TEXT-DEPENDENT SPEAKER RECOGNITION WITH LONG-TERM FEATURES BASED ON FUNCTIONAL DATA ANALYSIS

Chenhao Zhang[1], Thomas Fang Zheng[1] and Ruxin Chen[2]*

[1] Center for Speech and Language Technologies, Division of Technical Innovation and Development, Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University, Beijing, China
[2]R&D, Sony Computer Entertainment America, Foster City, CA, USA

ABSTRACT

Text-Dependent Speaker Recognition (TDSR) is widely used nowadays. The short-term features like Mel-Frequency Cepstral Coefficient (MFCC) have been the dominant features used in traditional Dynamic Time Warping (DTW) based TDSR systems. The short-term features capture better local portion of the significant temporal dynamics but worse in overall sentence statistical characteristics. Functional Data Analysis (FDA) has been proven to show significant advantage in exploring the statistic information of data, so in this paper, a long-term feature extraction based on MFCC and FDA theory is proposed, where the extraction procedure consists of the following steps: Firstly, the FDA theory is applied after the MFCC feature extraction; Secondly, for the purpose of compressing the redundant data information, new feature based on the Functional Principle Component Analysis (FPCA) is generated; Thirdly, the distance between train features and test features is calculated for the use of the recognition procedure. Compared with the existing MFCC plus DTW method, experimental results show that the new features extracted with the proposed method plus the cosine similarity measure demonstrates better performance.

Index Terms— Text-dependent speaker recognition, functional data analysis, functional principle component analysis, distance metrics

1. INTRODUCTION

In recent years, there has been an increasing interest in speech research filed. As one of the most popular speech research technologies, text-dependent speaker recognition (TDSR) [1] has the feature of dependable performance. TDSR can be applied in many practical applications such as security check and economic business. Using voice to recognize the identities can be more comfortable and convenient for users than other biometrics. Therefore TDSR is worthy of further research.

In traditional TDSR, speech feature dynamics are exploited to identify different speakers. These methods compare the feature vector sequence extracted from the speaker's test utterance with the "feature-dynamics-models", in other words, the templates. For template matching, a speaker model consists of a sequence of vectors extracted from the training utterance. During the recognition phase, the distance between each test utterance' feature sequence and the speaker's template is calculated, and these distances are used to determine whether the test utterance and the training utterance are from the same speaker. Traditionally, Dynamic Time Warping (DTW) [2-3] or HMM [4] are used for classification.

From the above, it can be easily found that the speech feature vector plays a key role in TDSR because it can significantly influence the comparison between the test utterance and the training template [5]. It is important to note that around the vast majority of the prevalent speaker recognition systems, the speech spectral envelope parameters like Mel-Frequency Cepstral Coefficient (MFCC) [6] are dominant features. MFCC offers a short-term representation of the speech spectral envelops or the impact of the vocal tract shape in extracting an utterance. The traditional MFCC captures the highly local portion of the significant temporal dynamics, thus they cannot reflect some certain overall statistical characteristics hidden behind the sentences. If there is a way to combine the short-term features and the extra information obtained from looking over the whole data, the TDSR performance can be improved. Recently, some researchers have been focusing on a proposed computational method collectively known as the Functional Data Analysis (FDA) [7-9]. FDA has the mathematical framework that allows statistics on datasets whose elements are entire curve, and it consists the dimensional generalization in Hilbert's space. It has been proved that the FDA theory shows good performance on the speech feature analysis and pitch re-synthesis [10]. Considering the above fact, a long-term feature extraction process is proposed based on the FDA theory and MFCC features. The most fundamental point is that the FDA theory provides a way to formulate the problem of the TDSR to the statistical thinking and take the temporal advantage of MFCC feature. Firstly, the conventional MFCC feature is extracted from the speech data; Secondly, the FDA theory is

applied to the MFCC features; Thirdly, the function coefficients is compressed based on Functional Principal Component Analysis (FPCA), and the resulted compressed coefficients of the fitting functions are taken as the new feature instead of MFCC feature. The new features contain the overall utterance information so there is no need to perform the alignment as in the DTW algorithm, instead, a distance calculation between the train utterance features and test utterance features (or template) is necessary to be performed for the recognition purpose. It is assumed that different distance measure may lead to different performance and therefore comparison is needed to find a suitable distance measure.

This paper is organized as follows. In Section 2 the framework of the proposed method will be detailed, especially the FDA feature extraction procedure. In Section 3, different distance measures will be compared experimentally. In Section 4, the experiments as well as results and discussions will be described. Conclusions will be presented in Section 5.

