# Robust Speech Recognition in Noisy Environments Based on Subband Spectral Centroid Histograms

Bojana Gajić, *Member, IEEE,* and Kuldip K. Paliwal, *Member, IEEE*

***Abstract*—We investigate how dominant-frequency information can be used in speech feature extraction to increase the robustness of automatic speech recognition against additive background noise. First, we review several earlier proposed auditory-based feature extraction methods and argue that the use of dominant-frequency information might be one of the major reasons for their improved noise robustness. Furthermore, we propose a new feature extraction method, which combines subband power information with dominant subband frequency information in a simple and computationally efficient way. The proposed features are shown to be considerably more robust against additive background noise than standard mel-frequency cepstrum coefficients on two different recognition tasks. The performance improvement increased as we moved from a small-vocabulary isolated-word task to a medium-vocabulary continuous-speech task, where the proposed features also outperformed a computationally expensive auditory-based method. The greatest improvement was obtained for noise types characterized by a relatively flat spectral density.**

***Index Terms*—Auditory models, dominant frequencies, feature extraction, noise robustness, speech recognition, subband spectral centroids (SCCs).**

## I. Introduction

A MAJOR limitation for the use of automatic speech recognition (ASR) in many practical applications is its lack of robustness against changes in the acoustic environment. This problem is especially pronounced when designing speech-based interfaces for mobile communication devices. The ever decreasing size of these devices makes a speech-based interface very attractive. However, since the acoustic environment is unpredictable and variable, it is not possible to account for it during ASR system training. This causes a mismatch between the trained speech models and the actual speech to be recognized, which results in a severe degradation of the recognition performance. Much research has been done during the last decade in order to find efficient methods for reducing the mismatch [1]–[3].

ASR is based on speech feature vectors that contain relevant information for discriminating between different speech sounds. Commonly used speech features, e.g., mel-frequency cepstral coefficients (MFCCs) and linear prediction cepstral coefficients (LPCCs), are highly affected by changes in the acoustic environment [4]. One way of increasing the robustness of ASR systems is to find speech features that are less sensitive to changes in the acoustic environment, while retaining good discriminative properties.

The exceptional ability of the human auditory system to recognize speech in noisy acoustic environments has inspired the use of knowledge on human speech perception in speech feature extraction for ASR. This research has resulted in a number of feature extraction methods based on detailed modeling of the processes in the human auditory system. Those auditory-based methods have generally shown increased robustness against environmental noise compared to standard feature extraction methods.

A closer look at several auditory-based methods has revealed that they utilize, in one way or another, the information about dominant frequencies in the speech signal (spectral peak positions). This information is not used explicitly in standard feature extraction methods. The spectral peak positions remain largely unaffected by environmental noise as long as the noise spectrum does not contain strong spectral peaks. This might explain the increased noise robustness of auditory-based features compared to standard speech features.

In this study, we investigate whether a simple method of incorporating dominant-frequency information into speech features gives a similar improvement of noise robustness as achieved by auditory-based methods. Reliable estimation of the spectral peak positions is a difficult task, especially in the presence of noise. However, it was shown in [5] that subband spectral centroids (SSCs), computed as the first moments of the speech power spectrum over different frequency subbands, are closely related to spectral peak positions and quite robust against additive white Gaussian noise. Consequently, we propose a new feature extraction method that combines the dominant-frequency information provided by the SSCs with the subband power information used by standard feature extraction methods. The new features are referred to as subband spectral centroid histograms (SSCHs).

The paper is organized as follows. Section II reviews the use of dominant-frequency information in several auditory-based feature extraction methods. Section III introduces the SSCH features. The recognition tasks and databases used for the evaluation of their ASR performance are described in Section IV. The results of an experimental study aimed at optimizing several parameters involved in the SSCH computation are presented in Section V. In Section VI, these features are compared to both standard and auditory-based features. Finally, the major conclusions from our study are summed up in Section VII.

Manuscript received February 8, 2004; revised November 21, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Bayya Yegnanarayana.

B. Gajić was with the School of Microelectronic Engineering, Griffith University, Brisbane, QLD 4111 Australia, on leave from the Department of Telecommunications, Norwegian University of Science and Technology, N-7491 Trondheim, Norway (e-mail: gajic@iet.ntnu.no).

K. K. Paliwal is with the School of Microelectronic Engineering, Griffith University, Brisbane, QLD 4111 Australia (e-mail: k.paliwal@me.gu.edu.au).

Digital Object Identifier 10.1109/TSA.2005.855834

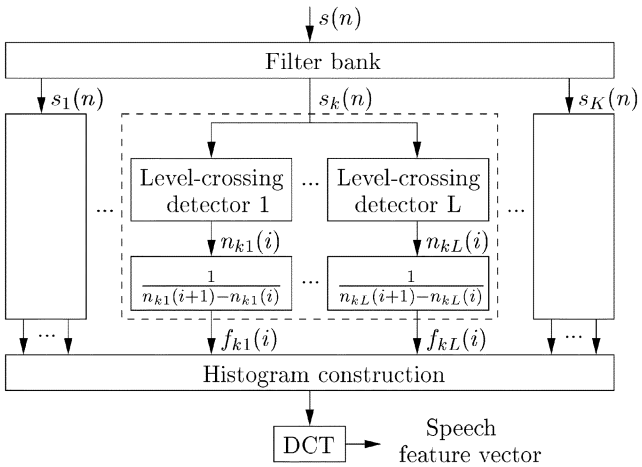


Fig. 1. EIH method for speech feature extraction.

