Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 50 (2008) 312–322
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification

Xugang Lu *, Jianwu Dang

Japan Advanced Institute of Science and Technology, 1-1, Asahidai, Nomi, Ishikawa 923-1292, Japan

Received 18 June 2007; received in revised form 23 October 2007; accepted 27 October 2007

Abstract

The features used for speech recognition are expected to emphasize linguistic information while suppressing individual differences. For speaker recognition, in contrast, features should preserve individual information and attenuate the linguistic information at the same time. In most studies, however, identical acoustic features are used for the different missions of speaker and speech recognition. In this paper, we first investigated the relationships between the frequency components and the vocal tract based on speech production. We found that the individual information is encoded non-uniformly in different frequency bands of speech sound. Then we adopted statistical Fisher's *F*-ratio and information-theoretic mutual information measurements to measure the dependencies between frequency components and individual characteristics based on a speaker recognition database (NTT-VR). From the analysis, we not only confirmed the finding of non-uniform distribution of individual information in different frequency bands from the speech production point of view, but also quantified their dependencies. Based on the quantification results, we proposed a new physiological feature which emphasizes individual information for text-independent speaker identification by using a non-uniform subband processing strategy to emphasize the physiological information involved in speech production. The new feature was combined with GMM speaker models and applied to the NTT-VR speaker recognition database. The speaker identification using proposed feature reduced the identification error rate 20.1% compared that with MFCC feature. The experimental results confirmed that emphasizing the features from highly individual-dependent frequency bands is valid for improving speaker recognition performance.

PACS: 01.30.−y

Keywords: Speaker identification; Physiological features; Speech production; Fisher's *F*-ratio; Mutual information; Frequency warping

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 algorithms are also based on modeling the LPC or MFCC feature using Gaussian mixture model (GMM) (Reynold, 1995). However, the purpose of speech recognition is quite different from that of speaker recognition, the former task needs to emphasize linguistic information and suppress individual information, while the later task needs to preserve individual information. This contradiction suggests that either LPC or MFCC may not meet the requirements of both speech and speaker recognition tasks.

1.1. Speaker feature extraction based on signal processing

For speaker recognition, the problem is how to extract and utilize the information that characterizes individual speakers. Generally speaking, individual information of speakers results mainly from two factors: physiological and social factors. The former are related to the speaker's gender, age, and oral morphology, which are inborn

* Corresponding author.
E-mail addresses: xugang@jaist.ac.jp (X. Lu), jdang@jaist.ac.jp (J. Dang).

0167-6393/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2007.10.005

characteristics; the latter concerns the speaker's dialect, idiolect, occupation, and so on, which result from his/her social environment. In this paper, we focus on the former factors, and investigate their acoustic characteristics in speech. In other words, we attempt to extract individual information which is involved in morphological details and acoustic characteristics, and implement it in speaker recognition.

Usually, when producing a speech sound, speakers' physiological and morphological features are encoded in acoustic characteristics of the sound. The diverse articulators contribute different physical properties in the acoustic spectrum, which are personalized in individual morphologies. In order to extract that information, some speech feature representations have been developed. The LPC feature can well model the vocal tract properties by using an all-pole model, which reflects the main vocal tract resonance property in the acoustic spectrum (Rabiner and Juang, 1993; Stevens, 1998). This feature emphasizes the formant structure that concerns major individual differences of the speakers, while some significant details of individuals such as the nasal, piriform fossa and other side branches are ignored. In contrast, the MFCC feature takes the mechanism of the auditory nonlinear frequency resolution into consideration, which improves the representation robustness (Rabiner and Juang, 1993). For extracting more direct physiological features, the fundamental frequency or pitch which reflects the vocal cord information of speakers is often used (Atal, 1976). The LPC residual signal has also been proposed for describing the speakers' glottal information (He and Liu, 1995). When these features were used for speaker recognition, the performance was improved to some extent (Atal, 1976; He and Liu, 1995). For speaker recognition, the essential goal is to find the non-linguistic information which is highly correlated with individual characteristics from speech sounds, those studies discussed above endeavored to extract the intrinsic individual information for speaker recognition task. However, how to find and extract the intrinsic individual-related acoustic features is still a difficult problem.

1.2. Speaker feature investigation based on speech production

The main focus of this study is to provide a systematic analysis of the relationship between acoustic frequency components and individual characteristics from both the speech production point of view, and the statistical information processing point of view, and to propose a physiological feature extraction method for speaker recognition. Actually, in essence, most of the previous studies try to extract the features for speaker recognition from the main vocal tract physical property, which is usually described by the phoneme-dependent dynamics of vocal tracts. However, the speaker-dependent features are more important for speaker recognition which is expected to be invariant in the articulation dynamics. In the vocal tract, during speech production, there are number of side branches, such as the nose, piriform fossa, etc., which have less variation during speech and introduce invariant features in some specific frequency regions (Suzuki et al., 1990; Dang and Honda, 1994, 1996a,b). The frequency regions concerned with the side branches may be related to paralinguistic or non-linguistic information. One important characteristic of those side branches is that they show large variation across speakers, but have small changes during speech production for the same speaker. In addition, they are not easily changed by conscious efforts. In other words, the frequency regions produced by those side branches are not easily disguised in the speech. The acoustic features around those frequency regions should be suitable for individual description.

Based on the analysis above, in this paper, we investigate new features which reflect the important speaker-specific information in frequency domain, design a subband processing strategy for feature extraction, and apply it to the speaker identification task. This paper is organized as follows. Section 2 analyzes encoding of speaker individual information from speech production point of view, i.e., analyzing the individual physiological features caused by speaker individual speech organs' morphology, which shows the non-uniform frequency dependency of individual characteristics. Section 3 introduces two methods, i.e., Fisher's *F*-ratio (Wolf, 1972) and mutual information (Cover and Thomas, 1991), to quantify the dependencies between frequency components and speaker identities. Based on the results from Sections 3 and 4 describes the proposed non-uniform subband processing algorithm, and extracts the new speaker physiological feature. In Section 5, the speaker model using HTK is first introduced, then speaker identification experiments are performed to test the new feature extracted based on the algorithm in Section 4. In addition, in Section 5, the performance of the system with the proposed feature is compared with those using traditional features. Finally, Section 6 gives conclusions and future directions.

