# Recognition of Noisy Speech Using Dynamic Spectral Subband Centroids

Jingdong Chen, *Member, IEEE*, Yiteng (Arden) Huang, *Member, IEEE*, Qi Li, *Member, IEEE*, and Kuldip K. Paliwal, *Fellow, IEEE*

***Abstract*—Despite their widespread popularity as front-end parameters for speech recognition, the cepstral coefficients derived from either linear prediction analysis or a filter-bank are found to be sensitive to additive noise. In this letter, we discuss the use of spectral subband centroids for robust speech recognition. We show that centroids, if properly selected, can achieve recognition performance comparable to that of the mel-frequency cepstral coefficients (MFCCs) in clean speech, while delivering better performance than MFCC in noisy environments. A procedure is proposed to construct the dynamic centroid feature vector that essentially embodies the transitional spectral information. We discuss some properties of the proposed dynamic features.**

***Index Terms*—Cepstrum, robust speech recognition, subband centroid.**

## I. Introduction

THE CEPSTRAL coefficients derived from either linear prediction (LP) analysis or a filter-bank are almost "standard" front-end features in currently available automatic speech recognition (ASR) systems. For example, the mel-frequency cepstral coefficients (MFCCs) are today being standardized for cellular phone systems [1]. Much evidence shows that the cepstral coefficients have served as very successful frontends for hidden Markov model (HMM) based speech recognition in the past decade. Many speech recognition systems based on these representations have achieved a very high level of accuracy in a clean speech environment.

Despite their de facto standardization as front-end features, the cepstral features are widely acknowledged not to cope well with noisy speech. To improve the robustness of frontends with respect to noise and distortion, there has been tremendous effort in searching for alternative features [2]–[4]. Observing that the higher amplitude portions (such as formants) of a spectrum are relatively less affected by noise, Paliwal proposed spectral subband centroids (SSC) as features [5]. If the short-time power spectrum of a speech signal is denoted by $P(t,\omega)$, where $t$ is the frame index and denotes the time dimension, and $\omega$ is radial frequency, the subband moment of order $p$ is defined as

$$M_i^p(t) = \int_0^\pi \omega^p w_i(\omega) P(t,\omega) d\omega \tag{1}$$

Manuscript received September 13, 2002; revised March 12, 2003. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mark Hasegawa-Johnson.

J. Chen, Y. Huang, and Q. Li are with Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974 USA (e-mail: jingdong@research.bell-labs.com).

K. K. Paliwal is with School of Microelectronic Engineering, Griffith University, Nathan 4111, Queensland, Australia.

Digital Object Identifier 10.1109/LSP.2003.821689

where $1 \le i \le Q$, $Q$ is the total number of frequency bands, and $w_i(\omega)$ is the frequency response of the $i$th bandpass filter. SSC, essentially the first-order normalized moment, is given by

$$C_i(t) = \frac{M_i^1(t)}{M_i^0(t)}. \tag{2}$$

In [5], the SSC representations were compared with the linear prediction cepstral coefficients (LPCCs) for an English e-set alphabet recognition task. It turned out that for an open test where the test condition is different from the training condition, only three SSCs yielded performance comparable to that of ten LPCCs, yet LPCCs delivered a better recognition rate than SSCs for a closed test where the training and test conditions are identical. This indicates the potential of the SSC features for robust speech recognition. SSC was further investigated in [6] where the speech signal is represented by SSC histogram-based cepstral coefficients. These new cepstral features yielded promising performance for speech recognition. Problems also discussed in [6] include what shape of bandpass filter $[w_i(\omega)]$ should be used, and how the subbands should be divided (linear scale, mel-scale, or Bark-scale?), etc. Other experiments using SSCs as supplementary features to the cepstral coefficients for speech recognition can be found in [7] and [8].

Although the SSC representation was experimentally used in speech recognition with certain success, it is important to further investigate its potential as an independent feature set. In this letter, we reexamine the SSC as an independent feature set for speech recognition. We address two issues that have not been covered by any of the previous studies: how many frequency bands should be used to achieve good recognition performance, and how should the dynamic SSC features be computed to augment the static SSC features?

