SPEAKER NORMALIZED SPECTRAL SUBBAND PARAMETERS FOR NOISE ROBUST SPEECH RECOGNITION

*Satoru Tsuge** *Toshiaki Fukada* *Harald Singer*

ATR Interpreting Telecommunications Research Laboratories
2–2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288 Japan
Tel: +81 774 95 1301, FAX: +81 774 95 1308, E-mail: stsuge@itl.atr.co.jp

ABSTRACT

This paper proposes speaker normalized spectral subband centroids (SSCs) as supplementary features in noise environment speech recognition. SSCs are computed as frequency centroids for each subband from the power spectrum of the speech signal. Since the conventional SSCs depend on formant frequencies of a speaker, we introduce a speaker normalization technique into SSC computation to reduce the speaker variability. Experimental results on spontaneous speech recognition show that the speaker normalized SSCs are more useful as supplementary features for improving the recognition performance than the conventional SSCs.

1. INTRODUCTION

All speech recognition systems include a signal processing front-end that converts a speech waveform into feature parameters. In real world applications, we often encounter situations in which mismatches between training and testing conditions exist (e.g., noise, speaker, or channel). In such cases, there is a dramatic degradation in the recognition performance. Therefore, the front-end is required to extract robust feature parameters from the speech signal that are relatively insensitive to these mismatches.

To this end, significant efforts have been made in research to compensate condition mismatches [1][2][3]. These approaches, however, usually require a knowledge or estimation of the properties of the current existing noise in advance (e.g., signal-to-noise ratio and/or noise spectrum). Another approach to cope with condition mismatches is to investigate robust features.

Recently, spectral subband centroids (SSCs) have been proposed as such features [4]. SSCs are computed as frequency centroids for each subband from the power spectrum of the speech signal. They can be obtained reliably even under noisy conditions, since SSCs roughly capture spectral peaks (such as formants) whose positions are almost unchanged in noisy environments.

*On leave from Tokushima University, Japan.

The effectiveness of SSCs is described in [4]. However, the task is very simple; speaker-dependent alphabet recognition in a clean environment. From now on, therefore, the effectiveness of SSCs has to be investigated in noisy environments. Moreover, because we consider that SSCs do not help very much in speaker-independent tasks, i.e., from the fact SSCs are highly dependent on formant peaks, in this paper, we incorporate a speaker normalization technique into SSC computation to reduce the speaker variability.

In the following section, we introduce spectral subband centroids. In Section 3, a method of combining SSCs with a speaker normalization technique is described. Section 4 shows recognition results on a Japanese spontaneous speech database in noisy environments.

2. SPECTRAL SUBBAND CENTROIDS IN A NOISY ENVIRONMENT

2.1. Spectral subband centroids [4]

Let us assume that the frequency band $[0,\ F_s/2]$(F_s is the sampling frequency) is divided into M disjoint subbands and that the shape of each subband filter is rectangular. Let the lower and higher edges of the m-th subband be l_m and h_m, respectively. Then, l_1=0, h_M=$F_s/2$, and $l_{m+1} = h_m = m * F_s/(2*M)$, for $m = 1, 2, \ldots, M-1$. The spectral subband centroid C_m for the m-th spectral subband is defined as follows:

$$C_m = \frac{\int_{l_m}^{h_m} f P^{\gamma}(f) df}{\int_{l_m}^{h_m} P^{\gamma}(f) df}, \qquad (1)$$

where f is frequency and $P(f)$ is the power spectrum. γ is a constant controlling the dynamic range of the power spectrum. By setting $\gamma < 1$, the dynamic range of the power spectrum can be reduced. The frequency band can be divided uniformly on the Hertz (Hz) scale or on the Mel (or Bark) scale. A smooth power spectrum (i.e., spectral envelope) can be used for the cen-