2. Encoding of speaker individual information in speech production

In order to extract intrinsic speaker features, we must know where and how the speaker features are encoded during speech production. In this section, we try to clarify distributions of speaker features in frequency domain, and explain the intrinsic connections between the frequency components and physical factors for speaker individual information representation based on several speech production studies (Suzuki et al., 1990; Dang and Honda, 1994, 1997).

From the view point of speech production, speech sound results from a sound source modulated by the vocal tract filter. The speaker-specific information should be involved in the invariant characteristics of the sources and the vocal tracts of the speakers. In this study, we are more concerned

with invariant factors in the physiology and morphology of the vocal tract. Although the vocal tract is approximately treated as a single tube in many cases, the vocal tract possesses a complicated shape consisting of the main tract and a number of side branches. Fig. 1 illustrates the human speech production system, and a diagram of a speech production model. From the upper panel of Fig. 1, one can see that the vocal tract configuration consists of complex side branches and cavities, where the piriform fossa shown by dashed lines are located bilaterally behind the larynx, and cannot be seen from the midsagittal plane. For a simplified simulation of speech production image, the vocal tract is modeled as a main tube with several side branch tubes as shown in lower panel of Fig. 1. The biggest side branch within the vocal tract is the nasal passage (Dang and Honda, 1994, 1997). In producing nasal and nasalized sounds, the nasal cavity is coupled with the oral cavity by lowering the velum. 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 (Suzuki et al., 1990; Dang and Honda, 1996b). In the nasal cavity, there are a number of paranasal cavities that contribute anti-resonances (zeros) to the transfer function of the vocal tract. The acoustical effects of the paranasal cavities have been experienced by every one who catches a cold, in which the entrances of the paranasal cavities are obstructed by mucosa. Since the nasal cavity has a complicated structure and quite large individual differences, the nasal sounds not only offer several distinguishing phonetic units, but also provide a lot of individual information of speakers.

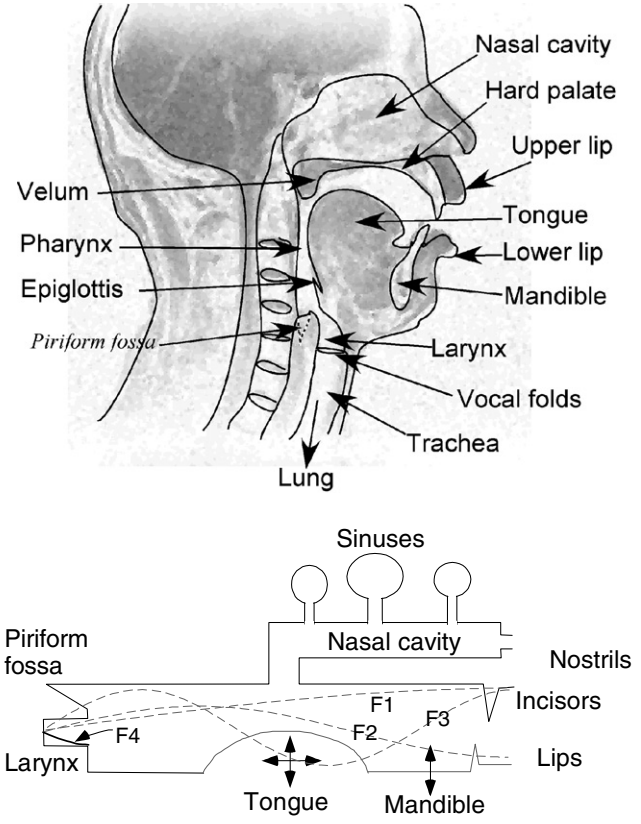

Fig. 1. Vocal tract of human speech production system (upper panel), and a model of the vocal tract with side branches (lower panel). The standing waveforms of four formants are also illustrated in the vocal tract.

The lower panel of Fig. 1 shows an important side branch of the vocal tract, the piriform fossa which is located behind the larynx bilaterally (Dang and Honda, 1996a, 1997). The piriform fossa is the entrance of the esophagus, and is shaped like twin cone-like cavities on the left and right sides of the larynx as shown in upper panel of Fig. 1 (small dashed cavities). These side branches have anti-resonances between 4 kHz and 5 kHz. In speech production model, introducing the piriform fossa module into the production model will cause spectral structure changes in frequency region between 4 kHz and 5 kHz, which can fit the real acoustic speech spectrum well. In addition, the piriform fossa cavities are speaker dependent and less changed during speech production, for finding invariant features for speaker recognition, the piriform fossa should be regarded as one important "cue" (Dang and Honda, 1996a, 1997).

Recently, Takemoto et al., conducted a number of studies to clarify the correspondence of acoustical features to specific parts of the vocal tract (Takemoto et al., 2006). They found that when producing vowels, the first three formants vary with the vocal tract, while the fourth one is almost constant. This phenomenon is illustrated in Fig. 1. As shown in upper panel of Fig. 1, the throat part of vocal tract consists of the pharynx and larynx, and the larynx tube connects the pharynx via the outlet of the larynx. In the acoustic system of the vocal tract, the laryngeal tube can be considered to be an independent subsystem, because the area ratio of the bottom of the pharynx to the outlet of the larynx is so large that the pharynx can be roughly treated as a free space for the larynx in the high frequency region. Since the larynx length is different for different speakers, this would be a physiological factor for generating individual differences in speech.