## II. Recognition Performance Versus Number of Subbands

In representing a speech signal by MFCC coefficients, the short-time power spectrum is divided into about 20 frequency bands. The logarithmic filter-bank outputs are then converted into about ten cepstral coefficients. This has been proven to give the highest performance in practical speech recognition systems. A natural question then arises in extracting SSCs: how many frequency bands should be used? Too few bands are not sufficient to preserve the important information in a speech signal that is necessary for distinguishing among phonetic units. On the other hand, too many bands will make the extracted SSCs sensitive to harmonics and noise. Here we

1070-9908/04$20.00 © 2004 IEEE

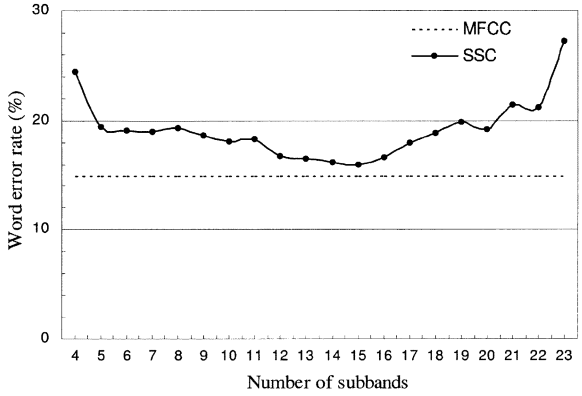


Fig. 1. Recognition performance of isolated spoken alphabet letters versus the number of subbands (without dynamic features). The short-time power spectrum is estimated using a 512-point fast Fourier transform. The frequency axis is then uniformly divided into $Q$ subbands with 50% overlapping.

perform a group of experiments to investigate the effect of the number of frequency bands on recognition performance.

The TI46 database is used for this task, which is an isolated spoken word database that was designed and collected by Texas Instruments (TI). The database consists of speech from 16 speakers, including eight males and eight females. The vocabulary consists of ten isolated digits from "zero" to "nine," 26 isolated English alphabet letters from "a" to "z," and ten isolated words ("enter," "erase," "go," "help," "no," "rubout," "repeat," "stop," "start," "yes."). There are 26 utterances of each word from each speaker: ten of these are designated as training, and the remaining 16 are designated as testing tokens. The speech signal is digitized at a sampling rate of 12.5 kHz. In our experiments, speech from all 16 speakers was used for recognition of the 26 letters. We will subsequently call these "words" since they are spoken in isolation.

The recognition system used is an HMM-based multispeaker isolated speech recognizer. The models are left-to-right with no skip-state transition. Eight states are used for each model. A mixture of eight multivariate Gaussian distributions with diagonal covariance matrices is used for each state to approximate its probability density function. Speech is analyzed every 10 ms with a frame width of 32 ms. Speech is preemphasized and Hamming windowed.

The word error rate as a function of the number of subbands is shown in Fig. 1. For this task, we have found that MFCC outperforms LPCC, and that 12 MFCCs ($C_0$ is neglected), which are derived from a filter-bank consisting of 24 mel-frequency triangular bandpass filters, give the highest recognition accuracy (without dynamic features). For the SSC case, it can been seen that the trend of the word error rate associated with the number of subbands is a saddle-shaped curve. In other words, as the number of subbands increases, the error rate first decreases and then increases. The lowest error rate is obtained using about 15 bands, which is almost as good as that of the MFCC features. We observe that for a large range, say between 10 and 20, the SSC features can yield performance comparable to that of MFCCs.