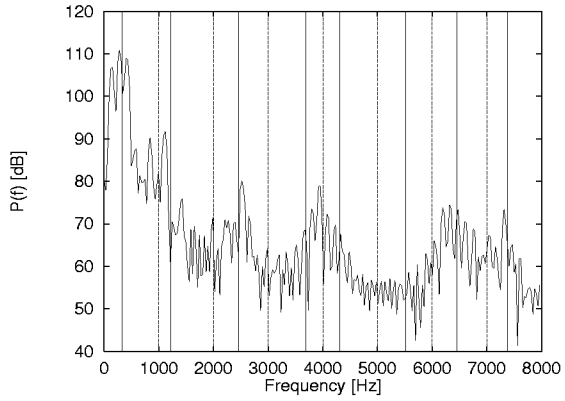


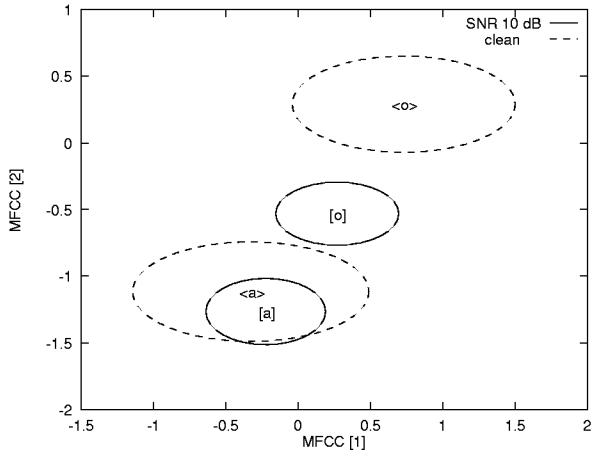
Figure 1: SSCs obtained from the FFT power spectrum of Japanese vowel /u/ ($M = 8$, $\gamma = 0.5$). The subband boundaries are shown by the vertical dotted lines, and the centroids by the solid lines.

troid computation. In this paper, we compute SSCs from a uniformly-divided unsmoothed (FFT) power spectrum on the Hz scale.

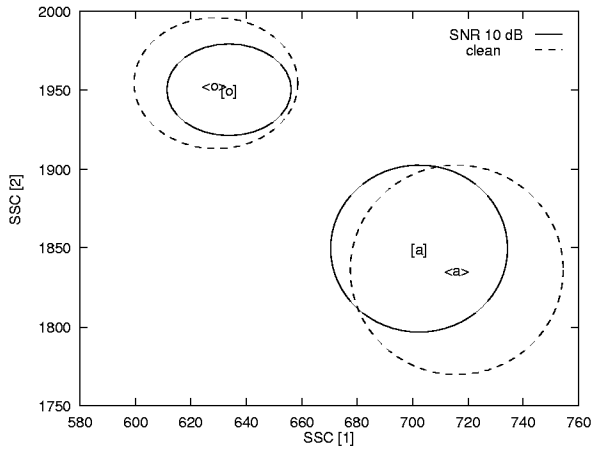
2.2. Analysis example

As an illustration, the FFT power spectrum for a Japanese vowel /u/ sampled at 16 kHz for clean speech are shown in Fig. 1. In this figure, spectral subband centroids with $\gamma = 0.5$, which are computed by eight uniformly-divided subbands (dotted lines), are shown by the solid lines. We can see that several SSCs (e.g., C_1, C_2, and C_3) are located around formant frequencies. That SSCs roughly capture formant frequencies in many cases has also been confirmed on other data.

Figure 2 illustrates the effectiveness of SSCs as robust features in a noisy environment. Figure 2(a) shows the distributions of the first and second order MFCCs for Japanese vowels /a/ and /o/. Figure 2(b) shows the distributions of the first and second order SSCs for the same vowels. In both figures, the dotted ovals indicate the areas of $\mu \pm 0.5\sigma$, where μ is the mean and σ is the standard deviation, in a clean environment. The solid ovals indicate the areas in a noisy environment (SNR=10dB). We can see from these figures that SSCs can be stably obtained compared to MFCCs even when the condition mismatches are observed.



(a) MFCCs



(b) SSCs

Figure 2: Distributions for Japanese vowels /a/ and /o/ for a single male speaker. The upper figure shows the distributions of the first and second order MFCCs for these vowels. The lower figure shows the distributions of the first and second order SSCs for the same vowels. <> and [ ] indicate the means in clean and noisy (SNR=10dB) environments, respectively. The solid ovals and the dotted ovals indicate areas of $\mu \pm 0.5\sigma$.

3. SSC WITH A SPEAKER NORMALIZATION TECHNIQUE

In the previous section, we showed that SSCs roughly capture formant frequencies of spectra. Accordingly, we can expect SSCs to provide useful information under noisy conditions, since formant frequencies are not usually changed by condition mismatches. However, we do consider that the distributions of SSCs computed from large amounts of speakers will be highly overlapped between different phones, since formant peaks are speaker dependent. We, therefore, introduce a speaker normalization technique (e.g., [5], [6]) into the SSC computation to reduce the speaker variability. In this paper, we use a speaker normalization technique based on frequency warping.