The vocal folds shown in upper panel of Fig. 1, locating between the trachea and larynx, modulates the air input from lung. The vibration frequency of the vocal folds, i.e., the fundamental frequency in acoustic speech ranging between 100 Hz and 400 Hz, depends on the length and stiffness of the vocal folds, which is a speaker-dependent characteristic. This property has already been used as an important speaker individual feature for speaker recognition (Atal, 1976).

In summary, as shown above, the speaker-specific features caused by different articulatory speech organs are distributed non-uniformly in high frequency bands. Traditional feature extraction methods focus on the large spectral peaks caused by the movements of the vocal tract and emphasize the lower frequency bands. Our analysis showed that some invariant parts of the vocal tract, such as the nasal cavity, piriform fossa, laryngeal tube, etc.,

which serve as speaker information "cues" are encoded in the spectrum of speech sounds. It is possible to emphasize speaker information by identifying and manipulating the locations of their frequencies.

3. Quantization of the dependencies between frequency components and speaker identities

From Section 2, we found that 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 phonemic 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 (Rabiner and Juang, 1993; Stevens, 1998). This kind of non-uniform distribution of speaker information in frequency bands was also confirmed in (Hayakawa and Itakura, 1995; Miyajima et al., 1999). In MFCC feature representation, the Mel frequency scale is used to get a high resolution in low frequency region, and a low resolution in high frequency region. This kind of processing is good for obtaining stable phonetic information, but not suitable for speaker features that are located in high frequency regions. Since the speech organs with a relatively invariant shape contribute acoustic properties to different frequency regions, we must find out the frequency regions that correspond to the individual information, and develop a decoding method that has higher frequency resolution in those specific frequency regions. Miyajima et al. (1999), used a frequency warping function rather than using Mel frequency warping function to process speech spectrum. However, their warping function was a monotonic warping function, which cannot reflect the non-uniform, non-monotonic distribution property of speaker information in frequency domain. For choosing a better frequency warping, we need to quantify the contribution of each frequency component to speaker individual information description.

3.1. Speaker information measurement based on F-ratio

For investigating the contribution of each frequency region to speaker recognition, we use triangle-shaped band-pass filters with linear frequency scale to process the speech power spectrum. The triangle-shaped filters are shown in Fig. 2. Each filter band gives an integrated energy around the center frequency of the filter band (total 60 frequency bands). 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 dependencies between frequency band outputs and speaker identities.

The Fisher's F-ratio is widely used to measure the discriminative ability of a feature for pattern recognition (Webb, 2002). We adopt it to measure the speaker discriminative score in each frequency band, which is used as an index of dependency or importance of speaker information (Wolf, 1972; Campbell, 1997). In this study, the average F-ratio is defined as

$$\text{F\_Ratio} = \frac{\frac{1}{M}\sum_{i=1}^{M}(u_i - u)^2}{\frac{1}{M\cdot N}\sum_{i=1}^{M}\sum_{j=1}^{N}(x_i^j - u_i)^2} \tag{1}$$

where x_i^j is one subband energy of the jth speech frame of speaker i with $j = 1, 2, \ldots, N$, and $i = 1, 2, \ldots, M$. u_i and u are the subband energy averages for speaker i and for all speakers, respectively, which are defined as

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

Eq. (1) is the ratio of the inter-speaker variance to intra-speaker variance in a given frequency band. A larger value obtained in a frequency band means that more speaker information is encoded in that band.

3.2. Speaker dependency measurement using mutual information

Mutual information is a powerful tool to measure the dependency between random variables (Cover and Thomas, 1991). Suppose a discrete speech feature variable and speaker class variable are Y and C, respectively, the mutual information between the two variables is defined as

$$I(Y;C) = H(Y) + H(C) - H(Y,C) \tag{3}$$

where $H(\cdot)$ is an entropy function. For variable Y it is defined as

$$H(Y) = -\sum_{y\in Y} p(y)\log_2 p(y) \tag{4}$$

From Eq. (3), one can see that the mutual information equals the sum of the entropies of the speech feature and speaker class variables, by reducing the joint entropy between them. For a better understanding of the mutual information of the two variables, Eq. (3) is rewritten as

$$I(Y;C) = H(Y) - H(Y|C) \tag{5}$$

One can see from Eq. (5) that the mutual information equals the entropy of the feature variable, reducing the conditional entropy under the condition of knowing the speaker identity. In other words, the mutual information of the feature variable equals the uncertainty reduction of the feature variable when the speaker identity is known. Also from Eq. (5), one can see that if the feature variable has no speaker information, the mutual information between them equals zero. For estimating the mutual information of two variables using Eq. (5), we only need to estimate the entropy and joint entropy, which requires the estimation of the probabilities of one and two variables. In this study, we use the histogram method to estimate the probabilities for the calculation of entropy and joint entropy (Webb, 2002). Based on this mutual information

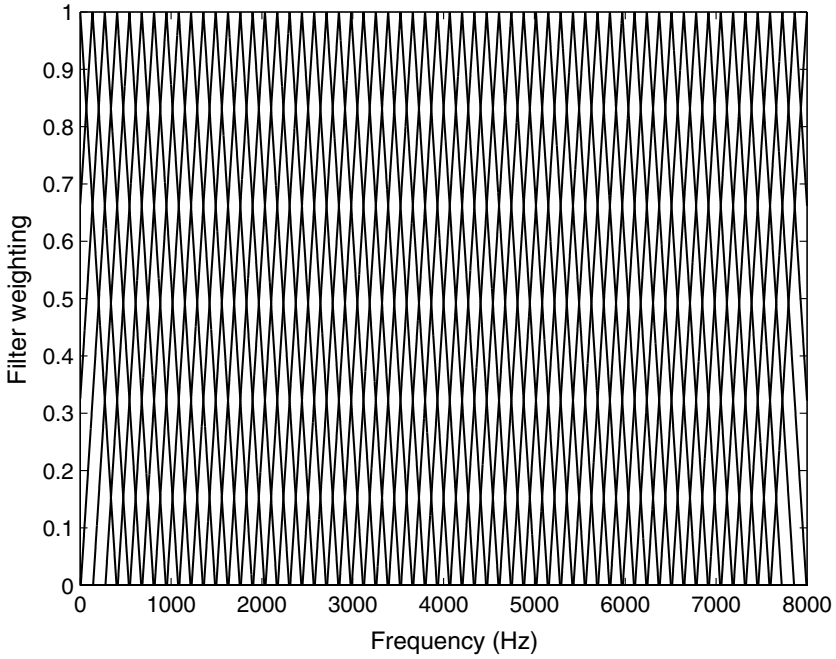

Fig. 2. Uniform subband filters with uniform bandwidth.