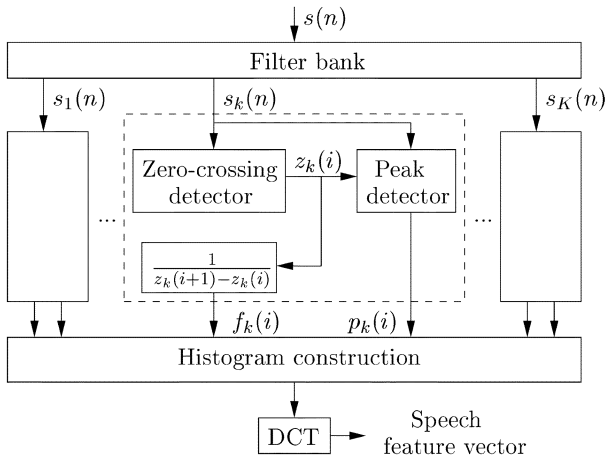


Fig. 2. ZCPA method for speech feature extraction.

## II. Dominant-Frequency Information in Auditory-Based Feature Extraction

This section reviews the use of dominant-frequency information in several earlier proposed auditory-based feature extraction methods. It makes the basis for our belief that the improved noise robustness of the auditory-based methods is mainly due to the use of dominant-frequency information rather than to the detailed modeling of the human auditory system.

### A. Synchrony Spectrum

The synchrony spectrum [6] is a speech feature type based on the detailed modeling of the processes in the human auditory system. It consists of the outputs from a set of generalized synchrony detectors, one for each subband, that measure the extent of dominance of the periodicities at subband center frequencies. Therefore, the subbands with center frequencies close to the spectral peaks obtain the highest scores. In that way, the information on the dominant frequencies in the speech signal is included into the feature vectors.

### B. Subband Autocorrelation

Subband autocorrelation (SBCOR) analysis [7], [8] is a simplification of the synchrony spectrum computation. It is based on a simple bandpass filtering followed by the computation of the autocorrelation coefficients for the subband signals at time $\tau = 1/F_c$, where $F_c$ is the subband center frequency. Generally, a spectral peak at frequency $F$ gives rise to peaks in the autocorrelation function at integer multiples of $1/F$. Consequently, the value of the subband autocorrelation coefficient at time $\tau = 1/F_c$ indicates the extent of dominance of the subband center frequency in the subband signal. SBCOR features were shown to outperform standard features based on subband power estimates [7], and their robustness against different types of speech distortions was demonstrated in [8].

### C. Ensemble Interval Histograms

Ensemble interval histograms (EIHs) [9], [10] are probably the best known auditory-based features that resulted from a detailed modeling of the human auditory system. They are computed by filtering the speech signal through a cochlear filter bank, and applying a set of level-crossing detectors to each subband signal, as shown in Fig. 1. For a given speech frame, the intervals between successive crossings of the same level are measured. The inverse interval lengths, computed for all levels and all subbands, are then collected in a single histogram. Finally, the discrete cosine transform (DCT) is performed for decorrelation purposes. Note that the level-crossing rates for a given subband are related to the dominant subband frequency. Furthermore, the number of levels crossed is related to the subband signal power. An EIH thus combines the dominant subband frequency information with the subband power information.

### D. Zero-Crossings With Peak Amplitudes

The zero-crossings with peak amplitudes (ZCPA) feature extraction method [11] is a simplification of the EIH computation. It is illustrated in Fig. 2. The cochlear filter shapes are replaced by simple filters designed by the windowing method. Furthermore, the set of level-crossing detectors is replaced by a single zero-crossing detector, while the subband power information is preserved by measuring the peak amplitudes between successive zero-crossings. Each speech frame is thus represented by a histogram of the inverse zero-crossing intervals corresponding to all the subband signals. Instead of increasing the histogram bin counts by one, they are increased by the logarithm of the corresponding peak amplitudes.

The dominant-frequency principle [12] states that if there is a significantly dominant frequency in the signal spectrum, then the inverse zero-crossing intervals tend to take values in the vicinity of this frequency. Therefore, the inverse zero-crossing intervals for a given subband signal can be seen as estimates of the dominant subband frequency. Furthermore, the peak signal amplitudes between subsequent zero crossings are related to the power of the subband signal. Consequently, the ZCPA histogram construction can be seen as assigning subband power estimates to the histogram bins corresponding to the dominant subband frequencies. The standard MFCC method, on the other hand, assigns subband power estimates to entire subbands, without taking into account the power distribution within subbands. ZCPA histograms can hence be seen as an alternative spectral representation of speech that emphasizes spectral peaks.

Since the processing is done on discrete-time signals, the zero-crossing intervals are measured in terms of integer number of samples between successive zero crossings. The resolution of dominant subband frequency estimates, computed as in-

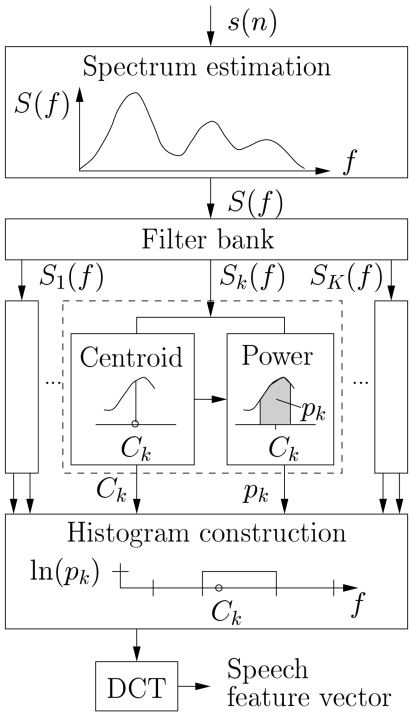


Fig. 3. SSCH method for speech feature extraction.

