An investigation of dependencies between frequency components and speaker characteristics based on phoneme mean F-ratio contribution

Songgun Hyon[*], Hongcui Wang[*], JianguoWei[*] and Jianwu Dang[*,†]
[*] School of Computer Science and Technology, Tianjin University,
92 Weijin Road, Nankai District, Tianjin 300072, China
E-mail: h_star1020@yahoo.com
[†]School of Information Science, Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
E-mail: jdang@jaist.ac.jp

***Abstract*—This paper proposes a new speaker feature extraction method, which is based on the non-uniformly distributed speaker information in frequency bands. In order to emphasize individual differences effectively, in this study, we first examine the differences of the distribution of individual information in the frequency region when a speaker utters different phonemes. Then we adopt an improved *F*-ratio, a phoneme mean *F*-ratio, to measure the dependences between frequency components and individual characteristics. According to the result of the analysis, we adopt an adaptive frequency filter to extract more discriminative feature. The new feature was combined with GMM speaker models and applied to the speaker recognition database which includes 50 persons. The experiment shows that the error rate using the proposed feature is reduced by 28.5% compared with the *F*-ratio feature, and reduced by 68.02% compared with the MFCC feature.**

I. INTRODUCTION

A. Nonlinear frequency scale transformations based on the contribution rate of speaker identification

Speaker recognition is to extract speaker's individual information from the speech, and recognize according to requirement [1]. For speaker recognition, the problem is how to extract and utilize the information that characterizes individual speakers.

Nonlinear frequency transforms such as Mel, Bark and ERB (Equivalent Rectangular Bandwidth) are usually applied to speech recognition and speaker recognition. However, these frequency scale transformations focus only on human auditory characteristics, not considering the semantics and individual information.

Most of the previous studies try to extract the features for speaker recognition from the contribution in different frequency domain. Hayakawa [2] used LPC features which are from different frequency bands. The experiment showed that speaker individual information existed in high frequency bands. Orman and Arsland [3] analyzed contribution of different frequency sub-bands for recognition performance and proposed a feature extraction method based on frequency domain filters. Miyajima and Watanaba [4] extracted features by using second-order all-pass function to normalize frequency spectrum, have achieved some results, but it can't reflect the property of the non-uniformly distributed in frequency bands. In order to efficiently reflect speaker's individuality in short-time spectrum, Yu [5] proposed a new non-linear frequency transform and feature extraction algorithm, which is based on analyzing contribution of short-time spectrum in different frequency sub-bands to speaker recognition and the technology of least square polynomial curve fitting. Miyajima [4] used a monotonic frequency warping function rather than using Mel frequency warping function to process speech spectrum.

Lu and Dang [6] adopted the Fisher's F-ratio method to analyze the dependencies between frequency components and individual characteristics, which further confirmed the non-uniformly distributed in frequency bands of the speaker information, but the linguistic information hasn't been totally removed in *F*-ratio feature.

B. The phoneme influence on the distribution of speaker information

The speaker information is not uniformly encoded in frequency bands. For example, the information of the glottis is mainly encoded in a low frequency band (between 100 Hz and 400 Hz), and the information of the piriform fossa in a high frequency band (between 4 kHz and 5 kHz), etc. In contrast, most speech discriminative information, such as the first three formants, is encoded in a low and middle frequency region from 200 Hz to 3 kHz, which is very important for speech recognition [7, 8]. Such kind of non-uniform distribution of speaker information in frequency bands was also confirmed in [2, 4]. The non-uniform distribution of speaker information in frequency has close relationship with vocal tract shape of the speaker.

However, these personality traits are obtained when human speak, it is difficult to avoid the effect of content. Most of the speaker information and linguistic information exist in the same frequency domain, and are difficult to isolate. For example, nasal cavity has the close relationship with special phonemes (nasal sound), like /n/, /m/, /ng/. In some

non-nasalized vowels such as /i/ and the voice bar of voiced stop consonants, a quite strong coupling takes place between the nasal and oral cavities via a transvelar coupling caused by the velum vibration [8, 9]. The formant can reflect the phoneme information, and also describe the individual inner oral cavity features such as vocal tract length and tongue size etc. Most speaker information as describe above has relation to vowel, however the distribution of speaker information on consonants is quite different. Generally, the speaker information has different distributions in different phonemes.