criterion, we can quantify the dependency between the feature variable from a given frequency band and speaker class variable.

3.3. Dependency measurement results based on F-ratio and mutual information

The NTT-VR speaker recognition database is used in this study, in which there were 35 speakers in total, including 22 male speakers and 13 female speakers (Matsui and Furui, 1992). The speech was collected in five sessions over a period of 10 months. In each session, each speaker was asked to speak sentences with normal, slow and fast speech rates. The average duration of the sentences is about 4 s. The speaker discriminative abilities, quantified using *F*-ratio and mutual information for each frequency band (totally, 60 frequency bands are used), are shown in Figs. 3 and 4, respectively. In Figs. 3 and 4, the session recorded in August, 1990 is denoted as session 90-8, and other sessions are denoted analogously. The results in Figs. 3 and 4 are consistent with peaks and valleys located in similar frequency regions on the curves. From both figures, one can see that the distribution of speaker discriminative information is almost invariant over the time span (different recording sessions). Of course besides these consistent common characteristics of the two figures, there are two main differences between Figs. 3 and 4. One is that in Fig. 4 the first large peak region in the fundamental frequency region (100–300 Hz) is comparatively higher than peaks in other frequency regions. The second is that the mutual information gives dependency measurement in bits, based on coding length, as used in information theory (Cover and Thomas, 1991). From the common characteristics of the two figures, one can see that, the speaker discriminative information is mainly concentrated in three regions in the frequency domain (three large peak regions). As it is known that the fundamental frequency of speech is caused by the vibration of the vocal cords. The vocal cord of different speaker has different properties, such as length, stiffness, etc., which results in varying fundamental frequency. So the first peak region with lowest frequency from 100 Hz to 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 (Dang and Honda, 1996a), introducing the piriform fossa module in the speech production model causes spectral structure changes in frequency region from 4 kHz to 5 kHz. In addition, the 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 (Dang and Honda, 1996a, 1997). The frequency region from 6.5 kHz to 7.8 kHz seems to be related to the consonants, probably to the location of their constrictions. 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 discriminative information is concentrated in this frequency region, which is consistent among all the speakers for phoneme description. This statistical result confirms our analysis in Section 2, i.e., speaker information is not uniformly distributed in each frequency band. The similar conclusion was also drawn in (Hayakawa and Itakura, 1995).

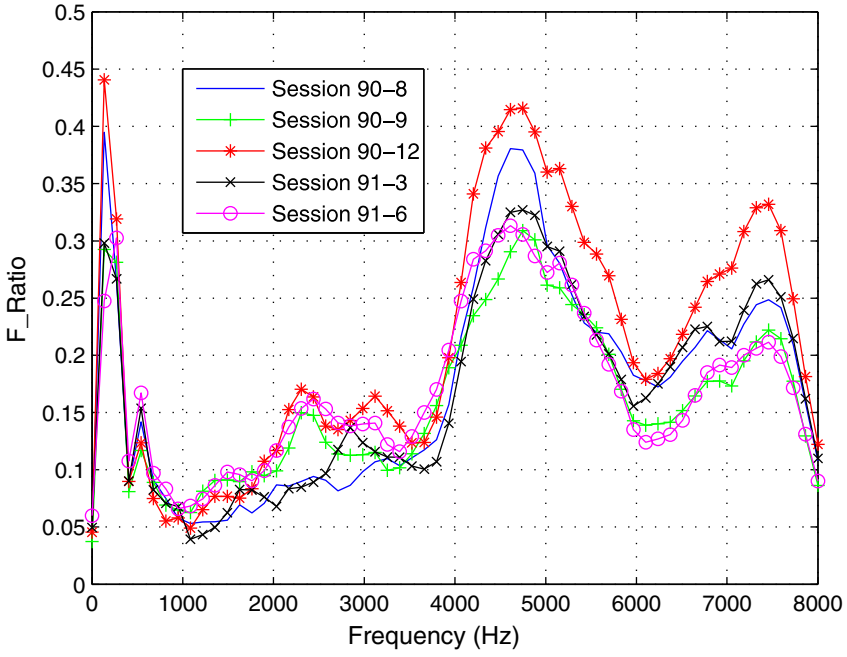

Fig. 3. Speaker discriminative score in frequency domain using *F*-ratio.