Warping factor α, which is used as a coefficient for

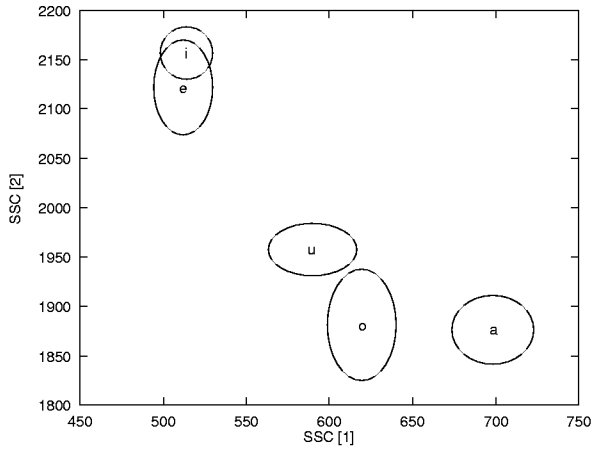
spectral warping to normalize the speaker's vocal tract length, can be computed as

$$\alpha = \frac{\sum_{p \in \mathbf{P}} F_{c,p}}{\sum_{p \in \mathbf{P}} F_{s,p}}, \quad (2)$$

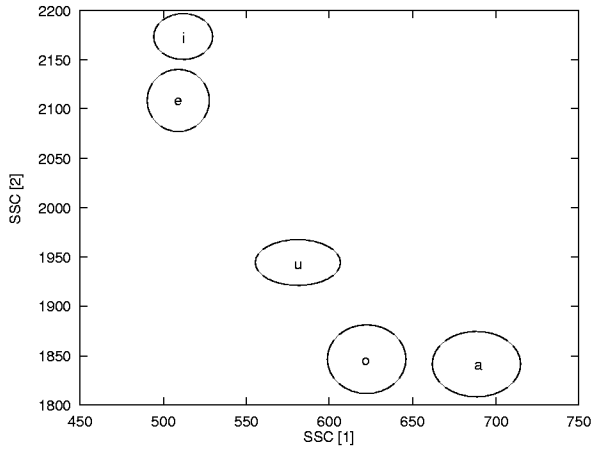
where $\mathbf{P}$ denotes the set of vowels {/a/, /i/, /u/, /e/, /o/}. $F_{c,p}$ and $F_{s,p}$ are the averages of the second formant frequencies for the vowel p for the training corpus c and the speaker s, respectively. By using warping factor α, speaker normalized SSCs, $\bar{C}_m$, can be computed as,

$$\bar{C}_m = \frac{\int_{l_m}^{h_m} f \cdot P^{\gamma}(\alpha^{-1} f) df}{\int_{l_m}^{h_m} P^{\gamma}(\alpha^{-1} f) df}. \quad (3)$$

Figure 3 illustrates the effectiveness of the speaker normalized SSCs. Figures 3(a) and (b) show the conventional (i.e., without speaker normalization) and the proposed (i.e., with speaker normalization) SSCs, respectively. The distributions ($\mu \pm 0.5\sigma$) of the first and second order SSCs for the five Japanese vowels are shown as ovals. These distributions were obtained from 230 speakers. We can see that the overlaps between vowels are successfully reduced by introducing the speaker normalization technique (e.g., /i/ and /e/). From these figures, we can expect speaker normalized SSCs to give a better recognition performance than the conventional SSCs.



(a) Conventional SSCs



(b) Speaker normalized SSCs

Figure 3: Distributions for five Japanese vowels for 230 speakers ($\mu \pm 0.5\sigma$). The upper figure shows the distributions without speaker normalization technique. The lower figure shows the distributions with speaker normalization technique.

4. EXPERIMENTS

The speaker normalized SSCs were evaluated through continuous word recognition on a Japanese spontaneous speech database [7].

4.1. Conditions

A total of 230 speaker (100 male and 130 female) dialogues sampled at 16 kHz were used for the training. For the open test set, 42 speaker (17 male and 25 female) dialogues were used. 12-dimensional MFCCs and log power, and their first and second derivatives (i.e., 39 dimensions in total), which were computed using a 20 msec window duration and a 10 msec frame period, were used as a conventional feature vector (**MFCC**). In addition, 6-dimensional SSCs and their first derivatives were used together with the conventional vector (i.e., 51 dimensions in total) (**MFCC+SSC**). In the SSC computation, the Nyquist frequency band (from 0 to 8000 Hz) was divided equally into six subbands (i.e., $M = 6$), and γ in Eq. (1) and Eq. (3) was set to 0.5. The speaker normalization technique described in Section 3 was applied to **MFCC** and to **MFCC+SSC** (referred to as **SN-MFCC** and **SN-MFCC+SN-SSC**, respectively). A warping factor was calculated for each speaker in Eq. (2) from clean speech. Formant frequencies were automatically obtained by using the commercially-available Waves+ package.