By considering that the distribution of speaker information is related to phoneme, it is necessary to find out a new frequency scale transform based on the distribution of the speaker information. For choosing a better frequency warping, we need to quantify the contribution of each frequency component to speaker individual information description.

II. CONTRIBUTION OF THE DIFFERENT FREQUENCY BANDS FOR SPEAKER RECOGNITION

A. Speaker information measurement based on the phoneme mean F-ratio score

In order to investigate the contribution of each frequency region for speaker recognition, we use triangle-shaped band-pass filters with linear frequency scale to process the speech power spectrum. Each filter band gives an integrated energy around the center frequency of the filter band. By using such filter bands, we get the band-pass energy spectrum with equal frequency resolution along the frequency axis. Based on the output of each frequency band, we can measure the dependences between frequency band outputs and speaker identities.

In order to get the speaker discriminative contribution in each frequency sub-band for speaker identification, we firstly considered the *F*-ratio for the same phoneme (Phoneme F-ratio, PF). Suppose there are M speakers, P phoneme. The speaker identification *F-ratio* of the l^{th} frequency sub-band for k^{th} phonemes is defined as the ratio of the variance of different speaker's mean and the mean of same speaker's variance under the conditions of k^{th} phoneme utterance.

$$PF_k^l = \frac{\frac{1}{M}\sum_{i}^{M}(u_k^i - u_k)^2}{\frac{1}{\sum_{i=1}^{M} N_{ik}}\sum_{i}^{M}\sum_{j}^{N_{ik}}\left[x_k^i(j) - u_k^i\right]^2} \quad (1)$$

Where N_{ik} is the frame number of the k^{th} phoneme in speaker i, $x_k^i(j)$ is the data of j^{th} speech frame of the k^{th} phoneme in speaker i, with $j=1,2,\ldots, N_{ik}$; u_k^i is the mean of k^{th} phoneme data in speaker i, u_k is the mean of k^{th} phoneme data for all speakers, as formulated in the following.

$$u_k^i = \frac{1}{N_{ik}}\sum_{j=1}^{N_{ik}} x_k^i(j); \quad u_k = \frac{1}{\sum_{i=1}^{M} N_{ik}}\sum_{i=1}^{M}\sum_{j=1}^{N_{ik}} x_k^i(j) \quad (2)$$

According to the phoneme *F*-ratio, the contribution rate of speaker recognition in l^{th} sub-band when product the k^{th} phoneme can be got as Eq. (3):

$$PFC_k^l = \frac{PF_k^l}{\sum_{i=1}^{L} PF_k^i} \quad (3)$$

According to the *F-ratio* contribution of each phoneme (PFC), the general *F-ratio* score of speaker recognition for the l^{th} sub-band is defined as the weighted average of the *F-ratio* contribution corresponding to all phonemes:

$$PMF^l = \sum_{k=1}^{P} \frac{N_k}{\sum_{i=1}^{P} N_i} \bullet PFC_k^l \quad (4)$$

Finally, based on the phoneme mean *F-ratio* score (PMF), the speaker recognition contribution rate of the l^{th} sub-band is defined as:

$$PMFC^l = \frac{PMF^l}{\sum_{i=1}^{L} PMF^i} \quad (5)$$

Where, L is the number of the sub-bands.

B. Phoneme Mean F-ratio Contribution (PMFC) measurement results for speaker identification

In order to find the effects of phoneme on speaker recognition, we need phoneme-labeled speech corpus which is different from ordinary speaker recognition. We conducted the experiments on "RyongNam2006" Korean speech recognition database in which phonemes are labeled. The data is recorded via the NT2 condenser microphone, with a sampling frequency of 22.05 kHz, 16 bit quantification and single-channel recording. In our experiments, the data set include 43 speakers (25 men, and 18 women). Each speaker is recorded more than 15 times. Each utterance is more than 10 seconds. The time interval of the same person's recording is 2~3 days.