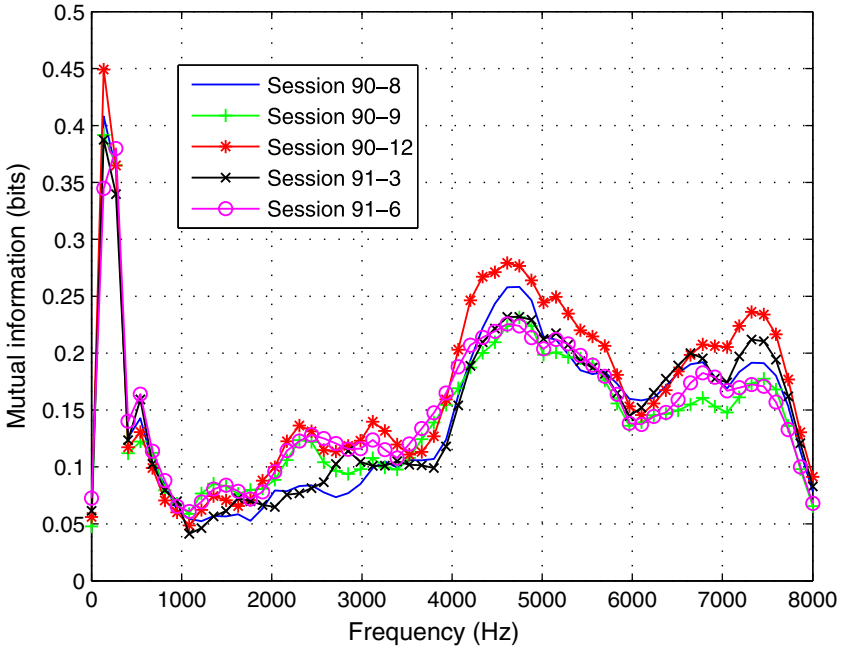

Fig. 4. Speaker discriminative score in frequency domain using mutual information.

4. Non-uniform subband filtering for speaker feature extraction

According to the results in Section 3, we can design subband processing algorithms to extract speaker-dependent features for speaker identification. There are two solutions to emphasize the contribution of those frequency bands with more speaker information. The first solution is to give weighting coefficients for different frequency bands, based on their dependencies on speaker identity class. The second solution is to use non-uniform frequency warping, which gives different frequency resolutions to different frequency regions according to their dependencies on speaker identity class. Thus the spectral structure or profile around different frequency regions may be refined or smoothed. The frequency warping function we used is quite different from that was used in (Miyajima et al., 1999), which used a monotonic frequency warping function. Our preliminary

experiments showed that the first solution does not provide significant improvement for speaker identification performance. In the following, we focus on the second solution for speaker feature extraction.

We designed the non-uniform subband filters by considering the dependency measurements in order to change frequency resolutions in different frequency regions. In real applications, the dependency measurement using *F*-ratio is easier than that using mutual information. In addition, the fundamental frequency can be changed by speakers' conscious efforts, it should not be emphasized so much. Therefore, in this study we adopt the dependency measurement calculated using *F*-ratio. The designed non-uniform subband filters for speech signal processing during the feature extraction stage are shown in Fig. 5. The bandwidth of each subband filter is assigned to be inversely proportional to the *F*-ratio. By this processing, the frequency resolution around a frequency region with a high *F*-ratio is improved. Comparing the filters in Fig. 2 with those in Fig. 5, one can see that the center frequencies of filter bands are not distributed uniformly along the frequency axis in Fig. 5. The scale for frequency warping is different from either a linear scale or log scale. The bandwidths of the filters are assigned equally in Fig. 2, while they are non-uniformly assigned in Fig. 5 which causes the non-uniform frequency resolution around the center frequencies for the proposed filters. In order to compare frequency resolution changes obtained by different methods, we plot the frequency resolution curves of the proposed non-uniform frequency warping processing method and the Mel frequency warping method in Fig. 6, where the frequency resolution curve of uniform filter bands method is also plotted for a reference. From Fig. 6, for the same subband number, one can see that Mel frequency warping method has high resolution in low frequency region, while the non-uniform subband processing method has high resolution in the frequency regions with high *F*-ratio values. In other words, the Mel frequency warping method has high resolution in the region which encodes phonetic information, while the proposed non-uniform frequency warping processing method shows high resolution in the frequency regions which encodes more speaker information.

Since the spectral envelope extracted with the non-uniform frequency warping filters designed based on *F*-ratio emphasizes speaker-specific information, it is expected that the feature extracted using these non-uniform subband filters will improve the speaker identification performance. In the following section, we extract the feature set using the non-uniform subband filters, and apply it to speaker identification experiments.

5. Speaker identification experiments

In a speaker identification system, each speaker is represented by a statistic modeling of the speech features. The model parameters for each speaker are estimated during the training process. A speaker class is identified when a speaker model can give the maximum likelihood probability for the input speech data. In this section, we train and test the speaker identification system using our proposed feature set, and compare its performance with those of uniform subband and Mel frequency subband-based features.

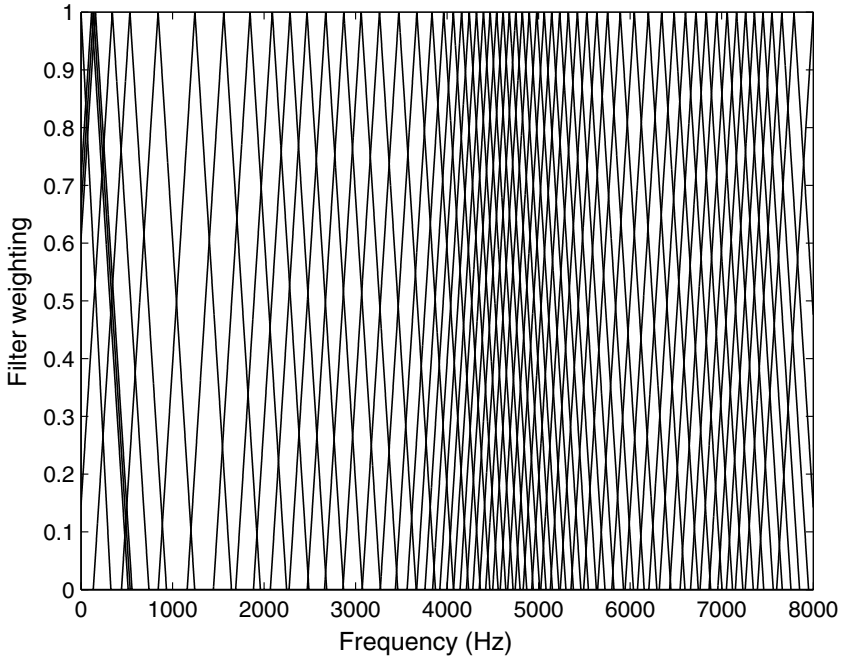

Fig. 5. Non-uniform subband filters with non-uniform bandwidth.

