, 32 ms and 64 ms for MFCC128, MFCC256, and 48 ms, 56 ms and 72 ms for LTF4, LTF6 and LTF8, respectively. Furthermore, we show in Fig. 4 a comparison of the amount of

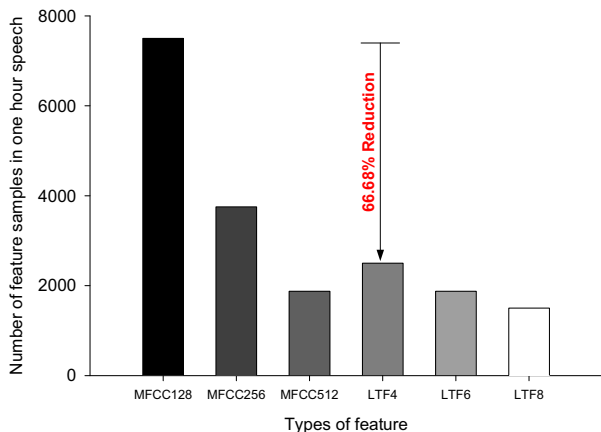


Figure 4: Number of feature samples for various types of features in one minute speech.

Table 2: Comparison of results of MFCC and LTF on NIST SRE-2008 (all English trials).

Feature	Male		Female		All	
	EER	100xDCF	EER	100xDCF	EER	100xDCF
MFCC128	3.11	1.18	3.45	1.43	3.34	1.38
MFCC256	3.29	1.38	3.97	1.73	3.72	1.62
MFCC512	4.19	1.89	5.23	2.19	4.87	2.09
LTF4	3.14	1.13	2.84	1.43	2.96	1.33
LTF6	3.29	1.35	3.52	1.57	3.43	1.51
LTF8	3.44	1.27	4.15	1.69	3.89	1.55

feature samples for one minute of speech data after VAD process.

We observe an average EER reduction of 11.38% from EER=3.34% of MFCC128 to EER=2.96% of LTF4. It was worth noting that the long-term feature also leads to feature sample reduction, thus improves computational efficiency given the same amount of input data. The number of feature sample in LTF4 was 2499. In Fig. 4, LTF4 leads to more than 66.68% feature sample reduction over the conventional MFCC128 (The number of feature sample was 7499).

The number of feature samples was greatly reduced with the larger window length L , such as LTF6 and LTF8. Note that the proposed long-term feature achieved a competitive performance at a reduced computational cost. Based on the similar frame size, we can find the feature type of LTF6 and LTF8 perform better results than MFCC256 and MFCC512. The above evaluations were plotted with the common Detection Error Tradeoff (DET) curves in Fig. 5.

4. Conclusions

We study the feature extraction techniques using long-term and temporal information for effective speaker recognition. Experiments were conducted on NIST SRE-2008. From the experiments, we find that the auto-regression moving average filter is useful in characterizing speakers. We observe 6.62% EER reduction with ARMA process. Compared to the conventional MFCC feature, LTF achieves 66.68% feature sample reduction and 11.38% EER reduction. We confirm that long-term and temporal information are helpful in speaker recognition.

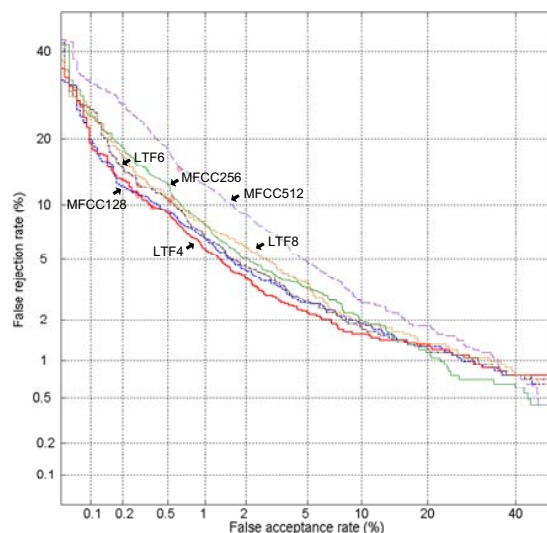


Figure 5: Comparison of MFCC and LTF features.

5. References

- [1] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovsk-Delactaz, and D. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP J. Appl. Signal Processing*, vol. 4, pp. 430–451, 2004.
- [2] NIST SRE (Online) <http://www.nist.gov/speech/tests/spk/>
- [3] M. J. F. Gales "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions on Speech Audio Processing*, vol. 7, pp. 272–281, 1999.
- [4] D. Reynolds, W. Campbell, J. Campbell, B. Dunn, T. Gleason, D. Jones, T. Quatieri, C. Quillen, D. Sturim, P. Torres-Carrasquillo, "Beyond cepstra: exploiting high-level information in speaker recognition," Workshop on Multimodal User Authentication, 2003.
- [5] C. P. Chen and J. Bilmes, "MVA Processing of Speech Features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 257–270, 2007.
- [6] W. M. Campbell, D. E. Sturim, D. A. Reynolds and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. ICASSP*, pp. 97–100, 2006.
- [7] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.
- [8] A. Torre, A. M. Peinado, J. C. Segura, J. L. Perez-Cordoba, M. C. Bentez and A. J. Rubio, "Histogram Equalization of Speech Representation for Robust Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 355–366, 2005.
- [9] O. Viikki, and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, pp. 133–147, 1998.
- [10] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. 2001: A Speaker Odyssey*, pp. 213–218, 2001.
- [11] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, J. G. Wilpon, "An improved endpoint detector for isolated word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 4, pp. 777–785, 1981.
- [12] R. Collobert and S. Bengio, "SVM-Torch: support vector machines for large-scale regression problems," *Journal of Machine Learning Research*, vol. 1, pp. 143–160, 2001.
- [13] R. Auckenthaler, M. Carey and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.
- [14] K. P. Li and J. E. Porter, "Normalizations and selection of speech segments for speaker recognition scoring," in *Proc. ICASSP*, vol. 1, pp. 595–598, 1988.