PHYSIOLOGICAL FEATURE EXTRACTION FOR TEXT INDEPENDENT SPEAKER IDENTIFICATION USING NON-UNIFORM SUBBAND PROCESSING

Xugang Lu, Jianwu Dang

School of Information Science
Japan Advanced Institute of Science and Technology
1-1, Asahidai, Nomi , Ishikawa 923-1211, Japan
{xugang, jdang}@jaist.ac.jp

ABSTRACT

The features used for speech recognition should emphasize linguistic information while suppressing speaker differences. For speaker recognition, features should have more speaker individual information while attenuating the linguistic information. In most studies, however, the identical acoustic features are used for the different missions of speaker and speech recognitions. In this paper, we propose a new physiological feature extraction method which emphasizes individual information for speaker identification. For the purpose, physiological features of speakers were analyzed from the point of view of speech production. It is found that the speaker individual information is encoded in different frequency regions of speech sound. The speaker discriminative information was quantified using Fisher's F-Ratio in each frequency region. Based on the F-Ratio, we proposed a non-uniform sub-band processing strategy to extract new feature which can emphasize or refine the physiological aspects involves in speech production. We combined the new feature with GMM for speaker identification task and applied on NTT-VR speaker recognition database. Compared with MFCC feature, by using the proposed feature, the identification error rate was reduced 20.1%.

Index Terms **—** Speaker identification, physiological feature, non-uniform subband.

1. INTRODUCTION

Linear Prediction Coefficient (LPC) and Mel Frequency Cepstral Coefficient (MFCC) features are widely used as acoustic features for speech recognition. The state of the art of text independent speaker identification algorithm is also based on modeling the LPC or MFCC feature using Gaussian Mixture Model (GMM) [1]. However, the purpose of speech recognition is quite different from that of speaker recognition, the former task needs to emphasize linguistic information and suppress speaker individual information, while the later task needs more speaker individual information. This contradiction suggests that LPC and MFCC may not meet both speech and speaker recognition tasks.

For speaker recognition, the problem is how to extract and utilize the information that characterizes an individual speaker. Individual speaker information results mainly from two factors: physiological and social factors. The former is related to the speaker's gender, age, and oral morphology which are inborn characteristics; the latter concerns the speaker's dialect, idiolect, occupation, and so on which results from his/her social environment. In this paper, we focus on the former factor, and investigate the individual information caused from speech production point of view.

When producing a speech sound, speakers' physiological and morphological features are contained in acoustic characteristics of the sound. The diverse articulators' physical properties are represented in acoustic spectra [2]. In order to extract that information, some speech feature representations were developed. The LPC feature can well model the vocal tract property by using an all-pole model which reflects the main vocal tract resonance property in acoustic spectra [2][3]. While MFCC feature takes the auditory nonlinear frequency resolution mechanism into consideration which makes the representation more robust [3]. For extracting more direct physiological features, the fundamental frequency or pitch which reflects the vocal cord information of speakers is often used [4]. The LPC residual signal for describing the speakers' glottal information [5] is also proposed. When these features are used for speaker recognition, the performance is improved as some researchers demonstrated [4][5]. In essence, most of the representations for speaker recognition want to catch the main vocal tract physical property which is usually described as acoustic resonance property. Actually, besides the main vocal tract, there are some side branches, such as the nose, piriform fossa, etc., which introduce specific features into speech [6][7]. These features are distributed in some specific frequency regions. For example, as revealed in [6], the side branches always produce anti-resonance which is reflected as zero-pole pairs on spectral profile. A relation between the physiological features and acoustic ones is shown in Fig. 1. It shows the transfer functions for vowel /a/. The top curve is the spectrum of transfer function of the oral cavity without side branches, which was calculated using a transmission line model. The bottom curve is the long-term average spectrum of real sound. The other curves were calculated for the oral cavity with nasal

coupling and/or the piriform fossa. As shown in Fig. 1, when both the nasal coupling and piriform fossa are taken into account, the simulation (the second curve from the bottom) has a nice matching with the real sound. The situation of pole-zero pairs depends on the coupling degree of the nasal and oral cavities. Because different speakers usually have diverse coupling degrees of the cavities according to their habits, the fine structures of those frequency regions reflect the different physiological information for the speakers. Those cavities produces acoustic feature in some frequency regions, such as the side branch of piriform fossa produces anti-resonance between 4kHz and 5kHz. Also, those cavities are less changed during speech production. If we emphasize the acoustic feature around those frequency regions, the feature should be more suitable for speaker individual characteristics description.