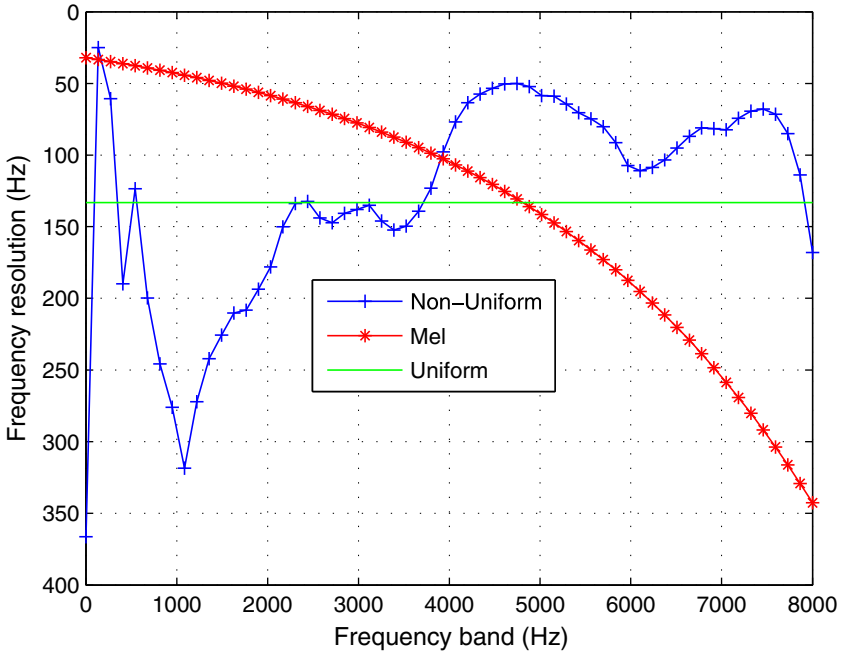

Fig. 6. Comparison of frequency resolutions of filter bands with non-uniform processing (solid curve with plus symbol), Mel scale processing (solid curve with circle symbol), and uniform processing (solid curve).

5.1. Speaker modeling using GMM

GMM is widely used for speaker modeling in context-independent speaker identification (Reynold, 1995). In our research, we use HTK (Young, 1992) to design the GMM speaker models. 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.

5.2. Speaker feature extraction

The processing diagram for speaker feature extraction is shown in Fig. 7. In the feature extraction process, 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 16 ms frame length and 8 ms shift is employed. Sixty 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 coefficient vectors (zeroth order cepstral coefficient was excluded). Finally, the proposed feature vectors are extracted for speaker modeling.

For comparison, the feature sets are extracted using the framework of Fig. 7 with Mel frequency subband filters, uniform frequency subband filters (in Fig. 2) as well as with the proposed non-uniform frequency subband filters (in Fig. 5). The features are denoted as MFCC, UFCC, and NUFCC, respectively.

5.3. Speaker identification experiments and comparisons

We conducted speaker identification experiments on NTT-VR database. For training speaker models, 10 sentences uttered at normal speaking rate were used for each

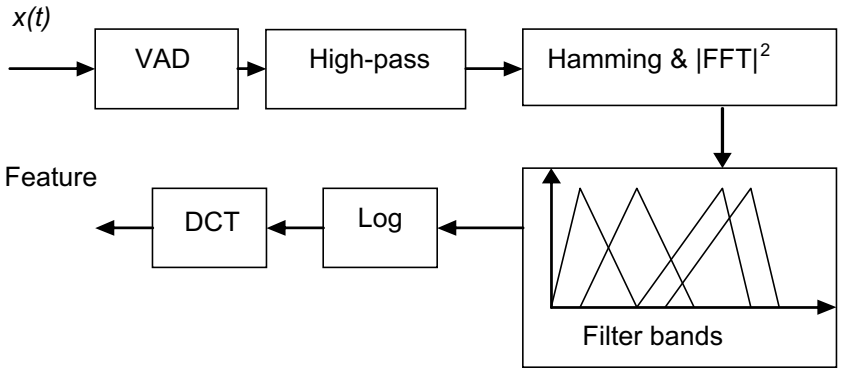

Fig. 7. Speaker feature extraction diagram.

speaker, from the session 90-8 which was recorded in August, 1990. All other utterances, including sentences and words in all sessions at different speaking rates were used in the identification. The feature sets were modeled by both the diagonal covariance matrix and full covariance matrix settings. The speaker identification experiments were first conducted by increasing the Gaussian mixture numbers for the speaker models. For full covariance matrix setting, however, the Gaussian mixture number was limited when the training data set was not large enough. By increasing the mixture number step by step, the highest identification rate was achieved when the mixture number was 32 for the diagonal covariance matrix setting, and 4 for the full covariance matrix setting, respectively. The performance using full covariance matrix setting is a little better than that using the diagonal covariance matrix setting. Therefore, the full covariance Gaussian mixture model with mixture number 4 was chosen for all identification experiments. The identification results are shown in Fig. 8 for the three feature sets.

In Fig. 8, the horizontal axis is the session index, and the Aver. is the average identification rate for all sessions. From Fig. 8, one can see that the identification rate is high in session 90-8, since the speaker models were trained using the data set from the same session, although the speech rate of some testing data is different from that of the training data set. And the identification rate for session 91-6 is slightly lower than for other sessions, which suggests that the speakers changed their speaking style more over the period of 10 months from session 90-8 compared with other sessions. The performance of feature MFCC is a little bit better than that of UFCC. Among the three feature sets, the proposed non-uniform frequency feature (NUFCC) demonstrated the best performance. Quantitatively, compared with the baseline feature set MFCC, the identification error reduction rate is about 20.1% on average for all testing sessions. Based on this result, we conclude that the feature extracted by emphasizing the individual induced frequency regions could improve speaker identification performance.

