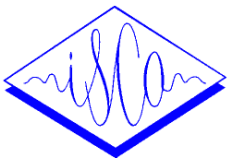

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

On the Use of Bandpass Liftering in Speaker Recognition[1]

Bin Zhen, Xihong Wu, Zhimin Liu, and Huisheng Chi

(National Key Lab on Machine Perception, Center for Information Science,

Peking University, Beijing, 100871, China)

Abstract

The measurements of speech spectral envelopes may not accurately characterize the true speech spectrum because of analysis model constraints, such as window position fluctuations, excitation interference, and measurement noise. Juang found that these undesirable spectral measurement variations could be partially reduced by bandpass liftering, and gained better results in speech recognition. In this paper, we propose a new bandpass liftering process for speaker recognition. The new liftering process reduces the error rate of 43% than those without the liftering and 9% than Juang's method in speaker recognition. From the experimental result, we found that the spectral contours contain both speech and speaker information and the fine spectral structures contain speaker-discriminating information.

Keywords: liftering, speaker recognition, feature extraction

1. Introduction

Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual information included in the speech waves.[4,5] Speaker identity is correlated with the physiological and behavioral characteristics of speaker. These characteristics exist both in the spectral envelope (vocal tract characteristics) and in the supra-segmental features (voice source characteristics and dynamic features spanning several segments). Speech recognition is the process of automatically recognizing what is speaking on the basis of common information included in the speech waves.[2,3] There are two kinds of variability in the speech waves.[2,3] Within-speaker variability can result from changes in the speaker's physical and emotional state, speaking rate or voice quality. Across-speaker variability can result from differences in sociolingustistic background, dialect, and vocal tract size and shape.

Until now, the speech and speaker recognition tasks have similar steps, such as speech analysis, similarity calculation, time normalization, and decision logic. The speech analysis procedure performed on the raw input speech waveform, results in some representation of the signal that characterizes the relevant features of the spoken speech. Juang said that the procedure can be regarded as a data reduction procedure that retains the vital characteristics of the signal and eliminates undesirable interference from irrelevant characteristics of the speech.[1] However, the measurements of speech spectral envelopes may not accurately characterize the true speech spectrum because of analysis model constraints, such as window position fluctuations, excitation interference and measurement noise.[1,3] Juang has found that these undesirable spectral measurement variations can be partially reduced by a bandpass liftering process in cepstrum domain, and gained better results in speech recognition.[1]

As it known, the characteristics of the speech signal are determined primarily by the linguistic message, the secondary speech message, including the speaker discriminations are encoded as nonlinguistic articulatory variations of the basic linguistic message.[5] Speech recognition must extract the linguistic message within the speech by removing the across-speaker variability and the within-speaker variability, while the speaker recognition must extract the distinctive speaker discrimination message.[4,5] Then, the question is, should bandpass liftering be used in speaker recognition? And if it were, should they be the same?

In this paper, we propose a new bandpass liftering window that is more suitable to speaker recognition. The paper was organized as follow: Section 2 introduces the new bandpass liftering window for speaker recognition. Section 3 presents the experimental results of the new liftering method in speech and speaker recognition. Summary is drawn in the final section.

2. Bandpass liftering windows

In [1], Juang analyzed the ratio of the variances of the simulated fixed filter data to the variances of the mixed data, and found that the variances of the higher cepstral terms are relatively large compared with those from the general overall statistical model of the mixed data. The increase in variance ratio with increasing coefficient index indicates the diminishing discriminating power of the higher terms. Thus, the variability of higher terms are inherent artifacts of the analysis procedure, and hence are less desirable in spectral similarity comparisons than the lower terms. Juang also analyzed the origin of the variability of the low terms of cepstrum[1] The variability of the low terms is primarily due to variations in transmission, speaker characteristics, and vocal efforts, etc., of the speech. Different

[1] This work is supported by the key project of National Nature Science Foundation of China grants 69635052 and the postdoctoral foundation.

transmission channels usually have different frequency responses and the differences generally affect the low cepstral terms much more than the high terms. In his speech recognition experiment, Juang applied a type of cepstral liftering window to suppress the undesirable variations presented in the higher and lower LPC cepstral coefficients.[1] The liftering window is:

$$w_1(k) = \begin{cases} 1 + h\sin(k\pi / L) & k = 1,2,...L, h = L/2 \\ 0 & otherwise \end{cases} \quad (1)$$

where h, k and L determine shape of the liftering window. We refer it as "Raised-Sine (RS)" window in the paper.