TABLE I
WORD ERROR RATES FOR THE MFCC AND SSC FRONTENDS. S, D, A STAND FOR STATIC, DELTA, AND ACCELERATION COEFFICIENTS, RESPECTIVELY

| | S (12 coefficients) | S+D (24 coefficients) | S+D+A (36 coefficients) |
|---|---|---|---|
| MFCC | 14.9% | 8.0% | 6.8% |
| SSC | 16.7% | 12.6% | 11.1% |

## III. DYNAMIC SPECTRAL SUBBAND CENTROIDS

It has been widely observed that temporal processing of short-time speech parameters can significantly improve speech recognition [9]–[11]. Thanks to Furui's work [9], a simple yet effective method to determine the dynamic (delta and acceleration) cepstral features in the vicinity of a given feature vector is popularly used in existing systems

$$\mathbf{D}(n) = \mathbf{S}(n+\Delta) - \mathbf{S}(n-\Delta) \tag{3}$$

$$\mathbf{A}(n) = \mathbf{D}(n+\theta) - \mathbf{D}(n-\theta) \tag{4}$$

where $\mathbf{S}(n)$, $\mathbf{D}(n)$, and $\mathbf{A}(n)$ stand for the static, delta, and acceleration feature vectors at time $n$, respectively.

Table I shows speech recognition results for MFCCs, and together with their dynamic features, where the delta coefficients are estimated from (3) with $\Delta = 2$ and the acceleration coefficients are computed according to (4) with $\theta = 2$. It can be seen that the use of delta MFCCs decreases the error rate from 14.9% to 8.0%. Acceleration coefficients further reduce the error rate to 6.8%. The corresponding error rate reductions, relative to the static MFCC baseline, are 46.3% and 54.4%, respectively. We likewise estimate the dynamic SSC features, and the recognition results are also presented in Table I. The delta SSCs decrease the error rate from 16.7% to 12.6%, and the acceleration features further reduce the error rate to 11.1%. Although they are able to improve the recognition performance, the computed dynamic SSCs are relatively ineffective when compared to the dynamic MFCC coefficients. This is mainly due to the fact that the SSC trajectory is rather flat. As a result, the differences among SSCs of neighboring frames are small. Thus, the dynamic SSC coefficient computed using (3) and (4) carries little information. The relative ineffectiveness of the dynamic SSC features raises a question: is there a better way to calculate the dynamic SSCs? In what follows, we describe a new procedure to compute dynamic SSCs.

The new dynamic SSC features are estimated through time-domain variation, which is represented by the differentiation of $C_i(t)$ with respect to time $t$, as shown at the bottom of the page. Since $P(t,\omega)$ usually does not have an analytic form, we approximate $(\partial P(t,\omega))/\partial t$ by a finite-order difference

$$\frac{\partial P(t,\omega)}{\partial t} \approx \Delta P(t,\omega) = \sum_{k=-o'}^{o} a_k P(t+k,\omega) \tag{6}$$

$$\frac{\partial C_i(t)}{\partial t} = \frac{1}{\left[\int_0^\pi w_i(\omega)P(t,\omega)\right]^2}\left[\int_0^\pi \omega w_i(\omega)\frac{\partial P(t,\omega)}{\partial t}d\omega \int_0^\pi w_i(\omega)P(t,\omega)d\omega - \int_0^\pi \omega w_i(\omega)P(t,\omega)d\omega \int_0^\pi w_i(\omega)\frac{\partial P(t,\omega)}{\partial t}d\omega\right] \tag{5}$$

TABLE II
WORD ERROR RATES FOR THE MFCC AND MODIFIED SSC FRONTENDS. S, D, A, AND L STAND FOR STATIC, DELTA, ACCELERATION, AND LONG-TERM DELTA COEFFICIENTS, RESPECTIVELY

| | S (12 coefficients) | S+D (24 coefficients) | S+D+A (36 coefficients) | S+D+L (36 coefficients) |
|---|---|---|---|---|
| MFCC | 14.9% | 8.0% | 6.8% | - |
| SSC | 16.7% | 9.5% | 8.3% | 6.9% |

where $o$ and $o'$ are the orders of the difference equation, and $a_k$'s are some real-valued coefficients. On substituting (6) into the right hand side of (5), it can be shown that

$$\frac{\partial C_i(t)}{\partial t} \approx \Delta C_i(t) = \sum_{k=-o'}^{o} b_k C_i(t+k) \tag{7}$$

where

$$b_k = \begin{cases} a_k \frac{M_i^0(t+k)}{M_i^0(t)}, & \text{for } k \neq 0 \\ a_0 - \sum_{j=-o'}^{o} a_j \frac{M_i^0(t+j)}{M_i^0(t)}, & \text{for } k = 0. \end{cases} \tag{8}$$

