화자 불변 특징추출을 위한 스펙트럼 정규화

오 광 철°, 이 황 수
한국과학기술원, 정보 및 통신 공학과

Spectral Normalization for Speaker-Invariant Feature Extraction

Kwang-Cheol Oh and Hwang-Soo Lee
Dept. of Information and Communication Eng., KAIST

Abstract

We present a new method to normalize spectral variations of different speakers based on physiological studies of hearing. The proposed method uses the cochlear frequency map to warp the input speech spectra by interpolation or decimation. Using this normalization method, we can obtain much improved recognition results for speaker independent speech recognition.

1 Introduction

Many researchers have tackled problems on speaker variability by using multiple representations for each model or by making a model out of data gathered from a number of talkers. These approaches sometimes perform well, however, they tend to fail in recognizing some people whose spectra are different from those of reference models.

In this paper, we propose a new spectral normalization technique to represent various speech signals from a lot of different speakers consistently. In the spectral normalization, a cochlear frequency map is allied with interpolation or decimation[1][2]. The cochlear frequency map reveals a relationship between characteristic frequencies and cochlear longitudinal locations in the inner ear, and we adopt the cochlear map derived by M.C. Liberman for the cat[3]. The digital frequency can be extended or shrunken by decimation or interpolation, respectively. And voice signals have the tendency that formants with a lower fundamental frequency(F0) are lower than those with a higher F0. Hence we can use interpolation or decimation to normalize individual speech spectrum with a factor obtained from the cochlear frequency map.

2 Spectral Normalization

One can increase or decrease the sampling rate of discrete signals by interpolation or decimation, respectively. As a result of the interpolation or decimation, we can obtain the warped frequency spectrum of the input signal. Let the spectrum of an input signal be $X(e^{jw})$, and the interpolated and the decimated spectra be $Y_I(e^{jw})$ and $Y_D(e^{jw})$, respectively. Then, they are related as follows.

$$Y_I(e^{jw}) = X(e^{jwL}), \tag{1}$$

$$Y_D(e^{jw}) = \frac{1}{M}X(e^{jw/M}), \tag{2}$$

where L and M are the interpolation and the decimation factors, respectively. These equations imply that the spectrum of the decimated signal is an extended version of the original spectrum and that of the interpolated signal is a shrunken one. Therefore, we can align different individual spectra by shrinking or extending each spectrum by interpolation or decimation, respectively.

In order to match the different spectra, first we must decide how much we should extend or shrink the spectrum. In other words, we have to choose the interpolation factor L or the decimation factor M. These factors should be selected carefully so that the spectra of voiced speech from various speakers could be properly normalized. In this paper, we use the perceptual knowledge obtained from the physiological and psychological studies of hearing in order to choose those factors.

A mammalian cochlea in the inner ear acts as a mechanical frequency analyzer, maximally sensitive to high frequencies in the basal turn and sensitive to lower frequencies at positions closer to the apex. Each location on the cochlea is equally important in listening and responses actively for a frequency called the characteristic frequency. The cochlear frequency map, which shows the relationship between the characteristic frequency f and the normalized cochlear location x from the apex, derived by M.C. Liberman[3] is expressed by the following equations:

$$f = 456 \times (10^{x \times 2.1} - 0.80) \qquad (3)$$

or

$$x = \frac{1}{2.1} \times \log_{10}(\frac{f}{456} + 0.80). \qquad (4)$$

In human perception, one can hardly identify a specific frequency absolutely but is just apt to compare its relative height, i.e. notices only the dissimilarities among several tones. We assume here that the gaps among the locations corresponding to $F0$ and formants are almost equal for each speaker. With this assumption and Eq. (4), we can decide L or M in order to normalize the spectrum of new speaker's voice having the fundamental frequency $F0_{new}$ to that of the reference spectrum with $F0_{ref}$. We use the fractional interpolation or decimation to normalize individual voices dedicately, and fix the cutoff frequency of the low-pass filter to 1/50 for easy implementation. If $F0_{new}$ is larger than $F0_{ref}$ then we fix L to 50 and set