Fig. 1 shows the *F*-radio contribution for the different phonemes. From Fig. 1, one can see that the distribution of speaker information changes with phonemes. For the vowel sound, most of the speaker information is distributed in the band below 300Hz and 4 ~ 5KHz. However, for some vowels, such as /æ/, there are obvious speaker characteristics in the high frequency band near about 10KHz. For the consonant sound, a lot of speaker information is distributed in the high frequency band above 6KHz. For nasal, in frequency band below 500Hz and around 2.5KHz.

According to the *F*-ratio contribution of each phoneme, the phoneme mean F-ratio contribution (PMFC) can be obtained from Eq. (4), Eq. (5). Fig. 2 shows the PMFC result for speaker information of all sub frequency bands and comparison with normalization value of the common F-ratio obtained from [6].

As shown in Fig.2, there are two apparent peaks below 5 kHz frequency. The first peak region with lowest frequency below 300 Hz is concerned with the glottal information, the fundamental frequency. The second peak region is located in the frequency range from 4 kHz to 5.5 kHz. As shown in [9, 10, 11, 12], the piriform fossa module in the speech production model causes spectral structure changes in frequency region from 4kHz to 5kHz. In addition, the

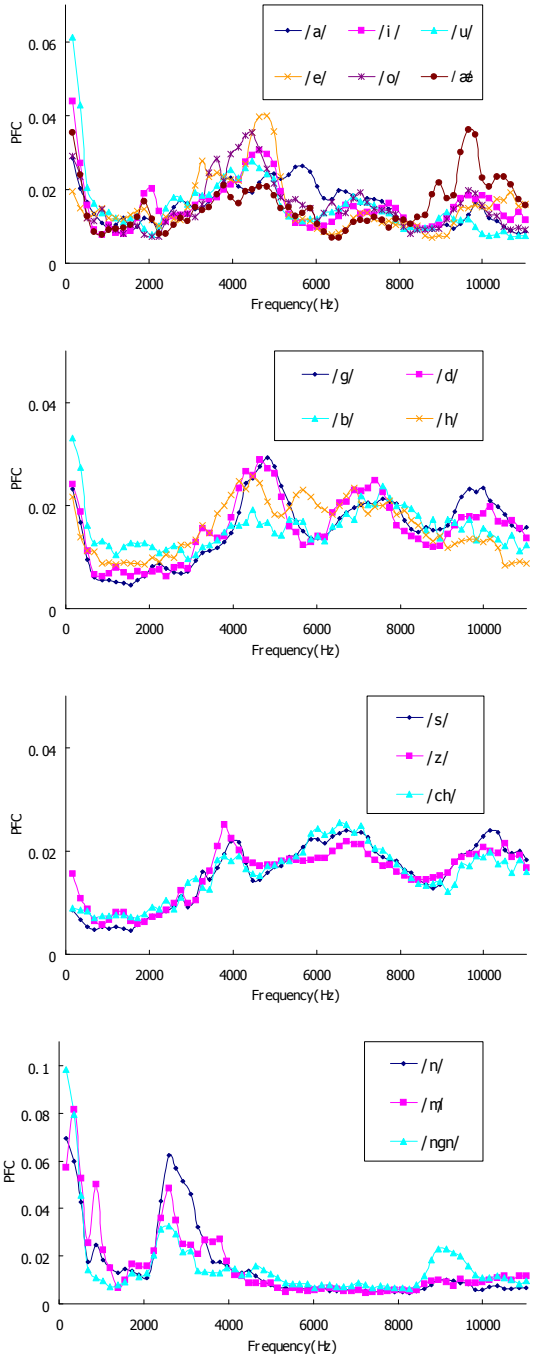

Fig. 1 Comparison of F-ratio contribution of the speaker information for different phonemes.

piriform fossa cavity is speaker dependent and less changed during speech production. Thus, the second peak region concerned with the piriform fossa is another one important speaker discriminative cue [6, 9, 10, 11, 12]. In contrast, there is less speaker discriminative information in the middle frequency region from 500 Hz to 3.5 kHz, especially near the 1 kHz frequency region. This is because most of the phonetic