For brevity, the intermediate derivation is eliminated. Comparing (3) and (4) with (6), one can readily see the difference between the conventional dynamic features and the proposed dynamic features. In brief, the conventional dynamic features are computed through a difference equation with constant coefficients, while the proposed dynamic features are estimated through a difference equation with coefficients varying according to $M_i^0(t)$ and $M_i^0(t+k)$, which are essentially the $i$th subband energies at time $t$ and $(t+k)$.

A paramount issue in computing dynamic SSCs using (7) is to obtain the $a_k$ coefficients and eventually the $b_k$'s. It is noted that the dynamic coefficient computed from (7) is unbounded. This large dynamic range of the dynamic coefficients may cause the succeeding HMM training process to diverge. However, this problem can be circumvented by limiting the dynamic range of the $b_k$'s. In this letter, we suggest a heuristic set of $b_k$

$$b_k = \begin{cases} \frac{M_i^0(t+k)}{M_i^0(t+k)+M_i^0(t-k)}, & \text{for } k = 2 \\ \frac{-M_i^0(t+k)}{M_i^0(t+k)+M_i^0(t-k)}, & \text{for } k = -2 \\ 0, & \text{else} \end{cases} \tag{9}$$

to compute the delta SSCs. It can be easily shown that $-1 \leq b_k \leq 1$. The resulting dynamic SSC will, therefore, be in a reasonable range. Once the delta SSC features are estimated, (4) can be used to compute the acceleration coefficients.

The first three columns of Table II show the results using the proposed delta SSC features for the same recognition task as in the previous experiment. The new delta SSCs decrease the error rate from 16.7% to 9.5%. The corresponding error rate reduction is 43.1%. Using the acceleration coefficients further reduces the error rate to 8.3%.

Through investigation, we found that better recognition performance can be obtained by using long-term delta centroid features to replace the acceleration coefficients. We suggest using (7) to compute the long-term delta centroids with

$$b_k = \begin{cases} \frac{M_i^0(t+k)}{M_i^0(t+k)+M_i^0(t-k)}, & \text{for } k = 4 \\ \frac{-M_i^0(t+k)}{M_i^0(t+k)+M_i^0(t-k)}, & \text{for } k = -4 \\ 0, & \text{else.} \end{cases} \tag{10}$$

The result is also shown in the last column of Table II. As seen, using long-term delta coefficients, the recognition error rate is reduced to 6.9%, which is as good as the performance of MFCCs. Comparing Table II with Table I, one can see the effectiveness of the proposed dynamic SSC features.

## IV. ROBUSTNESS OF THE SSC FEATURES WITH RESPECT TO NOISE

This experiment compares the SSC and MFCC features for noisy speech recognition. The test bed and recognition system are the same as in the previous experiments. To control the SNR, we directly add some noise to speech signals in the test set while the training speech is kept clean. Two types of noise are used: computer-generated white Gaussian noise and babbling noise recorded from a New York City train station. The SSC feature vector consists of 36 coefficients, including 12 static, 12 delta [computed according to (7) and (9)], and 12 long-term [computed according to (7) and (10)] coefficients. The MFCC vector also consists of 36 coefficients: 12 static, 12 delta, and 12 acceleration coefficients [computed from (3) and (4)]. The recognition results are presented in Fig. 2.

The MFCC coefficients are derived from the filter-bank energies. They are sensitive to the level of noise. As the SNR decreases, recognition performance using MFCCs decreases dramatically. Compared with MFCCs, SSCs mainly represent spectral peaks of speech sounds. Since it has a flat spectrum, the Gaussian noise does not affect the peak positions of the speech spectrum very much. Hence the SSC features perform significantly better than the MFCC coefficients in Gaussian noise.