For these four feature vectors (i.e., **MFCC**, **MFCC+SSC**, **SN-MFCC**, and **SN-MFCC+SN-SSC**), shared-state context dependent HMMs with five Gaussian mixture components per state were trained from clean speech [8]. The total number of states was set to 800. We used a spontaneous speech recognizer using cross-word context constrained word graphs [9].

Table 1: Recognition results (word accuracy in %).

feature vector	SNR [dB]			
	10	15	20	30
MFCC	19.7	44.4	63.1	72.5
MFCC+SSC	30.1	51.2	65.0	71.0
improvement (%)	13.0	12.2	5.2	-5.5

Table 2: Recognition results with speaker normalization technique (word accuracy in %).

feature vector	SNR [dB]			
	10	15	20	30
SN-MFCC	20.0	47.6	66.5	73.7
SN-MFCC+SN-SSC	32.9	58.2	68.7	74.1
improvement (%)	16.1	20.3	6.5	1.5

The test vocabulary consisted of about 7,000 words, and the variable-length N-gram [10] was used for the language model.

Workstation noise was added to clean speech utterances at four kinds of signal-to-noise ratios (SNR=10, 15, 20, and 30 dB). Each SNR was measured by calculating the ratio of speech energy to noise energy at the utterance level.

4.2. Results

The recognition results for **MFCC** and **MFCC+SSC** are shown in Table 1. We can see from these results that SSCs are useful feature parameters especially for lower SNRs. Actually, we observed a relative improvement in the error rate by 13.0% at SNR=10dB. **MFCC+SSC**, however, degraded the recognition performance at SNR=30dB compared to **MFCC**.

The recognition results for **SN-MFCC** and **SN-MFCC+SN-SSC** are shown in Table 2. First, we can see from this table that SSCs help improve the recognition performance for all SNRs. Especially, we observed a significant improvement in the error rate by 20.3% at SNR=15dB. Furthermore, by comparing Tables 1 and 2, the improvement for each SNR in Table 2 is consistently larger than the corresponding one in Table 1. This implies that incorporating the speaker normalization technique into the conventional SSC computation is effective for improving the recognition performance.

5. CONCLUSIONS

In this paper, we have proposed speaker normalized spectral subband centroids (SSCs) as supplementary features in noise environment speech recognition. SSCs can be readily computed from the power spectrum without any knowledge of the current noise. To reduce the speaker variability, a speaker normalization technique was successfully incorporated into the conventional SSC computation. Experimental results on spontaneous speech recognition showed that the proposed SSCs give a consistently better performance under several noise conditions than the conventional SSCs. We could also confirm large improvements in the word accuracy (16.1% at SNR=10dB and 20.3% at SNR=15dB) by using the proposed SSCs together with the conventional feature parameters.

In our experiments, noisy speech was artificially generated by adding noise signals to clean speech. As it is widely known that formant frequencies change depending on the SNR (i.e., the Lombard effect), we plan to perform further experiments by using real noisy speech.

6. REFERENCES

[1] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. In *IEEE Trans. Acoustics, Speech and Signal Processing*, pages 113–120, April 1979.

[2] M. J. F. Gales and S. Young. An improved approach to the Hidden Markov model decomposition of speech and noise. In *Proc. ICASSP*, pages 233–236, 1992.

[3] H. Hermansky and N. Morgan. Rasta processing of speech. In *IEEE Trans. Speech and Audio Processing*, volume 2, pages 578–589, October 1994.

[4] K. Paliwal. Spectral subband centroids features for speech recognition. In *Proc. ICASSP*, pages 617–620, 1998.

[5] E. Eide and H. Gish. A parametric approach to vocal tract length normalization. In *Proc. ICASSP*, pages 346–348, 1996.

[6] L. Lee and R. C. Rose. Speaker normalization using efficient frequency warping procedures. In *Proc. ICASSP*, pages 353–356, 1996.

[7] A. Nakamura, S. Matsunaga, T. Shimizu, M. Tonomura, and Y. Sagisaka. Japanese speech databases for robust speech recognition. In *Proc. ICSLP*, pages 2199–2202, Philadelphia, 1996.

[8] M. Ostendorf and H. Singer. HMM topology design using maximum likelihood successive state splitting. *Computer Speech and Language*, 11(1):17–41, 1997.

[9] T. Shimizu, H. Yamamoto, S. Matsunaga, and Y. Sagisaka. Spontaneous dialogue speech recognition using cross-word context constrained word graphs. In *Proc. ICASSP*, pages 145–148, 1996.

[10] H. Masataki and Y. Sagisaka. Variable-order n-gram generation by word-class splitting and consecutive word grouping. In *Proc. ICASSP*, pages 188–191, Atlanta, 1996.