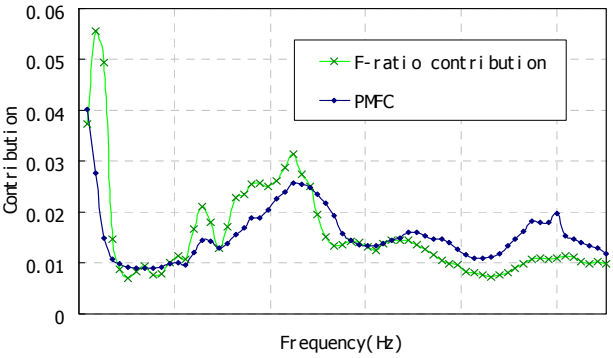

Fig. 2 Comparison of the contribution for speaker information with phoneme mean F-ratio and common F-ratio.

discriminative information is concentrated in this frequency region, which is consistent among all the speakers for phoneme description.

However, from 7 kHz to 11 kHz frequency domain, the result is a bit different. The contribution from the proposed method is obviously higher than Lu's. The experiment shows the contribution of consonant in high frequency above 6 kHz is much higher compared with the common F-ratio. Furthermore, some vowels with this high frequency domain have a certain degree of speaker identification contribution.

That is to say, the frequency region above 9 kHz also carries individual information, which may contribute to speaker recognition.

III. SPEAKER IDENTIFICATION EXPERIMENTS

A. Speaker feature extraction and modeling

According to the phoneme mean *F-ratio* contribution results, we can design the non-uniform sub-band filters by considering the dependency measurements in order to change frequency resolutions in different frequency regions. We extract the feature set using the non-uniform sub-band filters, and apply it to speaker identification experiments.

The processing diagram for speaker feature extraction is shown in Fig. 3.

In the process of the feature extraction, a voice activity detector (VAD) is used to delete the silence and pause periods within speech sentences. The signal is then pre-emphasized using an emphasizing coefficient of 0.98. Fast Fourier transform (FFT) is used for each frame in which a hamming window with 25 ms frame length and 10 ms shift is employed.

64 band-pass filters are used to integrate each frequency band to get power spectrum. According to the $PMFC^l$ results in Eq. (5), we can reset the sub-band filter based on the different contribution of every sub-band for speaker information. There are $L \bullet PMFC^l$ filters in every frequency sub-band. When updating, the higher contribution of the sub-band, we can get the more closely distribution of triangular filter.

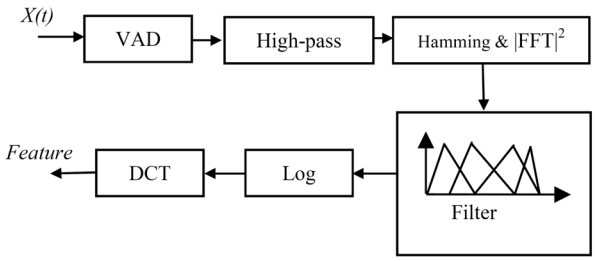

Fig. 3 Speaker feature extraction diagram

After applying the triangle filters and logarithm transform, the Discrete Cosine Transform (DCT) is adopted to get 32 order cepstral coefficient vectors (zeroth order cepstral coefficient was excluded). The feature vector is defined as:

$$c_{PMFFCC}(i)=\sqrt{\frac{2}{L}}\sum_{l=1}^{L}\log m(l)\cos\left\{\frac{\pi i}{L}(l-\frac{1}{2})\right\} \qquad (6)$$

Where, *m(l)* denotes the outputs of triangular filters.

Each speaker is modeled using a GMM. The parameters of the models are estimated during the training stage, based on which the likelihood probability is calculated for identification during the testing stage.

B. Speaker identification experiment

We conducted speaker identification experiments on "RyongNam2007" Korean speech database which includes 50 persons (25 men and 25 women). Sampling frequency is 22.05 kHz, 16 bit quantification, single-channel recording. For training speaker models, 2~3 sentences uttered at normal speaking rate were used for each speaker, every sentence costs about 10s. For testing speaker models, 30 sentences were used. Every sentence cost about 2~5s.