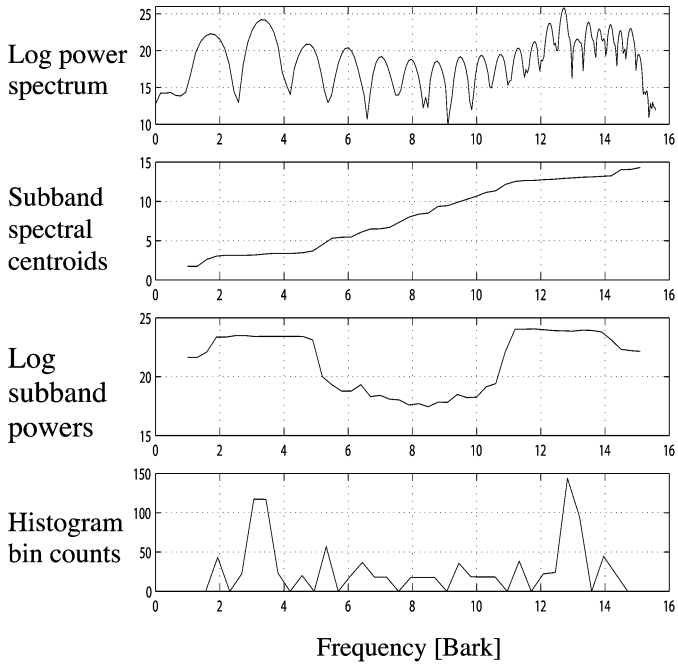


Fig. 4. Example of the SSCH computation.

verse zero-crossing intervals, decreases therefore rapidly with increased frequency (e.g., the difference between 1/8 and 1/9 is 90 times greater than the difference between 1/80 and 1/81, while the frequency is only 10 times greater). This leads to low accuracy of dominant frequency estimates for high-frequency subband signals. The problem can be circumvented by up-sampling the high-frequency subband signals using frequency-dependent interpolation factors.

The computational cost of the ZCPA method is considerably lower than that of the EIH method and other auditory based methods. However, compared to the standard MFCC method, the computational complexity is still very high. This is due to the use of time-domain filtering, and the need for heavy interpolation of the high-frequency subband signals.

The ZCPA features were shown to greatly outperform LPCC, MFCC, PLP, SBCOR, and EIH features on a small-vocabulary isolated-word ASR task in the presence of different types of additive background noise [11]. The superior robustness of the ZCPA features compared to the MFCC features was confirmed in [13], on both a small-vocabulary isolated-word task, and a medium-vocabulary continuous-speech task.

## III. Subband Spectral Centroid Histograms

The robustness of the ZCPA features against additive background noise has indicated the positive effect of integrating dominant subband frequency information and subband power information into speech features. However, this method is less attractive for practical applications due to the high computational cost. Our aim was to develop a new feature extraction method that extracts the same conceptual information from the speech signal with a reduced computational complexity.

Subband spectral centroids (SSCs) have been shown to be closely related to spectral peak positions, both for clean and noisy speech [5]. Moreover, they can be computed efficiently from a short-term speech power spectrum estimate. Several studies [5], [14]–[16] have investigated the effect of augmenting standard speech feature vectors by SSCs, but the results have not been consistent.

In our work, we seek for a better method of integrating the dominant-frequency information provided by the SSCs with the subband power information. This is achieved by constructing a histogram in a way similar to that of the ZCPA method. The resulting features are referred to as subband spectral centroid histograms (SSCHs) [17]–[19]. Similar histograms have been used earlier with average instantaneous frequencies computed from individual subband signals to derive features for speech recognition [20], [21].

### A. Algorithm Description

The procedure for the SSCH computation is illustrated in Fig. 3. It starts by computing a fast Fouier transform (FFT)-based power spectrum estimate $S(f)$ for the given speech frame and passing it through a set of K highly overlapping bandpass filters with amplitude responses $H_k(f)$, $k = 1,\ldots,K$. SSCs are then computed as

$$C_k = \frac{\sum f H_k(f) S^\gamma(f)}{\sum H_k(f) S^\gamma(f)}, \qquad k = 1,\ldots,K \tag{1}$$

where $\gamma$ is a parameter that controls spectral dynamic range, and the summation is performed over all frequency samples in the FFT. Subband power estimates are computed as

$$p_k = \sum H_k(f) S(f), \qquad k = 1,\ldots,K \tag{2}$$

where the summation is performed either over entire subbands, or over smaller frequency ranges centered around subband centroids. This is discussed further in Section V-C.

Next, a histogram of the SSCs is constructed by dividing the speech frequency range into bins $R_j$, $j = 1, \ldots, J$, and computing the bin counts as

$$\text{count}(j) = \sum_k \Psi_j\{C_k\}, \qquad j = 1, \ldots, J \tag{3}$$

where

$$\Psi_j\{C_k\} = \begin{cases} \ln\left(\frac{p_k}{N_k}\right), & C_k \in R_j \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

and $N_k$ is the number of frequency samples in the $k$th subband. For each SSC, the corresponding bin count is thus increased by the logarithm of the subband power estimate normalized by the subband bandwidth. Finally, the DCT of the histogram is computed for decorrelation purposes.

Fig. 4 illustrates the outputs from different stages of SSCH computation plotted against Bark frequency, for a frame of vowel /i:/. Notice that the flat portions in the SSC plot represent spectral peaks, with the length indicating the degree of spectral dominance. Furthermore, the prominent histogram peaks closely follow the peaks in the power spectrum.

The SSCH features are somewhat similar to the DOMIN auditory model [22]. In this model, histograms of the subband peak frequencies are computed, but no explicit subband power measurements are incorporated.

### *B. Relationship to ZCPA*

Both SSCH and ZCPA features are based on combining the dominant-frequency information and power information from the subband signals in a single histogram representation. However, they differ in the way the information is obtained from the speech signal. While the ZCPA method operates in the time domain by measuring the zero-crossing rates and peak amplitudes between zero crossings, the SSCH method computes the SSCs and subband powers from the short-term spectral estimates. Consequently, the SSCH method provides a more computationally efficient way of extracting the same type of information. The computational complexity of the SSCH method was found to be two orders of magnitude lower than that of the ZCPA method [17].