What kind of liftering windows should be used for speaker recognition? Here we use the Fisher discrimination criteria to calculate the Fisher ratio (F-ratio) of different components of MFCC in both speech and speaker recognition tasks. The F-ratio is defined as

$$F\text{-}ratio(i) = inter\text{-}class(i) / intra\text{-}class\ (i) \quad (2)$$

Where i is the component index of MFCC, the inter-class distance and intra-class distance are based on Euclidean distance definition. For speech recognition, the inter-class distance is the character vector distance of different words and the intra-class distance is mean variance of the character vector for the same word but from different speakers and different speaking. For speaker recognition, the inter-class distance is the character vector distance of the different speakers, and the intra-class distance is the mean variance of character vector from the different speaking of the same speaker. A higher F-ratio indicates better resolution capability. The F-ratio of MFCC terms for speech and speaker recognition calculated from TI46 database is shown in Fig. 1.

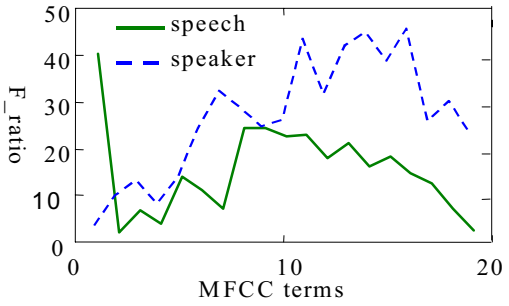

Fig. 1 F-ratio of components of MFCC in speech (solid line) and speaker (dashed line) recognition

For speech recognition, the F-ratio of middle terms is larger than that of the low and the high terms, and the F-ratio contour is very similar to the shape of RS liftering window function. For speaker recognition, the F-ratio of MFCC terms tend to increase with the increase of the MFCC index, which indicates that the high MFCC terms can separate speaker better than the low MFCC terms. Thus, according to the shape of F-ratio contour for speaker recognition, we define a new liftering window

$$w_2(k) = \begin{cases} 1 + h\sin(k\pi / L) & k = 1,2,...L/2, h = L/4 \\ 0 & otherwise \end{cases} \quad (3)$$

The shape of the window is like quarterly sinusoid function. We refer it as "Half Raised-Sine (HRS)" window in the paper. The shape of the RS and HRS liftering windows are drawn in Fig. 2

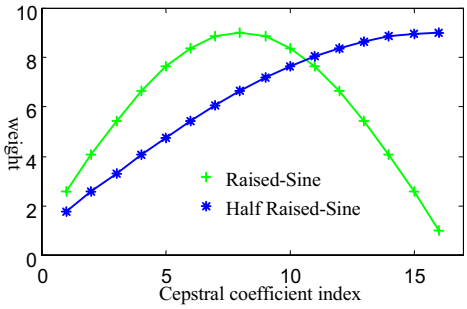

Fig. 2 Raised-Sine and Half Raised-Sine liftering windows

The effect of the liftering process can be visualized by transforming the windowed MFCC vector back to logarithmic spectrum domain. Fig.3 shows a series of liftered log spectra from speech "two" of a man, implied before and after RS and HRS window liftering. Through RS window liftering, the average level, as well as the slow and the fast variations with frequency of spectrum are de-emphasized, leaving components which change with that frequency. This process emphasizes both spectral peaks and valleys and smoothes the spectral contours. The peaks of the spectral contours represent the formants of speech and are important in characterizing the sound, especially in speech recognition task. Through the HRS window liftering, the average level and the slow variations with frequency of spectrum are de-emphasized as in RS liftering, while the middle and fast variations of spectrum are emphasized. Thus, the fine ripple fluctuated around the spectral contour, as well as the spectral peaks and valleys, are enhanced. The fine ripples may result from the detail information of the speech and numerical artifacts.[1] Note the part of log spectrum indicated by arrows. The ripples in the original spectrum were smoothed after RS liftering window processing, but were enhanced by the HRS liftering window.

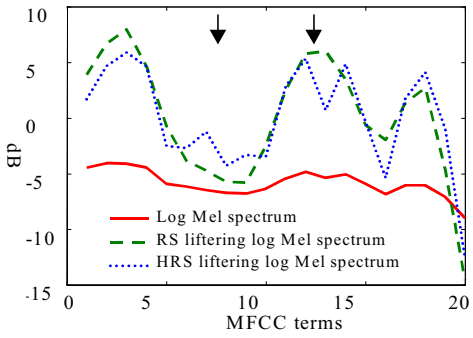

Fig. 3 Log Mel-frequency spectrum (solid line), RS window liftering log spectrum, L=12 (dashed line) and HRS window liftering log spectrum, L= 32 (dotted line).