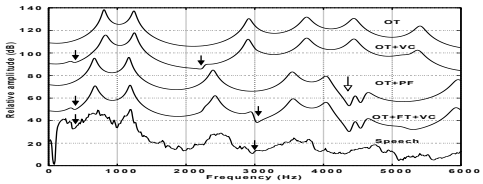

Fig.1 Transfer functions of the vocal tract and a long-term average spectrum for vowel /a/. Pole-zero pairs in regions of 300Hz, 3000Hz and 4500Hz were caused by the nasal coupling and piriform fossa. (After [6])

From the analysis above, we find that the speaker individual information is not distributed uniformly in each frequency band. The Mel frequency analysis for speaker individual information extraction is not suitable for speaker recognition task. In this paper, we investigate the new feature representations which reflect the importance of the speaker specific information in different frequency regions, and design a subband processing strategy for feature extraction and apply it for speaker identification task. The paper is organized as follows. In the second section, the physiological feature extraction method is given. In the third section, the GMM acoustic model for modeling each speaker using HTK is introduced. In the fourth section, speaker identification experiments are done to test the proposed feature. Lastly, some discussions and conclusions are given.

2. NON-UNIFORM SUBBAND PROCESSING FOR PHYSIOLOGICAL FEATURE EXTRACTION

As discussed in Section 1, speaker individual characteristics are not uniformly encoded in each frequency band. Such as glottal information is encoded in low frequency regions (between 50Hz and 500Hz), the piriform fossa information is encoded in high frequency regions (between 4000Hz and 5000Hz), etc. For speaker recognition, we need to investigate which frequency band possesses more speaker individual information, i.e., the importance of each frequency band for speaker recognition. For investigating the importance of each frequency band for speaker recognition, we use linear frequency scale triangle filters to process speech power spectrum. The triangle filters are shown in Fig.2. Each filter band gives an output which integrates the frequency energy around the center frequency of the filter band. We adopt the Fisher's F-Ratio of each frequency band to measure the speaker discriminative ability of each frequency band which is used as an index of importance of speaker individual information [9]. The F-Ratio is defined as (1):

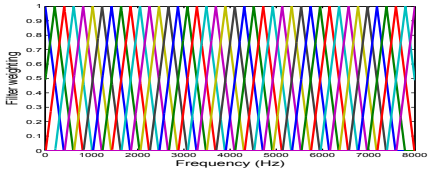

Fig.2 Uniform Sub-band filters with uniform bandwidth

$$F-Ratio=\frac{\sum_{i=1}^{M}(u_i-u)^2}{\frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(x_i^j-u_i\right)^2} \quad (1)$$

where x_i^j is the $j-th$ sample of speaker i with $j=1,2,...,N$, and $i=1,2,...,M$. u_i and u are the averages for speaker i and for all speakers respectively, which are defined as:

$$u_i=\frac{1}{N}\sum_{j=1}^{N}x_i^j \quad ; \quad u=\frac{1}{M}\sum_{i=1}^{M}u_i \quad (2)$$

Formula (1) is the ratio between the inter-speaker variance and intra-speaker variance of the feature in one frequency band. This measurement is often used to measure the discriminative ability of a feature for pattern recognition. We adopt it for measuring the speaker discriminative score in each of frequency bands.

We use the NTT-VR speaker recognition database in which there were 35 speakers in total including 22 male speakers and 13 female speakers [8]. The speech was collected in 5 sessions over a period of 10 months. In each session, each speaker was asked to speak the sentences with normal, slow and fast speed. The average duration of each sentence is 4s. The speaker discriminative ability is calculated using formula (1) for each frequency region, and shown in Fig.3. In Fig.3, the session recorded in August, 1990 is denoted as 90.8. Other sessions are denoted analogously. From Fig.3, one can see that, the discriminative information is mainly concentrated in three regions in frequency domain. The lowest region from 50Hz to 300Hz is concerned with the glottal information, the fundamental frequency. The dominant region is located in the range from 4 kHz to 5.5 kHz, which is concerned with the piriform fossa [7]. The region from 6.5 kHz to 7.8 kHz seems to be related to the consonants, probably the location of their constrictions. It is interesting to see that the distribution of speaker discriminative information is

invariant over the time span. In contrast, there is less speaker discriminative information in the middle frequency region from 0.5 kHz to 3.5 kHz. This because that the phonetic discriminative information is concentrated in this region which is consistent among the speakers for phoneme recognition. This statistical result confirms our speculation in Section 1, i.e., speaker individual information is not uniformly distributed in each frequency band. We use this result to design sub-band processing strategy for feature extraction for speaker identification.