### *C. Relationship to MFCC*

The SSCH and MFCC feature extraction methods have several common processing steps: spectral estimation, subband filtering, and subband power computation. However, the SSCH method incorporates two additional steps, namely centroid computation and histogram construction. Consequently, while the MFCC method assigns a subband power estimate to an entire subband, the SSCH method assigns it to the histogram bin that contains the subband centroid. In this way, the information on the spectral peak positions is preserved better. Nevertheless, it is important to remember that the SSCs are only estimates of the spectral peak positions computed from the speech spectra. They are thus affected by noise even if the true spectral peaks remain unchanged. The computational complexity of the SSCH method is of the same order of magnitude as that of the MFCC method [17].

## IV. Recognition Tasks

Two different recognition tasks were used for evaluation of the proposed method: a small-vocabulary isolated-word task based on the ISOLET Spoken Letter Database [23] and a medium-vocabulary continuous-speech task based on the speaker-independent part of the DARPA Resource Management (RM) database [24]. Both databases were recorded in quiet conditions using close-talking noise-canceling microphones. They were down-sampled to 8 kHz in our study in order to reduce the processing time needed for the different feature extraction methods. Model training and evaluation were performed using the HTK 3.0 program package [25]. The ASR performance was measured in terms of word accuracy

$$WAC = \frac{N - S - D - I}{N} \cdot 100\% \tag{5}$$

where $N$ is the total number of words in the test set, $S$ is the number of substitution errors, $D$ is the numbers of deletion errors, and $I$ is the number of insertion errors.

The first task consisted of recognizing spoken letters from the English alphabet. Each vocabulary word was modeled by a five-state left-to-right hidden Markov model (HMM) with five Gaussian components per state and no skip transitions. Utterances from 90 speakers (subsets ISOLET-1, ISOLET-2, and ISOLET-3) were used for model training, while utterances from an additional 30 speakers (subset ISOLET-5) were used for evaluation.

The second task involved recognizing queries about ships and ports along with commands to control a graphics display system. The number of vocabulary words was 991. A set of tied-state cross-word triphone models was trained by following the RM Recipe that is supplied with the HTK distribution. Each model consisted of three states and six Gaussian components per state. State tying was performed using decision tree clustering. The training was performed on 3999 sentences spoken by 109 speakers, while 300 sentences spoken by additional 10 speakers (February'89 test set) were used for the evaluation. The word-pair language model supplied with the RM database was used in our recognition experiments.

For the purpose of evaluating the robustness against environmental noise, three different noise types were added to the test data at several SNRs, namely white Gaussian noise, factory noise, and background speech. White Gaussian noise was generated using a random noise generator, while the factory noise and background speech were taken from the NOISEX database [26], where they are referred to as factory1 and babble noise, respectively.

Noisy speech was generated in the following way. For each speech file in the test database, a noise segment of length equal to the length of the speech file was randomly extracted, multiplied by a gain factor $g$ and added to the speech file. The gain factor was computed in accordance with the required SNR, defined as

$$SNR[dB] = 10\log_{10}\left(\frac{p_s^{\max}}{g^2 p_n}\right) \tag{6}$$

where $p_s^{\max}$ is the maximal frame power of the given speech file and $p_n$ is the noise power estimated over the noise segment. The SNR is thus measured as the ratio between the maximal speech power and the average noise power. This computation method

TABLE I
ASR PERFORMANCE OF THE SSCH METHOD AS A FUNCTION OF THE DYNAMIC RANGE PARAMETER $\gamma$

| $\gamma$ | Word accuracy [%] | | | | |
|---|---|---|---|---|---|
| | clean | SNR [dB] | | | |
| | speech | 25 | 20 | 15 | 10 |
| 0.5 | 86.5 | 77.9 | 70.8 | 58.1 | 34.0 |
| 1 | 86.2 | 79.7 | 73.4 | 59.9 | 42.2 |
| 2 | 85.5 | 77.0 | 71.5 | 60.8 | 44.6 |
| 4 | 83.7 | 75.3 | 71.4 | 60.5 | 44.6 |

makes the SNR independent of both the phonetic content of the speech utterance and the length of silence intervals surrounding the speech utterance.

## V. OPTIMIZING PARAMETER VALUES

This section presents the results of an experimental study aimed at optimizing several parameters involved in SSCH computation which were considered to be of a particular importance for the ASR performance. The choice of the remaining parameters is described in Section VI-A1. The study was performed on the ISOLET database, both on clean speech and in the presence of white Gaussian noise at various SNRs.

### A. Spectral Dynamic Range

The spectral dynamic range used in the SSC computation is controlled by parameter $\gamma$ in (1). If $\gamma$ is too small (close to 0), SSCs would approach the centers of their subbands, and thus contain no information. If it is too large (close to infinity), the SSCs would correspond to the locations of the subband peak values of the FFT-based power spectrum, and would thus be noisy estimates.

Table I shows the recognition performance of the SSCH method as a function of the parameter $\gamma$. We observe that increasing the dynamic spectral range up to a certain level had a positive effect on the recognition performance in the presence of noise. This result is reasonable, since increasing $\gamma$ makes spectral peaks more prominent, and thus reduces the effect of additive noise. In the rest of this study we used $\gamma = 1$, since this value gave the best overall performance across all SNRs.

The results in Table I were obtained using the optimized values for the filter bank and histogram parameters found in the remainder of this section. Similar trends were also observed with a different choice of the parameters that was used in the beginning of the optimization process.

### B. Filter-Bank Design and Histogram Construction

The filter bank used in this study consists of highly overlapping filters with rectangular frequency responses and center frequencies uniformly distributed on the Bark scale between 100 and 3800 Hz. The rectangular frequency responses were chosen since any other shape (e.g., triangular), would favor some frequencies within the subband more than others and thus give biased SSC estimates.