2. MFCC-FDA FEATURE EXTRACTION FRAMEWORK

In this Section, The MFCC-FDA coefficient feature extraction framework is described as below:

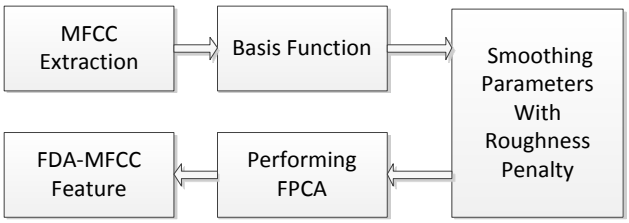

Fig. 1. FDA coefficient feature extraction framework

FDA represents the discrete points as the smooth curves or the continuous function $x_i(t),\ i = 1,\ \cdots,\ N$, so the basic model of FDA is built as below:

$$y_{ij} = x'_i(t_{ij}) = x_i(t_{ij}) + \varepsilon_i(t_{ij}), j = 1, \cdots, T_i \quad (1)$$

Where $y_i = (y_{i1}, \cdots, y_{iT_i})'$ is the discrete data, in this paper, it is MFCC feature series. $x_i(t_{ij})$ are fitting functions. $\varepsilon_i(t_{ij})$ are residuals that the fitting process causes. t represents the time.

2.1. The Basis Functions

A set of building blocks $\phi_k, k = 1, \cdots, K$ are called as basis functions, which are combined linearly. The fitting function $x_i(t)$ which will be represent the data is defined in mathematical notation as

$$x_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t) = \boldsymbol{c_i}'\boldsymbol{\phi(t)} \quad (2)$$

And this notation is called as the basis function expansion, $c_{i1},\ c_{i2},\ \cdots,\ c_{iK}$ are coefficients of the expansion.

In this paper, the Fourier basis functions will be used to simulate the MFCC feature series, as $\phi_0(t) = 1,\ \phi_{2r-1}(t) = \sin r\omega t,\ \phi_{2r}(t) = \cos r\omega t$. These basis functions will be uniquely determined though defining the number of the basis function K and the period ω.

2.2. The Calculation of Coefficients c

After the basis functions are decided, the $x_i(t)$ will be defined by the parameters $c_{i1},\ c_{i2},\ \cdots, c_{iK}$. As being common used, the sum of squared errors or residual is often to be represent the data fitting level:

$$SSE(y_i|c) = \sum_{j=1}^{n}\left[y_{ij} - \sum_{k=1}^{K} c_{ik}\phi_k(t_j)\right]^2 \quad (3)$$

The classic least square method can be used to solve this minimization problem.

2.2.1. Rough Penalty Method

When the number of the basis function K is decided not so appropriate, as too big or too small, it will cause to the overfitting problem or the underfitting problem when the least square method is used, so a roughness penalty method is introduced to improve the functional fitting problem. The roughness penalty method tries to solve the fitting job through two aspects:

- Make sure the closeness of the fit is good enough.
- Make sure the overfitting will not exist, that is, no dramatic changes in a local range.

The first aspect can be settled well by minimize the squared errors, then for the second aspect, the integration of square of the second derivate can be used to measure the curve change level, which is

$$PEN_2(x) = \int \{D^2x(s)\}^2 ds = \|D^2x\|^2 \quad (4)$$

These two goals are opposite, so the middle ground of *SSE* and PEN_2 should be taken, the criterion is built as:

$$PENSSE_\lambda = \sum_j \{y_i - x(t_j)\}^2 + \lambda * PEN_2(x) \quad (5)$$

Where λ is a smoothing parameter to control the level between *SSE* and PEN_2. When λ is small, the estimation will be toward to *SSE*; When λ is bigger, the estimation will get higher roughness penalty, the curve will be smoother.

2.2.2. Generalized Cross-Validation

The Generalized Cross-Validation measure (GCV) [11] is designed to locate a best value for parameters corresponding to define the basis function and the residual criterion, like the number of basis function K and the smooth parameter λ. The smaller the GCV value is, the better the parameters will be. The definition of the GCV values is:

$$GCV(\lambda) = \left(\frac{n}{n - df(\lambda)}\right)\left(\frac{SSE}{n - df(\lambda)}\right) \quad (6)$$

This GCV value will give the direction which value λ and the basis number K take better fitting level, the details can be found in [8].