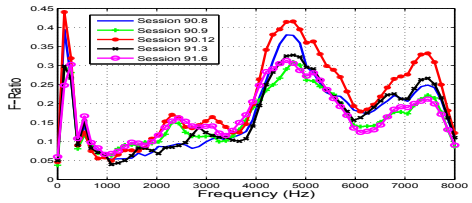

Fig.3 Speaker discriminative score in frequency domain using F-Ratio

In order to enhance the contribution of those frequency bands with more speaker individual information in spectral profile, we conduct the following procedure on sub-band filters' design. We improve frequency resolution in those frequency regions with high F-Ratio values. In the design of the sub-band filters, the bandwidth of each sub-band is inverse proportional to the F-Ratio of each frequency band. By this processing, the resolution of the spectral structure around frequency regions with high F-Ratio is improved. The designed filters are shown in Fig.4.

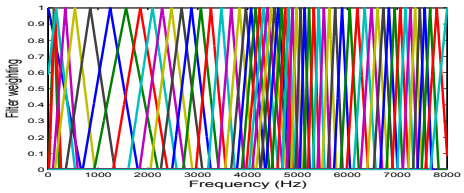

Fig.4 Non-uniform sub-band filters with non-uniform bandwidth

Comparing the filters in Figs. 2 and 4, one can see that the center frequencies of filter bands are distributed differently along frequency axis. The center frequencies are uniformly sampled along frequency axis in Fig.2, while are non-uniformly sampled in Fig.4. In order to compare the proposed non-uniform with the Mel frequency description, we plot the resolution of each description in frequency domain in Fig.5, where the uniform method is also plotted for a reference. From Fig.5, one can see that Mel frequency sampling has high frequency resolution in low frequency regions, while the non-uniform frequency sampling has high frequency resolution in the frequency regions with high F-Ratio values. Since the spectral envelope extracted with the non-uniform filters designed based on F-Ratio emphasizes the speaker specific information, it is possible that the feature extracted using these non-uniform filters improves the speaker identification performance. In the following sections, we test the feature set which is extracted using the non-uniform filter bands by doing speaker identification experiments.

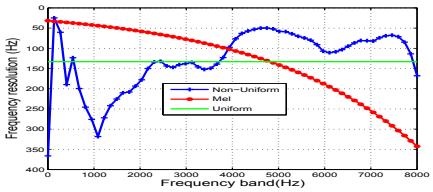

Fig.5 Comparison of frequency resolutions of filter bands

3. SPEAKER MODELING FOR IDENTIFICATION BASED ON HTK

GMM is widely used for speaker modeling in context independent speaker identification [1]. In our research, we use HTK to design the speaker models [10]. Each speaker is modeled using a three-state HMM in which only one state is modeled with Gaussian mixture distributions, the other two states are dummy states. This HMM speaker model is almost the same as the GMM speaker model except that the former has state self-transition involved in the calculation of the likelihood probability. The identification rate is defined as the ratio between the number of correctly identified speech segments and total number of speech segments for each speaker. Our purpose in this study is to test whether the proposed feature has more speaker individual information, thus a three-state with two dummy states HMM is used for each speaker modeling to evaluate the proposed method.

4. SPEAKER IDENTIFICATION EXPERIMENTS

We conducted speaker identification experiments on NTT-VR database [8]. For training speaker models, 10 speech sentences uttered at normal speaking rate were used for each speaker from session 90.8 which was recorded in August, 1990. For identification, we used all other utterances in all sessions at different speaking rates. The processing diagram for speech feature extraction is shown in Fig.6. In the feature extraction processing, a voice activity detector (VAD) is used to delete the silences and pause periods within speech sentences. The signal is then pre-emphasized using an emphasizing coefficient of 0.97. Short-term fast Fourier transform (SFFT) is used for each frame in which a hamming window with 16ms frame length and 8ms shift was employed. 60 band pass filters are used to integrate each frequency band to get power spectrum. After applying the logarithm, the Discrete Cosine Transform (DCT) is adopted to get 32 order cepstral coefficients vectors (zero-th order cepstral coefficient was excluded). Finally, the proposed feature vectors are extracted for speaker modeling. In the experiments, three kinds of feature sets are extracted under the conditions in which the filter bands block in Fig.6 was Mel frequency scale filter bands, uniform linear frequency scale filter bands (in Fig.2), or the proposed non-uniform filter bands (in Fig.4). The features are denoted as

MFCC, LFCC, and NUFCC, respectively. These feature sets were modeled by both the diagonal covariance matrix and full covariance matrix. The performance of using full covariance matrix is better than that of using the diagonal covariance matrix. For full covariance setting, however, the Gaussian mixture number was limited when the training data set is not large enough. The Gaussian mixture number in our experiment was chosen as 4 for all identification experiments. The identification results are in Fig.7 for the three feature sets.