Filter bandwidths should ideally be chosen such that each subband contains exactly one dominant spectral peak. In this case, SSCs serve as good estimates of spectral peak positions. Too small filter bandwidths result in a number of subbands that do not contain any dominant spectral peak. Centroids of such subbands are sensitive to random variabilities in speech. On the other hand, if filter bandwidths stretch over several dominant spectral peaks, SSCs will no longer represent reasonable estimates of subband peak locations.

TABLE II
ASR PERFORMANCE OF THE SSCH METHOD FOR DIFFERENT CHOICES OF FILTER BANK PARAMETERS AND HISTOGRAM BIN ALLOCATION

| Filter bw/no | Filt bw / Bin bw | No. bin | Word accuracy [%] | | | | |
|---|---|---|---|---|---|---|---|
| | | | Clean speech | SNR [dB] | | | |
| | | | | 25 | 20 | 15 | 10 |
| 1/48 | 4 | 57 | 88.46 | 78.08 | 69.36 | 55.96 | 30.83 |
| 1/143 | 4 | 57 | 87.95 | 78.27 | 70.19 | 55.38 | 28.25 |
| 1/143 | 6 | 86 | 88.44 | 77.56 | 67.50 | 53.85 | 27.76 |
| 2/72 | 2 | 14 | 84.29 | 73.53 | 67.44 | 54.81 | 35.90 |
| 2/72 | 4 | 29 | 86.28 | 77.24 | 70.90 | 59.55 | 38.33 |
| 2/48 | 6 | 43 | 87.18 | 78.91 | 71.67 | 59.17 | 38.97 |
| 2/72 | 6 | 43 | 86.86 | 79.17 | 72.50 | 58.65 | 38.21 |
| 2/72 | 8 | 57 | 86.47 | 78.46 | 71.22 | 58.85 | 37.31 |
| 3/48 | 4 | 19 | 83.65 | 76.86 | 69.74 | 58.40 | 40.45 |
| 3/24 | 6 | 29 | 85.38 | 78.65 | 73.40 | 60.64 | 43.97 |
| 3/48 | 6 | 29 | 86.47 | 78.46 | 72.44 | 59.81 | 41.67 |
| 3/72 | 6 | 29 | 85.96 | 78.72 | 72.37 | 59.23 | 41.41 |
| 3/24 | 8 | 38 | 86.41 | 79.87 | 72.82 | 59.87 | 43.21 |
| 3/48 | 8 | 38 | 86.15 | 79.74 | 73.40 | 59.87 | 42.24 |
| 4/36 | 6 | 21 | 84.29 | 75.45 | 66.60 | 54.81 | 33.21 |
| 4/36 | 8 | 29 | 86.60 | 78.08 | 69.10 | 56.79 | 36.22 |
| 4/36 | 10 | 36 | 85.19 | 77.69 | 68.85 | 56.22 | 36.03 |

TABLE III
ASR PERFORMANCE OF THE SSCH METHOD FOR TWO DIFFERENT METHODS OF SUBBAND POWER ESTIMATION

| Power integration range | Word accuracy [%] | | | | |
|---|---|---|---|---|---|
| | clean speech | SNR [dB] | | | |
| | | 25 | 20 | 15 | 10 |
| whole subband | 86.15 | 79.74 | 73.40 | 59.87 | 42.24 |
| 1 Bark | 86.35 | 80.45 | 74.10 | 61.60 | 42.50 |

Histogram bins should be sufficiently small to provide a good frequency resolution, but if they become too small, the resulting speech parameterization will become sensitive to small fluctuations in spectral peak positions (e.g., due to speaker differences). In this study, histogram bins having equal lengths on the Bark scale were used. This provides better frequency resolution in low-frequency subbands than in high-frequency subbands, which is in agreement with the processing in the human auditory system. In order to capture the information about subband power distribution, filter bandwidths must stretch over several histogram bins.

A series of recognition experiments was performed in order to optimize the filter bandwidths, the number of filters and the number of histogram bins. The results are shown in Table II.

We observe that the choice of the filter bandwidths had a significant effect on the recognition performance, especially at low SNRs. The best results in the presence of noise were achieved using filter bandwidths equal to 3 Bark (302–1927 Hz), while the performance on clean speech was best for filter bandwidths equal to 1 Bark (101–642 Hz). The recognition performance was not very sensitive to the particular choice of the number of filters.

The number of histogram bins had to be chosen large enough to provide a sufficiently good frequency resolution, but the recognition performance was not very sensitive to the particular

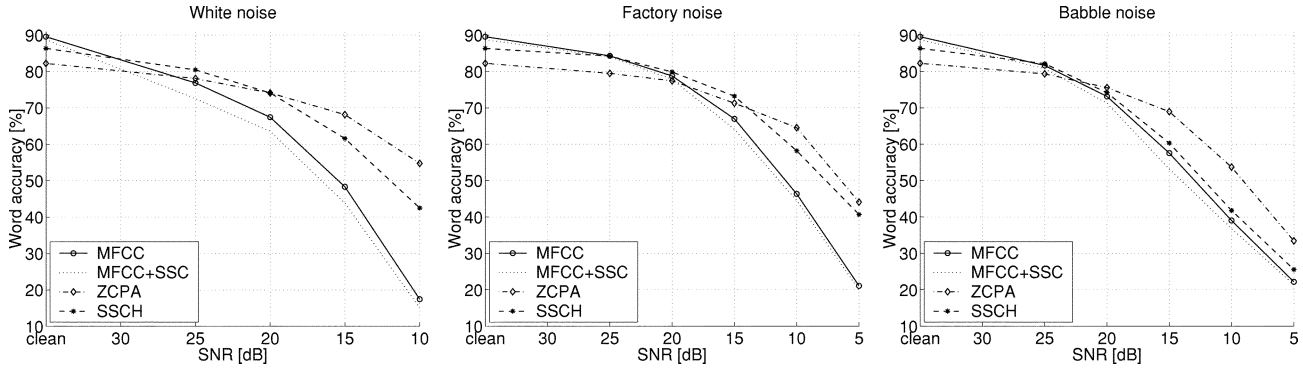


Fig. 5. ASR performance of different feature types on the ISOLET database in the presence of white, factory, and babble noise, respectively.