2.3. Functional Principle Component Analysis (FPCA)

After the data pre-processing phase as the MFCC feature smoothing, the original data $x_{i1}, x_{i2}, \cdots, x_{ip}$ is converted into function form $x_i(t)$, so a continuous smoothing coefficient feature is obtained to represent the data information of different time from all the speaker. Compared with traditional discrete MFCC feature points, the functional data contains the overall time dynamic information. Only parts of these coefficients provide the dominant characteristic of the speech. So FDA theory provides a way to compress the functional data information - FPCA. The traditional PCA uses the linear combination as below:

$$f_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = \sum_{j=1}^{p} \beta_j x_{ij} \tag{7}$$

f_i is the i-th principal component for the data. Each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to (uncorrelated with) the preceding components. The meaning of FPCA is very similar with the traditional PCA. In FPCA, the continuous function $x_i(t), t \in [0, T]$ into one variable by the formula below:

$$f_i' = \int_o^T \beta(s)\, x_i(s) ds = \int \boldsymbol{\beta x_i} \tag{8}$$

The function $\beta(s)$ is corresponding to the linear weighting coefficients $(\beta_1, \beta_2, \cdots \beta_p)$, and f_i' is the i-th functional principal component for the functional data $x_i(t)$, so this problem can be abstracted as the formulate as:

$$\begin{cases} \max\ Var(f') = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\left(\int \boldsymbol{\beta x_i}\right)^2 \\ s.t \displaystyle\int [\beta(s)]^2 ds = \|\boldsymbol{\beta}\|^2 = 1 \end{cases} \tag{9}$$

FPCA scores contain the compressed data information from the original functional data $x_i(t)$, so it is the new MFCC-FDA feature, for each speech sentence, there will be a MFCC-FDA feature.

3. THE DISTANCE MEASURE

Because the new MFCC-FDA feature contains the overall dynamic information of the MFCC feature series, the MFCC of each utterance has been converted to a single fixed dimension vector, there is no need to train the template models like DTW or HMM algorithm, we can directly calculate the distance between the features from train utterances and test utterances for the purpose of the classification. Different distance measures have different space description properties. Finding a suitable distance measure or a good similarity measurement [12] between the new feature vectors will greatly influence the performance of the FDA-MFCC feature based TDSR.

In the Section 4, the experiments part, the common distance measures, like Minkowski Distance and the cosine similarity, will be used in the classification phase. Mainly, this paper focuses on the regular amplitude distance measurements and the angle between the feature vectors.

$$\begin{cases} d_{Mink} = \sqrt[P]{\displaystyle\sum_{k=1}^{n} |x_{1k} - x_{2k}|^P} \\ S_{cos} = \dfrac{\sum_{k=1}^{n} x_{1k} x_{2k}}{\sqrt{\sum_{k=1}^{n} {x_{1k}}^2}\sqrt{\sum_{k=1}^{n} {x_{2k}}^2}} \end{cases} \tag{10}$$

4. EXPERIMENTS

4.1. Database and Conditions

The noisy data "for-THU-psi3m-2", which was recorded by Sony Computer Entertainment America with PlaystationEye devices at three-meter distance in an office fan noise environment. It contains 5 different speakers. Each speaker uttered about 240 different short words and each word 3 times. The length of word is about 2 seconds in average, and all the words are sampled at 16 kHz sampling rate with 16-bit width.

For each word, one recording was selected for training and the other two recordings were used for verifying. Acceptance occurs only when the same speaker utters the same word.

The 16-dimensional MFCC features were extracted from the utterances with 30 triangular Mel filters used in the MFCC calculation. For each frame, the MFCC coefficients and their first derivative formed a 32-dimentional feature vector.

The Fourier basis functions are chosen to do the smoothing work. For the parameters selection, as being introduced in Section 2, the GCV value figure is showed in Fig. 2.

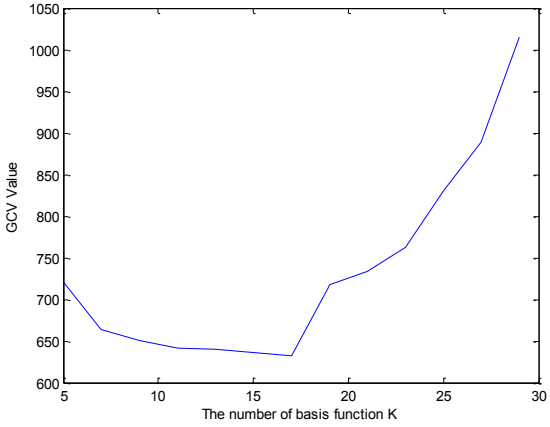

Fig. 2. GCV criterion for choosing the smoothing parameters

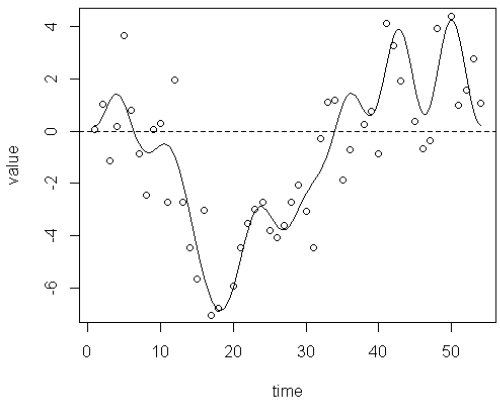

Fig. 3. One of the simulation results from one dimension of MFCC feature series

According to the GCV values, the functional parameters are chosen as: the number of basis functions K is 17, the smoothing parameter λ is 10e-4. By defining the functional parameters, the roughness penalty criterion can be used to calculate the coefficient set $c_{i1}, c_{i2}, \cdots, c_{iK}$, Fig. 3 is one of the simulation results from one dimension of MFCC feature series.