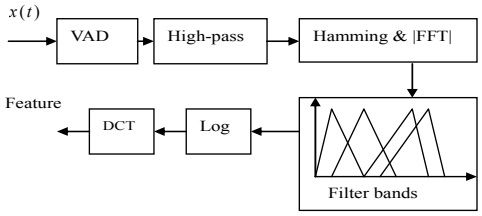

Fig.6. Speech feature extraction diagram

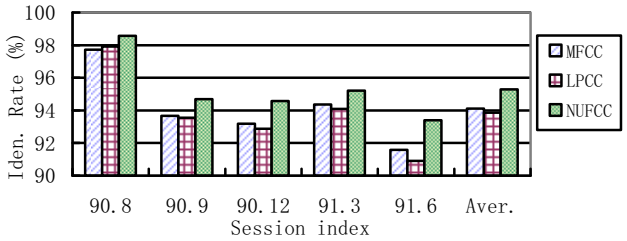

Fig. 7 Identification rate for the three feature sets

In Fig.7, the horizontal axis is the session index. From Fig.7, one can see that the identification rate is high in session 90.8 since the speaker models are trained using the data set from the same session although the testing data is different from the training data set. Comparing with the linear feature LPCC, the performance of feature MFCC is a little bit better. Among the three feature sets, the proposed non-uniform frequency feature (NUFCC) performs the best. Compared with the baseline feature set MFCC, the error reduction rate is about 20.1% on average for all testing sessions.

5. DISCUSSIONS AND CONCLUSIONS

In this study, we first analyzed speaker specific physiological information from the speech production point of view. The analysis showed that speaker individual information is partially concerned with physiological difference of speech organs. For quantitative analysis, we investigated the frequency band dependent distribution of speaker individual information using F-Ratio. The results showed that the piriform fossa causes one dominant discriminative speaker information in frequency region between 4kHz and 5.5 kHz, and the glottis and the constriction of the consonants would be the secondary factor in the lower and higher frequency domains, while almost no discriminative speaker information in the region of 0.5-3.5 kHz. According to the results, we designed non-uniform filter bands to extract speaker physiological-dependent features. Speaker identification experiments showed that feature extracted using the proposed non-uniform sub-band processing improved speaker identification performance. The error reduction rate was 20.1% compared with the baseline model with MFCC.

For further applying and improving the speaker identification performance based on the non-uniform distribution of speaker specific information in frequency domain, some problems need to be investigated further. One problem is how to quantify the speaker specific information in each frequency band more efficiently, and integrate this information using a statistical framework. Another problem is how to adopt different feature extraction methods for different speech categories. Because for different speech categories, such as for vowels or consonants, they excite different aspects of speaker individual information, we need to extract the speaker specific information by considering the different speech pattern categories. All these problems remain as our future work.

Acknowledgement: This study is supported in part by Grant-in-Aid for Scientific Research of Japan (No. 17300182) and grant of young scientist B of Japan (No.18700172).

6. References

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

[2] Kenneth N.Stevens, *Acoustic phonetics,* The MIT press, 1998.

[3] L. Rabiner, B.H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall PTR, 1993

[4] B.Atal, "Automatic recognition of speakers from their voices", *Proc. IEEE 64*, pp. 460-475, 1976.

[5] J. He, L. Liu, and G. Palm, "On the use of features from prediction residual signals in speaker identification", *Proc. EUROSPEECH95*, Vol. 1, pp. 313-316, Madrid, Spain, 1995.

[6] J. Dang, K. Honda, "An Improved vocal tract model of vowel production implementing piriform fossa resonance and transvelar nasal coupling", *Proc. ICSLP'96*, pp.965-968, Philadelphia, USA, 1996.

[7] J. Dang, and K. Honda. "Acoustic characteristics of the piriform fossa in models and humans," *J. Acoust. Soc. Am*. 101, 456-465, 1997.

[8] T.Matsui, S.Furui, "Comparison of text-independent speaker recognition methods using VQ distortion and discrete/continuous HMMS', *Proc. ICASSP 92,* Vol.II, pp. 157-160, 1992.

[9] J.J.Wolf, "Efficient acoustic parameters for speaker recognition", *J. Acoust. Soc. Am,* vol. 51, No.6, pp. 2044-2056, 1972.

[10] HTK tutorial book, http://htk.eng.cam.ac.uk/