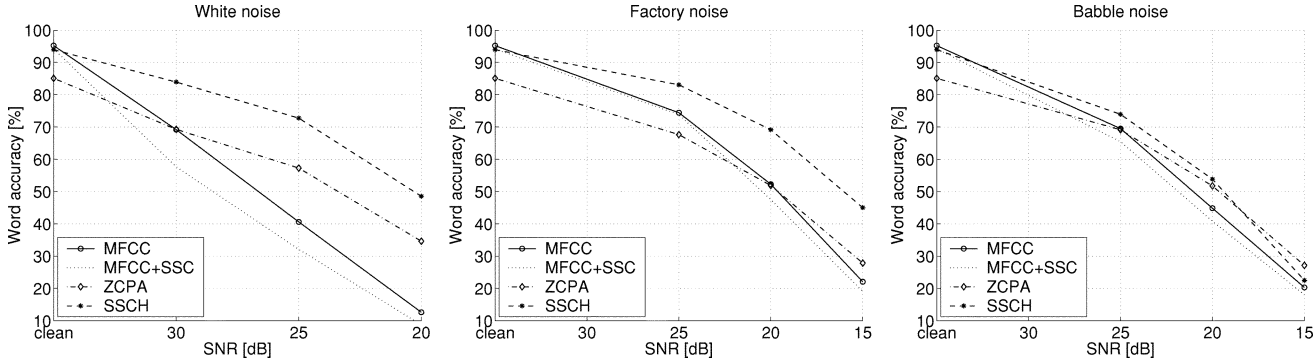


Fig. 6. ASR performance of different feature types on the RM database in the presence of white, factory and babble noise, respectively.

choice of the number of bins—values between 30 and 60 gave good results in all test conditions.

In the rest of this study, SSCH features were computed using 48 subband filters with bandwidths equal to 3 Bark. The number of histogram bins was set to 38. This parameter choice provided the best overall performance across all SNRs. (Note that equally good performance was obtained using 24 filters, so this value might have been chosen as well.)

### C. Subband Power Computation

Subband power estimates are computed by integrating the speech spectrum across the different subbands. Alternatively, the integration can be done over smaller frequency ranges centered around the SSCs. This might provide more robust estimates since the frequency range around the dominant frequency is less influenced by noise than the other parts of the subband. On the other hand, smaller integration ranges lead to less reliable estimates.

In the experimental study, we investigated the effect of reducing the integration range to 1 Bark around the centroids. The results are shown in Table III. It can be seen that slightly better results were obtained for the reduced integration range, but the difference was not statistically significant. In the rest of the study we used the integration range of 1 Bark centered around the centroids.

## VI. Comparison to Other Methods

This section presents the results obtained by evaluating the SSCH method on the recognition tasks described in Section IV. The performance of the SSCH features was compared to that of the standard MFCC features, MFCC features augmented by three SSCs, and ZCPA features.

### A. Implementational Details

*1) Subband Spectral Centroid Histograms:* The SSCH features were computed according to the procedure described in Section III-A. Speech analysis frames were 25 ms long, with a 10-ms frame shift. They were first passed through a first-order preemphasis filter with filter coefficient 0.97, followed by Hamming windowing and 512-order FFT computation. (The relatively high FFT-order was used in order to be able to implement sufficiently small spacing between the bandpass filters which was needed in some of the experiments described in Section V-B.) The power spectrum estimates were subsequently obtained by squaring the magnitudes of the FFT coefficients. The spectral dynamic range parameter was set to $\gamma = 1$.

The filter bank and histogram parameters were chosen in accordance with the results given in Section V. The filter bank consisted of 48 rectangular filters with bandwidths equal to 3 Bark (302–1927 Hz), and center frequencies uniformly distributed on the Bark scale between 100–3800 Hz. The frequency range between 100–3800 Hz was divided into 38 histogram bins which were uniformly distributed on the Bark scale.

Finally, the first 12 DCT coefficients (not including the 0th coefficient) were computed from the histograms and augmented by the delta and delta-delta coefficients. The delta coefficients were computed using the following regression formula:

$$d_t = \frac{\sum_{\theta=1}^{2} \theta (c_{t+\theta} - c_{t-\theta})}{2 \sum_{\theta=1}^{2} \theta^2} \tag{7}$$

where $d_t$ is a delta coefficient at time $t$, and $c_{t+\theta}$ and $c_{t-\theta}$ are corresponding static coefficients. Delta-delta coefficients were computed from delta coefficients using the same formula.

*2) Mel-Frequency Cepstral Coefficients:* The MFCC features were computed using a fairly standard procedure. The

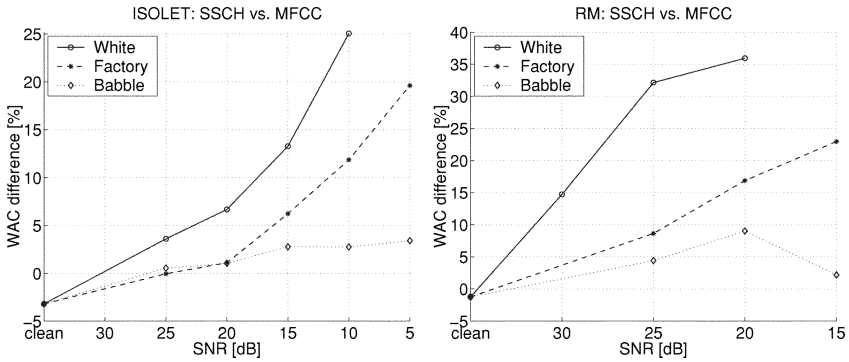


Fig. 7. Difference in ASR performance of SSCH and MFCC features in the presence of different noise types for ISOLET and RM databases, respectively.