6. Discussions and conclusions

In this study, we first analyzed speaker physiological information from the speech production point of view. The analysis showed that speaker information is deeply concerned with physiological and morphological differences of the speech organs. Speaker information was investigated quantitatively in the frequency domain using the *F*-ratio and mutual information measurements. The results showed that the glottis causes one important feature of speaker information in a low frequency region, the piriform fossa causes one dominant speaker discriminative information feature in frequency region between 4 kHz and 5.5 kHz, and the constriction of the consonants would be another factor in the higher frequency region around 7.5 kHz. In contrast, there is less speaker discriminative information in the region from 0.5 kHz to 3.5 kHz. According to these results, we designed non-uniform subband filters to extract speaker physiology-dependent features. Speaker identification experiments showed that the feature extracted using the proposed non-uniform subband processing improved speaker identification performance. The error reduction rate was 20.1% compared with the baseline model with MFCC.

Our study demonstrated the usefulness of the proposed feature extraction method, based on the evidence of

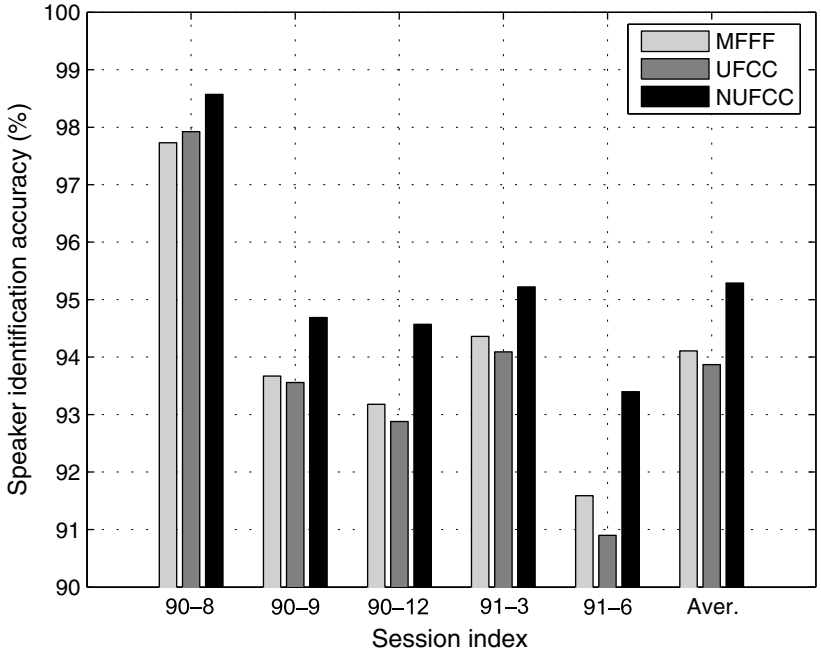

Fig. 8. Identification rates for the three feature sets.

non-uniform distribution of speaker-specific information in frequency domain. The evidence was derived from speech production point of view and was confirmed and quantified using statistical analysis methods. However, some expectations were not confirmed. For example, the nasal or nasalized characteristic is a speaker information which is used for speaker recognition by human beings (Atal, 1976). But the statistical analysis could not find the correlation of nasal characteristics with speaker identity. The reason may be that the statistical analysis methods using many acoustic signals average out the nasalized characteristic. Also, our experimental results suggest that besides considering the distribution of the speaker information in different frequency regions, we need to think about the distribution of speaker information in different time periods. For a speaker uttering sentences, sometimes it is easy to identify the speaker using a few utterances, but sometimes difficult to identify the speaker using some other utterances. Therefore, there is a problem of how to identify the time periods in speech sentences which encode more speaker information for speaker modeling and recognition.

In our experiments, the training data set has 10 * 35 sentences from one recording session (35 speakers, and 10 sentences for each speaker), which is used for F_Ratio calculation to design non-uniform subband filters, and training the GMM speaker models. The testing data set has 116 * 35 * 5 sentences (35 speakers, 116 sentences for each speaker, and five recording sessions). In our study, we not only investigated the dependencies between frequency components and speaker individual characteristics using information-theoretic measurement, but also associated the frequency components with clear physical meanings, i.e., the relationship between frequency components and vocal tract of individual speakers. In addition, in our study, we found that even if the frequency warping filter may differ a little when using separate data set and separate speaker set for calculation, consistent tendency of the relationship between frequency components and speaker individual characteristics is shown. We tested two data sets to quantify the dependency relationship between frequency components and speaker individual characteristics. Data set one has 10 * 35 sentences (35 speakers, and 10 sentences for each speaker). Data set two has 5 * 17 sentences (17 speakers, and five sentences for each speaker). We calculated the F_Ratios for the two data sets, and showed the results in Fig. 9. From Fig. 9, one can see that the F_Ratio curves of data sets one and two show almost consistent tendency in peaks and valleys, i.e., the dependency relationship between frequency components and speaker individual characteristics is consistent for the two data sets. This result shows that the physiological feature we focused on is quite stable.

In our study, because the speech recognition database was recorded in the same recording room, we have not examined the robustness of our proposed method to some external distortions, for examples, the distortion of environment noise, of transmission line channels, etc. However, based on our study, if the speech is distorted because of environmental changes, the frequency components should be adaptively selected to be emphasized for speaker recognition. Based on this consideration, the ideal case is to find an adaptive "universal" dependency relationship function, which can be used to emphasize speaker recognition "cues" to process acoustic speech adaptively (such as to adaptive to external distortions, phoneme categories, etc.). Our study is one important step towards to this goal.