The babbling noise consists of some background speech signals. In such a condition, noise does not only alter the power level of a speech signal, but also dramatically affects the peak positions of the speech spectrum. In this case, both MFCCs and SSCs deviate from their counterparts in the clean condition. They produce similar recognition performance in this noise condition.

## V. AURORA EVALUATION FOR THE SDC-AURORA SPANISH DATABASE

The purpose of this test is to evaluate the SSC features for the connected digits recognition task on the SpeechDat-Car (SDC) Aurora Spanish database, which is used by the Aurora consortium to test the performance of the frontend with well-matched training and testing as well as its performance in mismatched conditions likely to be encountered in deployed ASR systems. It contains more than 160 speakers and 4914 recordings (files) from a close-talking microphone and a hands-free microphone installed in a car under various driving conditions. The files are classified into three categories: quiet, low noisy, and highly noisy conditions, depending on the driving condition. For the consistency of comparison of results among different researchers, the database is designed by the Aurora consortium

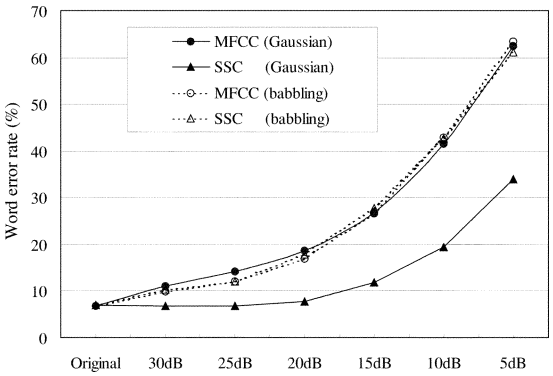


Fig. 2. Recognition performance of isolated spoken alphabet letters as a function of SNR in white Gaussian noise and babbling noise.

into three experimental configurations: well-matched (WM), medium-mismatch (MM), and high-mismatch (HM) experiments, respectively. See [12] for a detailed description of this database and how the WM, MM, and HM experiments are configured.

Experiments have been performed on this database at an 8-kHz sampling rate to compare the SSC and MFCC features. The recognizer used is a Bell Labs baseline recognition system. Here we use the head-body-tail (HBT) model, which assumes that the context-dependent digit models can be built by concatenating a left-context-dependent unit (head) with a context-independent unit (body) followed by a right-context-dependent unit (tail). In other words, each digit consists of 1 body, 12 heads, and 12 tails (representing all left/right contexts). In total, we have 276 units: 11(digits) $\times$ [1(body) + 12 (heads) + 12 (tails)] + 1 (silence). A three-state HMM is used to represent each head and tail and a four-state HMM for each body. Overall, this corresponds to a ten-state digit model for each variation of each spoken digit, with a total number of 837 states (including a one-state silence model). See [13] for more details.

The speech signal is analyzed every 10 ms with a 30-ms window. Each frame is represented by a feature vector consisting of 39 coefficients: 1 energy, 12 static features, 13 delta, and 13 acceleration (or long-term delta) coefficients. For MFCC and energy, the delta and acceleration coefficients are computed using (3) with $\Delta = 2$ and (4) with $\theta = 2$, respectively. The delta SSC features are calculated according to (7) and (9). The long-term delta SSC coefficients are calculated according to (7) and (10). The recognition results are shown in Table III. For the purpose of comparison, the recognition results using the long-term delta MFCCs and the acceleration SSCs are also presented, where the long-term delta MFCCs are computed from (3) with $\Delta = 4$ and the acceleration SSCs are computed from (4) with $\theta = 2$. Table III also shows the Aurora baseline performance using MFCCs (from [12]).