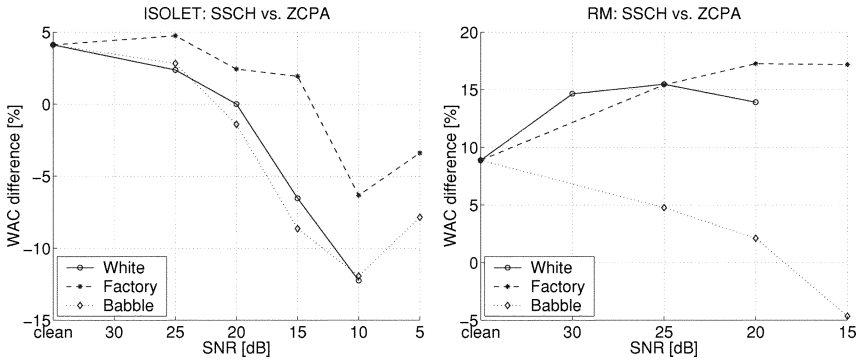


Fig. 8. Difference in ASR performance of SSCH and ZCPA features in the presence of different noise types for ISOLET and RM databases, respectively.

speech analysis frames were 25 ms long, with a 10-ms frame shift. They were first passed through a first-order preemphasis filter with filter coefficient 0.97, followed by Hamming windowing and 256-order FFT computation.

Subband filtering was performed next using a set of 20 triangular filters with 50% overlap between neighboring filters. Filter center frequencies were uniformly distributed on the mel scale between 66 and 3592 Hz, with bandwidths equal to 102 mel (i.e., 70–389 Hz). Subband power estimates were then computed by integrating the corresponding subband spectra.

Finally, 12 DCT coefficients were computed from the subband power estimates along with the delta and delta-delta coefficients in the same way as for the SSCH features.

*3) MFCC Augmented by SSCs:* In this method, the 36-dimensional MFCC feature vectors described above were augmented by three SSCs similarly as in [14]. The computation of the SSCs was similar to that of the SSCH method. The only difference was in the choice of the filter bank, which in this case consisted of three disjoint rectangular filters uniformly distributed on the linear frequency scale between 0–4000 Hz. The bandwidth of each filter was thus of approximately 1333 Hz. The resulting speech features are referred to as MFCC+SSC.

*4) Zero Crossings With Peak Amplitudes:* The ZCPA feature extraction method was summarized in Section II-D. The parameters involved in the computation were chosen according to the results in [13] and [17] so as to obtain a good compromise between the performances at low and high SNRs.

Frequency-dependent analysis-frame lengths were used, ranging between 33 ms at the high-frequency end and 134 ms at the low-frequency end. The analysis frames were extracted using rectangular windows with a frame shift equal to 10 ms. No preemphasis was used.

The filter bank consisted of 16 Hamming finite impulse response filters designed by the windowing method, with center frequencies uniformly distributed on the Bark scale between 200–3400 Hz. The bandwidths of the ideal prototype filters were equal to 2 Bark. The interpolation factor ranged from 1 (i.e., no interpolation) for the first four subbands to 16 for the last three subbands. The average interpolation factor was equal to 6.

The frequency range between 0–4000 Hz was divided into 60 histogram bins which were uniformly distributed on the Bark scale. Finally, 12 DCT coefficients along with the delta and delta-delta coefficients were derived from the histograms in the same way as for the SSCH and MFCC features.

### *B. Experimental Results*

The recognition performance of the feature types described above is respectively illustrated in Figs. 5 and 6 for the ISOLET and RM databases. Different acoustic environments (white noise, factory noise and babble noise) are considered. The results are discussed hereafter.

*1) SSCs as Additional Speech Features:* By comparing the performance of the MFCC and MFCC+SSC features we observe that the addition of the three SSCs to the standard MFCC feature vectors led to a consistent small performance degradation. This is in contrast to the results of the earlier studies [5], [14]–[16], which reported positive effects when appending SSCs to the MFCC feature vectors. However, the previous results were inconsistent. While some researchers observed a positive effect when adding SSCs only in the case of clean speech, others reported increased positive effects with reduced SNR.

The poor robustness of the SSC features can be explained by the fact that SSCs serve as reasonable estimates of speech spectral peak positions only in the subbands that contain a single

TABLE IV
PERFORMANCE COMPARISON BETWEEN STANDARD MFCC FEATURES, MODIFIED MFCC FEATURES AND SSCH FEATURES ON THE ISOLET DATABASE IN THE PRESENCE OF WHITE GAUSSIAN NOISE

| Feature type | Word accuracy [%] | | | | |
|---|---|---|---|---|---|
| | Clean speech | SNR [dB] | | | |
| | | 25 | 20 | 15 | 10 |
| MFCC | 89.6 | 76.9 | 67.4 | 48.3 | 17.4 |
| MFCC modified | 86.1 | 74.4 | 64.4 | 48.6 | 21.2 |
| SSCH | 86.4 | 80.5 | 74.1 | 61.6 | 42.5 |

spectral peak. However, since spectral peak positions vary according to the particular speech sound, it is not possible to design a filter bank that would produce suitable subband locations for all speech sounds.

*2) Comparing SSCH and MFCC Performance:* We can see from Figs. 5 and 6 that SSCH features outperformed MFCC features in the presence of additive noise, while MFCC features performed slightly better on clean speech. Fig. 7 shows the difference in the performance of the two feature types in various background conditions for the ISOLET and RM databases respectively. The advantage of using the SSCH features was largest in the case of white noise, followed by factory noise, while only a small improvement was observed in the case of babble noise. This can be explained by the presence of prominent spectral peaks in babble noise, that make dominant subband frequency information less reliable. This problem is less pronounced in the presence of factory noise, where intervals characterized by prominent spectral peaks interchange with those characterized by a relatively flat spectrum. Another interesting observation drawn from Fig. 7 is that the maximal improvement achieved by using SSCH features instead of MFCC features was larger for the more complex recognition task.