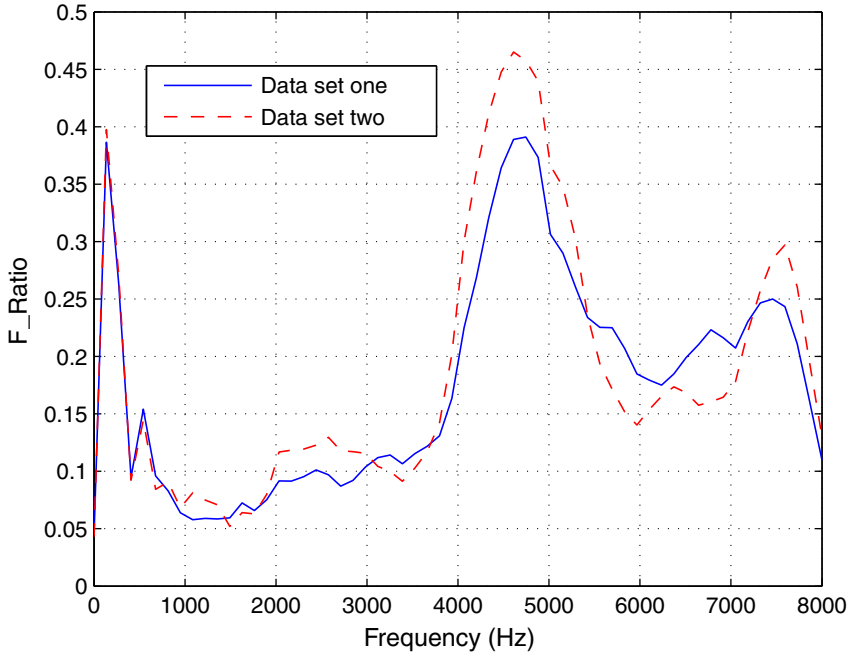

Fig. 9. F_Ratio curves of data sets one and two.

Since both the speaker information and linguistic information are encoded in the same acoustic speech characteristics, for speaker recognition, it is necessary to identify the parts contributed by speaker characteristics. For improving the speaker identification performance, in essence, it is necessary to find out common linguistic information from acoustic signals, and discard or attenuate that information during speaker modeling, since the involvement of the linguistic information in speaker modeling may degrade the discriminative ability of the models. In addition, it is necessary to find out the invariant speaker information, for instance, human being can identify a speaker by using one coughing, which is an impulse response and includes the morphological information of the vocal tract. However, we have no idea yet how human learns and uses this information. How to find out this information, and how to integrate this information using a statistical framework for speaker identification, are remained for our future work.

Acknowledgement

This study is supported in part by SCOPE (071705001) of Ministry of Internal Affairs and Communications (MIC) of Japan, Grant-in-Aid for Scientific Research of Japan No. 17300182 and Grant No. 18700172.

References

Atal, B., 1976. Automatic recognition of speakers from their voices. Proc. IEEE 64, 460–475.

Campbell, J., 1997. Speaker recognition: A tutorial. Proc. IEEE 85 (9), 1437–1462.

Cover, Thomas M., Thomas, Joy A., 1991. Elements of Information Theory. Wiley-Interscience.

Dang, J., Honda, K., 1994. Morphological and acoustical analysis of the nasal and the paranasal cavities. J. Acoust. Soc. Am. 96, 2088–2100.

Dang, J., Honda, K., 1996a. An improved vocal tract model of vowel production implementing piriform fossa resonance and transvelar nasal coupling. In: Proc. ICSLP1996, pp. 965–968.

Dang, J., Honda, K., 1996b. Acoustic characteristics of the human paranasal sinuses derived from transmission characteristic measurement and morphological observation. J. Acoust. Soc. Am. 100, 3374–3383.

Dang, J., Honda, K., 1997. Acoustic characteristics of the piriform fossa in models and humans. J. Acoust. Soc. Am. 101, 456–465.

Hayakawa, S., Itakura, F., 1995. Text-dependent speaker recognition using the information in the higher frequency band. In: Proc. ICASSP1994, pp. I-137–I-140.

He, J., Liu, L., 1995. On the use of features from prediction residual signals in speaker identification. In: Proc. EUROSPEECH95, Madrid, Spain, Vol. 1, pp. 313–316.

Matsui, T., Furui, S., 1992. Comparison of text-independent speaker recognition methods using VQ distortion and discrete/continuous HMMS. In: Proc. ICASSP 92, Vol. II, pp. 157–160.

Miyajima, C., Watanabe, H., Kitamura, T., Katagiri, S., 1999. Discriminative feature extraction – Optimization of Mel-cepstral features using second-order all-pass warping function. In: Proc. EUROSPEECH1999, pp. II-779–I-782.

Rabiner, L., Juang, B.H., 1993. Fundamentals of Speech Recognition. Prentice Hall PTR.

Reynold, D.A., 1995. Speaker identification and verification using Gaussian mixture speaker models. Speech Comm. 17, 91–108.

Stevens, Kenneth N., 1998. Acoustic Phonetics. The MIT press.

Suzuki, H., Nakai, T., Dang, J., Lu, C., 1990. Speech production model involving subglottal structure and oral–nasal coupling through closed velum. In: Proc. ICSLP90, pp. 437–440.

Takemoto, H., Adachi, S., Kitamura, T., Mokhtari, P., Honda, K., 2006. Acoustic roles of the laryngeal cavity in vocal tract resonance. J. Acoust. Soc. Am. 120, 2228–2239.

Webb, Andrew R., 2002. Statistical Pattern Recognition. John Wiley & Sons.

Wolf, J.J., 1972. Efficient acoustic parameters for speaker recognition. J. Acoust. Soc. Am. 51, 2044–2056.

Young, Steven, 1992. HTK Tutorial Book. <http://htk.eng.cam.ac.uk/>.