Apparently, our system yields a much better performance than the Aurora baseline. This is mainly due to a more accurate acoustical modeling in our system. Comparing the results of MFCC and SSC, one can see that in the well-matched case, MFCC slightly outperforms SSC. In the medium-mismatch situation, both SSC and MFCC deliver similar performance. However, in the high-mismatch condition, SSC is superior to MFCC. This demonstrates the robust nature of the SSC features.

TABLE III
WORD ACCURACIES OF SDC-AURORA SPANISH DATABASE (PERCENT). S, D, A, AND L STAND FOR STATIC, DELTA, ACCELERATION, AND LONG-TERM DELTA COEFFICIENTS, RESPECTIVELY

| | | WM | MM | HM |
|---|---|---|---|---|
| BL system | MFCC+D+A | 95.7 | 89.2 | 81.0 |
| | MFCC+D+L | 95.7 | 90.1 | 80.7 |
| | SSC+D+A | 95.0 | 88.1 | 78.4 |
| | SSC+D+L | 94.9 | 89.1 | 82.7 |
| Aurora baseline | MFCC | 86.9 | 73.7 | 42.2 |

## VI. CONCLUSION

In this letter, the spectral subband centroid features have been investigated for speech recognition. It is demonstrated that in clean speech conditions SSCs can produce performance comparable to that of MFCCs, provided that the number of subbands is properly selected. A procedure is proposed to compute dynamic SSC features, which can significantly augment the basic static SSCs for speech recognition. Experiments were performed to compare SSC with MFCCs for noisy speech recognition. The results showed that the centroids and the new dynamic SSC coefficients are more resilient to noise than the MFCC features.

## REFERENCES

[1] ETSU, "Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; front-end feature extraction algorithm; Compression algorithm,", ETSU Draft Std., ETSI ES 201 108 v0.08, 1999.
[2] J. W. Pitton, K. Wang, and B. H. Juang, "Time-frequency analysis and auditory modeling for automatic recognition of speech," *Proc. IEEE*, vol. 84, pp. 1199–1214, Sept. 1996.
[3] A. Potamianos and P. Maragos, "Time-frequency distributions for automatic speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 196–200, Mar. 2001.
[4] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 115–132, Jan. 1994.
[5] K. K. Paliwal, "Spectral subband centroid features for speech recognition," in *Proc. ICASSP*, vol. 2, 1998, pp. 617–620.
[6] B. Gajic and K. K. Paliwal, "Robust feature extraction using subband spectral centroid histograms," in *Proc. ICASSP*, vol. 1, 2001, pp. 61–64.
[7] S. Tsuge, T. Fukada, and H. Singer, "Speaker normalized spectral subband parameters for noise robust speech recognition," in *Proc. ICASSP*, vol. 1, 1999, pp. 285–288.
[8] D. Albesano, R. De Mori, R. Gemello, and F. Mana, "A study of the effect of adding new dimensions to trajectories in the acoustic space," in *Proc. EUROSPEECH*, vol. 4, 1999, pp. 1503–1506.
[9] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 254–272, Apr. 1981.
[10] F. K. Soong and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," in *Proc. ICASSP*, vol. 2, 1986, pp. 17.5.1–17.5.4.
[11] B. A. Hanson, T. H. Applebau, and J.-C. Junqua, "Spectral dynamics for speech recognition under adverse conditions," in *Automatic Speech and Speaker Recognition: Advanced Topics*, C. H. Lee, F. K. Soong, and K. K. Paliwal, Eds. Boston, MA: Kluwer, 1996.
[12] D. Macho, "Spanish SDC-Aurora database for ETSI STQ Aurora WI008 advanced DSR front-end evaluation: Description and baseline results,", STQ Aurora DSR Working Group Input Doc. AU/271/00, Nov. 2000.
[13] A. Afify, H. Juang, F. Korkmazsky, C.-H. Lee, Q. Li, O. Siohan, F. K. Soong, and A. Surendran, "Evaluating the Aurora connected digit recognition task—A Bell Labs approach," in *Proc. Eurospeech*, vol. 1, 2001, pp. 633–636.