4.2. Results and Analysis

The baseline system is Dynamic Time Warping method with MFCC feature, this method is proved as a classic method for text-dependent speaker recognition. Table 1 shows the performances of the baseline system and the experiment that just use the FDA coefficients as the feature without FPCA compression, the distance measure is traditional Euclidean Distance, the Equal Error Rate (EER) [13] was used to evaluate the system performance.

Method	EER
MFCC-DTW (Baseline)	6.13%
MFCC-FDA	9.54%

Table 1. Results without FPCA compression

It can be found that compared with the classic MFCC-DTW system, the performance of directly using the FDA coefficients as the new feature is not good enough, that is because this coefficients contain parts of redundant information, the improvement of FPCA can be expected. Table 2 shows these results, *nharm* represents the number of harmonics or principal components to be computed.

MFCC-FPCA (*nharm*)	1	3	5	7	9
EER (%)	11.80	7.95	6.31	6.15	6.01

Table 2. Results with FPCA compression

FPCA shows great improvement over FDA coefficients, it effectively reduces the redundant information, and the MFCC-FPCA plus Euclidean Distance can achieve an equivalent performance as the classic MFCC-DTW TDSR system.

At last, the influence of the similarity measurement experiments was processed. Table 3 shows the results with *nharm* = 5.

Similarity Measurement	Euclidean Distance	Manhattan Distance	Chebyshev Distance	Cosine Similarity
EER (%)	6.31	7.06	9.62	2.49

Table 3. The performance of FPCA with different similarity measurement

MFCC-FPCA plus cosine similarity gave the best performance, according to the MFCC-FPCA feature represents the coefficients of the functions, the angle between the coefficient vectors shows more significance than the difference between the amplitude of the vectors, so compared with the distance measure like Euclidean Distance, cosine similarity fits the MFCC-FPCA system much better.

5. CONCLUSIONS

In this paper we propose a text-dependent speaker recognition method based on the FDA theory and the MFCC feature. This method combines the advantage of short-term speech dynamics and the global statistical information. Although the FDA theory is not designed for speech analysis and recognition, but it substantively provides a better overall speaker characterization. The experimental results show that this feature with the cosine similarity measure can achieve a better result than the conventional MFCC-DTW method.

6. REFERENCES

[1] F. Bimbot, J. Bonastre and et al, "A Tutorial on Text-Independent Speaker Verification", Eurasip J. Appl. Speech Proc. 4 (2004)

[2] V. Ram, A. Das, and V. Kumar, "Text-dependent speaker-recognition using one-pass dynamic programming," Proc. ICASSP'06, (2006)

[3] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification", IEEE Transactions on Acoustics, Speech and Signal Processing, 29:254-272, 1981.

[4] T. Matsui and S. Furui, "Concatenated phoneme models for text-variable speaker recognition", ICASSP 1993, Minneapolis, Minnesota, vol. 2, pp: 391-394.

[5] D. A. Reynolds, "An Overview of Automatic Speaker Recognition Technology", ICASSP 2002, Orlando, Florida, pp.4072-4075, 2002.

[6] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. on ASSP. 1980, Vol. 28, pp: 357-366

[7] J. O. Ramsay and B. W. Silverman, "Functional Data Analysis" (2nd Ed.) New York: Springer, 2005.

[8] J. O. Ramsay and B. W. Silverman, “Applied Functional Data Analysis - Methods and Case Studies”, Springer, 2002.
[9] J. O. Ramsay, G. Hookers, and S. Graves, “Functional Data Analysis with R and MATLAB”, Springer, 2009.
[10] M. Gubian, “Functional Data Analysis for Speech Research”, InterSpeech 2011, Florence, Italy
[11] G. H. Golub, M. Heath and G. Wahba, “Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter”, Technometrics ,Vol. 21, No. 2, May, 1979
[12] R. Duda, P. Hart, and D. Stork, “Pattern Classification”, New York: Wiley, 2000.
[13] NIST Speaker Recognition Evaluation Plan, Online Available http://www.nist.gov/speech/tests/sre/