The feature sets were modeled by the diagonal covariance matrix. The highest identification rate was achieved when the Gaussian mixture number was 32.

Our test data is divided into three parts, speaker recognition experiment is done on each part. Each data set has about 10 * 50 sentences (50 speakers, and 10 sentences for each speaker).

The speaker identification results are shown in Table I for the three feature sets. In Table I, the proposed PMFFCC (Phoneme Mean *F-ratio* Frequency Cepstrum Coefficient) demonstrated the best performance. The error rate is reduced by 28.5% compared with the baseline model with FFCC (*F-ratio* Frequency Cepstrum Coefficient) and 68.02% compared with the MFCC (Mel Frequency Cepstrum Coefficient).

TABLE I
SPEAKER IDENTIFICATION RATES FOR THE THREE FEATURE SETS

Kind of parameter	MFCC	FFCC	PMFFCC
Database1	96.73%	98.11%	98.66%
Database2	95.07%	97.84%	98.39%
Database3	93.83%	97.42%	98.21%
total	95.06%	97.79%	98.42%

IV. CONCLUSION

In this study, to emphasize individual differences more effectively, we proposed a nonlinear frequency scale transformation based on phoneme mean F-ratio score. According to the phoneme mean F-ratio contribution of different frequency bands for speaker identification, the adaptive frequency warping is used to extract the discriminative features. Speaker identification experiments showed that the phoneme mean *F-ratio* method can well emphasized speaker information, to some extent, subtract the linguistic information.

However, we conducted this study on the Korean speech corpus, and we will test the efficiency of PMFFCC characteristics for speaker identification with other languages and investigate the experimental results in the future.

REFERENCES

[1] J. P. Campbell, "Speaker recognition: a tutorial", Proceedings of the IEEE, vol. 85, pp. 1437-1462, September 1997.

[2] S. Hayakawa and F. Itakura, "Text-dependent speaker recognition using the information in the higher frequency band", in Proc. IEEE int. Conf. Acoust. Speech, Signal Process., Adelaide, Australia, 1994, pp. 19-22.

[3] O. Orman and L. Arslan, "Frequency analysis of speaker identification", Proc. of Speaker Odyssey: The Speaker Recognition Workshop, pp. 219-222, Crete, Greece, 2001.

[4] C. Miyajima, H. Watanable, K. Tokuda and S. Katagiri, "A new approach to designing a feature extractor in speaker identification based on discriminative feature extraction", Speech Communication, vol. 35, no. 3-4, pp. 203-218, October 2001.

[5] Y. Yu, D. Yuan and F. Xue, "A non-linear frequency transform for speaker recognition", ACTA ACUSTICA, vol. 33, no. 5, pp. 450-455, 2008.

[6] X. Lu and J. Dang, "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification", Speech Communication, vol. 50, pp. 312-322, 2008.

[7] L. Rabiner and B. H. Juang, "Fundamentals of Speech Recognition", Prentice Hall PTR, 1993.

[8] H. Suzuki, T. Nakai, J. Dang and C. Lu, "Speech production model involving subglottal structure and oral-nasal coupling through closed velum", Proc. ICSLP90, vol. 1, pp. 437-440.

[9] J. Dang and K. Honda, "Acoustic characteristics of the human paranasal sinuses derived from transmission characteristic measurement and morphological observation", J. Acoust. Soc. Am., vol. 100, pp. 3374-3383, 1996.

[10] J. Dang and K. Honda, "An improved vocal tract model of vowel production implementing piriform fossa resonance and transvelar nasal coupling", Proc. ICSLP1996, pp. 965-968, 1996.

[11] J. Dang and K. Honda, "Acoustic characteristics of the piriform fossa in models and humans", J. Acous. Soc. Am., vol. 101, pp. 456-465, 1997.

[12] T. Kitamura, K. Honda and H. Takemoto, "Individual variation of the hypopharyngeal cavities and its acoustic effects", Acous. Sci. Tech., vol. 26, no. 1, pp. 16-26, 2005.