Fig. 4 lists the log Mel-frequency spectrogram of an utterance "two" of a male before and after bandpass liftering process. The x-axis is frame number index (time) and the y-axis is the index

of Mel-frequency filter. The peaks and valleys are enhanced after both liftering processes. The RS window liftering extracts contour of the spectrogram (the thick stripes), while the HRS liftering window extracts the detail spectra structures (the thin stripes).

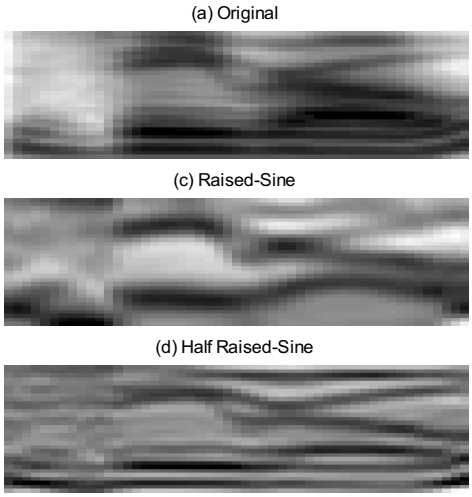

Fig. 4 Log Mel-frequency spectrogram of an utterance "two". (a) no liftering; (b) RS window liftering, L=12; (c) HRS windows liftering, L= 32.

Auditory physiology and psychology research manifests that the spectral structures, including the peaks of the spectral contour and the symmetry of the peaks are importance to auditory perception. The coarse scale analysis of the Shamma's primary auditory extracts the spectral contour information. The RS liftering windows can be regarded as an engineering implementation of spectral contour extraction. Lateral inhibition is a prevail mechanism in the perception path, which indicates that the spiking of a neuron is inhibited by the exciting of the neighbor neuron. In auditory, the lateral inhibition is to sharpen the edge of auditory spectra to code more robust information.

3. Experimental Results

3.1 The speech database

The speech database used in experiments was the ten isolated digits of standard speech database TI46. Each digit was spoken by sixteen speakers (eight females and eight males). The data of these digits were divided into two sets (training and testing). In training set, each digit was repeated 10 times by each speaker in one session. In the testing set, each digit was repeated 16 times by each speaker in eight different sessions. In each session, two utterances were recorded.

The training set was used for training and the recognition was performed in different Signal-to-Noise Ratio (SNR). A zero-mean white Gaussian noise was added to the test utterance at each specified SNR. The SNR is defined as

$$SNR = 10\log_{10}(P_S / P_N) \quad (4)$$

where P_S and P_N represent the average power of speech signal and noise respectively.

MFCC was used as feature vector. The MFCC are derived directly from the FFT power spectrum after pre-emphasis of speech. The power spectrum are weighted by a triangle filter shape and then summed. The filters have a half-bandwidth of 100Hz up to center frequency 1KHz and a bandwidth of 1.149 times the center frequency above 1KHz.

We used the DTW as the classifier of the recognition. During training phase, we use DTW to obtain one template for each digit from training sets. Thus, what we do is speaker-dependent speech recognition and text-dependent speaker recognition.

3.2 Speech recognition experiment

The speech recognition results using MFCC, RS liftering window and HRS liftering window are drawn in Fig. 5. The error rate of speech recognition is the average error rate of 16 people. The speech recognition error rates are sensitive to the change of speech quality. Compared with the MFCC, the speech recognition error rates reduced after both types of liftering process in all SNR conditions. The total error rate after the RS window liftering (L=16) decreases 46% compared with MFCC. The HRS liftering window performs almost as good as RS liftering window in different SNR and order. Increasing MFCC order from 12 to 16 does not have significant effects on the performance for both liftering window.

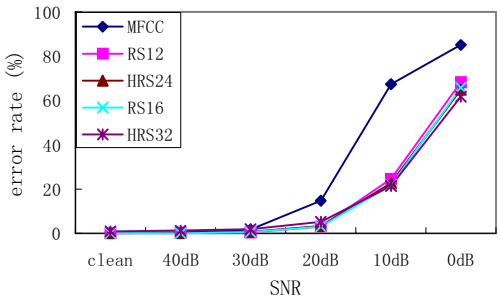

Fig. 5 Speech recognition error rates with different liftering in cepstral domain. The RS12 and RS16 indicates the RS window liftering with L=12 and L=16, respectively. The HRS24 and HRS32 indicates the HRS window liftering with L=24 and L=32, respectively.