$$M = 50 \times \frac{F0_{ref}/456 + 0.8}{F0_{new}/456 + 0.8}. \qquad (5)$$

On the contrary, if $F0_{new}$ is smaller than $F0_{ref}$, then fix M to 50 and set

$$L = 50 \times \frac{F0_{new}/456 + 0.8}{F0_{ref}/456 + 0.8}. \qquad (6)$$

Figure 1 illustrates the proposed spectral normalization method. Speech signal enters the pitch extractor, which extract pitch and voiced information. The voiced information controls switches to normalize voiced region only. $L\&M$ determinator chooses the interpolation and the decimation factors from the pitch information. The sampling-rate converter translates speech signal to normalized signal. Conventional feature extractor can be to extract feature vectors.

3 Experimental Results

Voices from different individuals reveal different formant structures which depend on some physical characteristics(for example, vocal tract length) and other emotional factors(speaking habits or accents etc.). Much of variations are mainly due to different sexuality such that with female $F0$ and formant frequencies of female speakers are generally higher than their male counterparts. Spectral differences among two men and two women are shown in Fig. 2.(a). These spectra are obtained by mel-scaled cepstrum analysis of order 16 for the vowel segment /i/. The second and third formant frequencies of men are lower than their counterparts, furthermore the third formant frequency of the man with $F0$ = 114 Hz is similar to those of the second formants of women. After applying the proposed spectral normalization method, the formants of men are shifted upward and those of women downward, consequently the formants of all four speakers are aligned as shown in Fig 2. (b).

Effects of the normalization are measured by experiments on speaker-independent digit recognition including 8 different voiced sounds /a, i , o, u, y∧, yu, m, l/. Speech data are extracted from 40 speakers, half of them are male speakers, and each talker utters every digit four times. These utterances are low-pass-filtered up to 7 kHz, then sampled at 16 kHz with 16-bit resolution. The proposed normalization technique is applied to these sampled speech. Fundamental frequencies are extracted by an autocorrelation pitch detector[4] and then manually corrected in advance. We obtained the LPC coefficients from the data by the autocorrelation method with a 30 ms Hamming window at a frame rate of 100 Hz. The mel-scaled cepstrum is extracted by applying the bilinear transformation[5].

To evaluate the performance of the proposed method, we used a DTW-based speech recognizer[6], and its reference pattern for each digit is selected from a pronunciation of one male speaker.

Data from 19 men and 20 women are tested by varying the analysis orders from 11 to 22, and the results shown in Fig. 3 demonstrate the performance improvement of the spectral normalization. Although the reference patterns are provided by a male speaker, recognition accuracy is dramatically increased for female's utterances. The performance elevation for the male talkers, however, is relatively lower than for the female. It seems that even if we can alter the positions of formants appropriately by applying the proposed method, the details of the spectrum is scarcely normalized as shown in Fig 2. This figure also describes the variations of recognition rates versus cepstrum orders. With the spectral normalization, the recognition performance is more rapidly improved and is settled down on the lower cepstrum order than that without the normalization.

4 Conclusions

We propose a new spectral normalization method, which warps the digital frequency by fractional interpolation or decimation whose factor is selected carefully from the cochlear frequency map by using the instantaneous pitch information. By applying the proposed method, the formant structures of different speakers are aligned together irrespective of the speaker's sexuality. In the experiments of speaker-independent speech recognition, the normalization surprisingly improves the recognition accuracy for the speakers different from the reference speaker in sex.

References

[1] Proakis, John G. and Manolakis, Dimitris G. : 'Introduction to Digital Signal Processing', Macmillan Publishing Company , 1988.

[2] Crochiere, Ronald E. and Rabiner, Lawrence R. : 'Multirate Digital Signal Processing', Prentice-Hall Inc. , 1983.

[3] Liberman, M. C.: 'The cochlear frequency map for the cat: Labeling auditory-nerve fibers of known characteristic frequency', *J. Acoust. Soc. Am.*, 1982, **72**, (5), pp. 1441-1449.