3.3 Speaker recognition experiment

The speaker recognition results of MFCC before and after different liftering windows are drawn in Fig. 6. The error rate of speaker recognition is the average error rate of digit from 0 to 9. Similarly, both two bandpass liftering windows have better performance than that of MFCC. Especially, the HRS liftering window (L=32) reduces error rate 43% compared with that of without liftering. Compared with RS liftering window, the HRS liftering window is more robust in noise environment. The total error rates decrease 12.5% and 9% in 12 and 16 orders of MFCC, respectively. Unlike in speech recognition, With the

increment of MFCC order from 12 to 16, the recognition rate increase 7% and 12% for RS and HRS liftering window respectively.

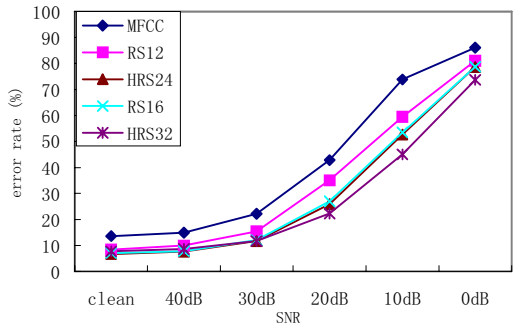

Fig. 6 Speaker recognition error rates with different liftering in cepstral domain. The legends are the same as in Fig. 5.

The low cepstral terms reflect slow variations of the spectral (spectral contour) and the high cepstral terms reflect the quickly varying ripple across the log spectral (fine structure of spectrum). As shown in Fig.3, Juang's RS liftering window emphasizes the spectral peaks and valleys and suppresses the fine detail spectrum, while the HRS liftering window enhances the fine ripple of the spectrum, as well as the spectral peaks and valleys. RS liftering window makes better performance in speech and speaker recognition indicates that enhancing the contrast of spectral peaks and valleys is helpful to both tasks. This signifies that the spectral contours (the formants) are important for both recognition tasks. RS liftering window and HRS liftering window perform almost equal in speech recognition indicates that the fine spectral ripples are of little use in speech recognition, at least under the condition of DTW classifier. Increment of MFCC orders does not increase recognition rate supports the conclusion from another hand.[2,3] For speaker recognition, the HRS liftering window outperforms RS window indicates that the fine spectrum ripples are help to speaker recognition.[8] The increasing recognition rate with more MFCC orders supports the conclusion from another hand because more MFCC orders denote more spectral details. The fine structure of spectrum contain speaker information is not new. It is well know that the residue in LPC contains useful speaker information.[2,8] The high order coefficients of LPC denote the fine structure of spectral.

4. Summary

In conclusion, we propose a new bandpass liftering process for speaker recognition. The error rate was reduced 43% with the new HRS liftering window than that obtained without the liftering process. Compared with Juang's RS liftering window, the HRS liftering window is more robust in noise environment, and it reduces the error rate 9%. The experiments on two liftering windows also indicate that the contours of the spectra are useful for both speech and speaker recognition, while the fine structures of spectrum denotes only the speaker-discriminating information. If we try to retain the spectrum details of speech and emphasize some components of features relating to speaker identity, we could obtain better results in speaker recognition.

References

1. B. H. Juang, L. R. Rabiner and J.G. Wilpon, "On the Use of Bandpass Liftering in Speech Recognition", IEEE Transactions on ASSP, Vol. 35, No.7, pp. 947-953, 1987.
2. X. Yang and H. Chi, "Digital Processing of Speech Signal", Press of Electric Industry, 1995.
3. L. R. Rabiner and B.H. Juang, "Fundamentals of Speech Recognition", 1993, second edition.
4. J. P. Campbell, "Speaker Recognition: a Tutorial", Proc. IEEE, vol.85, No.9, pp. 1437-1462, 1997.
5. G. R. Doddington, "Speaker Recognition- Identifying people by their voice," Proc. IEEE, vol. 73, No. 11, pp. 1651-1664, 1985.
6. J. O. Pickles, "An Introduction to the Physiology of Hearing", Academic Press, 1988.
7. S. A. Shamma, "Speech Processing in the Auditory System II: Lateral Inhibition and Central Processing of Speech Evoked Activity in the Auditory Nerve", J. Acoustic. Soc. Am., vol.78, pp. 1622-1632, 1985.
8. D. A. Reynolds, "Experimental evaluation of features for robust speaker identification ", IEEE Trans. Speech and Audio Proc., vol. 2, no. 4, pp. 639-643, 1994.