[4] Krubsack, D. A., and Niederjohn, R. J.: 'An autocorrelation pitch detector and voicing decision with confidence measures developed for noise-corrupted speech', *IEEE Trans.*, 1991, **ASSP-39**, (2), pp. 319-329.

[5] Oppenheim, A. V., Johnson, D. H.: 'Discrete representation of signals', *IEEE proc.*, 1972, **60**, (6), pp. 681-691.

[6] Sakoe, H., and Chiba, S.: 'Dynamic programming algorithm optimization for spoken word recognition', *IEEE Trans.*, 1978, **ASSP-26**, (1), pp. 43-49.

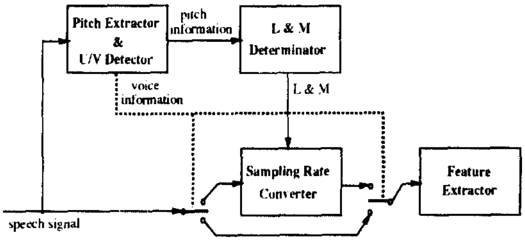

Fig. 1: Block diagram for the spectral normalization.

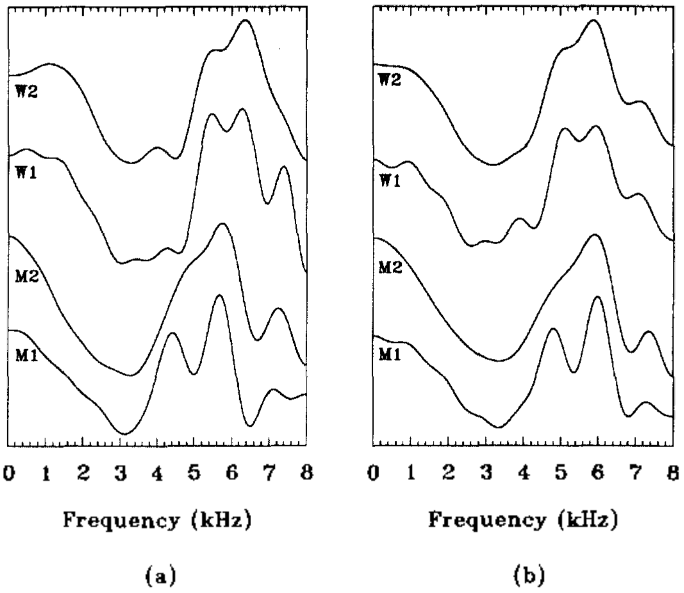

Fig. 2: Series of log spectra for the vowel segment /i/. Spectra in (a) are obtained from the original speech data without normalization and spectra in (b) from the data with normalization

M1 : man with 114 Hz pitch W1 : woman with 222 Hz pitch
M2 : man with 142 Hz pitch W2 : woman with 266 Hz pitch

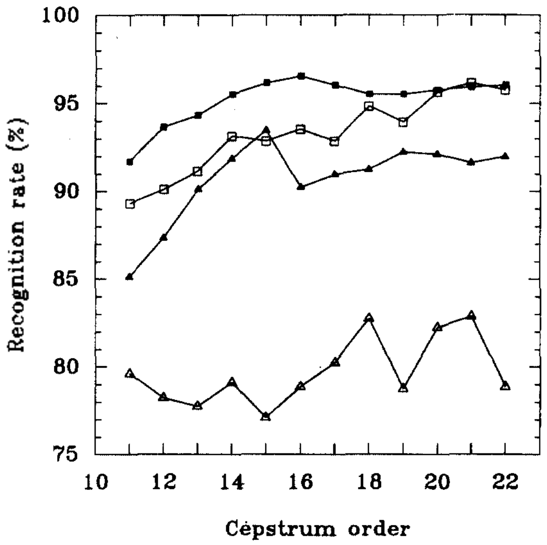

Fig. 3: Speaker independent recognition rates with and without normalization corresponding to the various cepstrum orders.

□——□ recog. results for male speakers without norm.
△——△ recog. results for female speakers without norm.
■——■ recog. results for male speakers with norm.
▲——▲ recog. results for female speakers with norm.