Furthermore, we wanted to find out whether the use of different filter banks in the MFCC and SSCH computation had a significant influence on the difference in their performance. We thus derived a set of modified MFCC features by replacing the filter bank by the one used in the SSCH computation. Table IV compares the recognition performance of the original MFCC, modified MFCC and SSCH features on the ISOLET database, both for clean speech and in the presence of additive white Gaussian noise. It can be seen that the performance of the modified MFCC features followed closely that of standard MFCC features, with only a small degradation at high SNRs, and a small improvement at low SNRs. This indicates that the superior robustness of the SSCH features compared to the MFCC features is mainly due to the use of the dominant subband frequency information provided by SSCs.

Note that the problem of the proper choice of the subbands is not as critical for the SSCHs as it is for SSCs that are used directly as speech features. The centroids of the subbands that do not contain any spectral peak are highly affected by noise, but since the power of such subbands is relatively low, their contribution to the histogram is small. Thus, the SSCH method incorporates an efficient weighting scheme, which assigns larger weights to more reliable SSCs. However, this is only true if the noise does not contain prominent spectral peaks.

*3) Comparing ZCPA and SSCH Performance:* Fig. 8 illustrates the difference in the performance of SSCH and ZCPA features in various background conditions for the ISOLET and RM databases respectively. It can be seen that, on the ISOLET database, SSCH features performed slightly better than ZCPA features at high SNRs, while ZCPA features performed better at low SNRs. On the RM database, on the other hand, SSCH features performed considerably better than ZCPA features in the presence of white and factory noise at all SNRs, while there was only a small difference in performance in the presence of babble noise. In addition, computational complexity of the SSCH method was two orders of magnitude lower than that of the ZCPA method.

## VII. SUMMARY AND CONCLUSION

In this paper, we demonstrated that the use of dominant-frequency information in speech feature extraction leads to increased ASR robustness against additive background noise. We started by reviewing several earlier proposed auditory-based features, and showed that their superior robustness against additive background noise might be mainly due to the use of the dominant-frequency information, rather than to a detailed modeling of the processes in the human auditory system. We then proposed a new feature extraction method that combines the dominant subband frequency information provided by SSCs with subband power estimates in a simple and computationally efficient way.

An experimental study was subsequently performed in order to optimize the parameters involved in the SSCH computation. Finally, the proposed features were compared to standard MFCC features, MFCC features augmented by three SSCs, and auditory-based ZCPA features on two different recognition tasks in various background conditions. The major results are summarized in the following.

- Augmenting SSCs to the standard MFCC feature vectors had a small negative effect on the recognition performance. This can be explained by poor robustness of the SSC features if subband positions are not appropriately chosen. This problem is effectively circumvented in the SSCH method.
- SSCH features outperformed standard MFCC features in the presence of additive noise. The improvement increased with increased task complexity and with reduced SNR. It was largest in the presence of noise types with relatively flat spectral characteristics.
- The advantage of using SSCH features compared to MFCC features was mainly due to the use of dominant subband frequency information.
- SSCH features were also considerably more robust than ZCPA features on the more complex recognition task, except in the case of babble noise.
- The computational complexity of the SSCH method is two orders of magnitude lower than that of the ZCPA method, and of the same order of magnitude as the MFCC method.

The major limitation of the SSCH method lies in the fact that it is designed to deal with additive noise only. Furthermore, it is implicitly assumed that spectral peaks belong to speech. As a consequence, the method is not effective in the case of additive background noise characterized by strong spectral peaks, such as babble noise.

An advantage of robust feature extraction methods compared to most other methods for increasing noise robustness in ASR is

the fact that they do not require any knowledge of the target environment. However, in situations where such knowledge is available or easy to obtain, a better recognition performance might be obtained by utilizing this knowledge. Thus, an important extension of the work presented in this paper would be to investigate whether the use of the SSCH features can be effectively combined with some of the other methods for increasing noise robustness. Note that such a combined approach could also circumvent the limitation of the SSCH method to additive noise.

## REFERENCES


[1] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Englewood Cliffs, NJ: Prentice-Hall, 2001.

[2] J.-C. Junqua and J.-P. Haton, *Robustness in Automatic Speech Recognition—Fundamentals and Applications*. Norwell, MA: Kluwer, 1996.

[3] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Commun.*, vol. 16, no. 3, pp. 261–291, Apr. 1995.

[4] J. P. Openshaw and J. S. Mason, "On the limitations of cepstral features in noise," in *Proc. ICASSP*, vol. 2, 1994, pp. 49–52.

[5] K. K. Paliwal, "Spectral subband centroid features for speech recognition," in *Proc. ICASSP*, vol. 2, May 1998, pp. 617–620.

[6] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *J. Phonetics*, vol. 16, no. 1, pp. 55–76, Jan. 1988.

[7] S. Kajita and F. Itakura, "Subband-autocorrelation analysis and its application for speech recognition," in *Proc. ICASSP*, vol. 2, 1994, pp. 193–196.

[8] ——, "Robust speech feature extraction using SBCOR analysis," in *Proc. ICASSP*, vol. 1, May 1995, pp. 421–424.

[9] O. Ghitza, "Temporal nonplace information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment," *J. Phonetics*, vol. 16, no. 1, pp. 55–76, Jan. 1988.

[10] ——, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 1, pp. 115–132, Jan. 1994.

[11] D.-S. Kim, S.-Y. Lee, and R. M. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 1, pp. 55–69, Jan. 1999.

[12] B. Kedem, "Spectral analysis and discrimination by zero-crossings," *Proc. IEEE*, vol. 74, no. 11, pp. 1477–1493, Nov. 1986.

[13] B. Gajić and K. K. Paliwal, "Robust speech recognition using features based on zero crossings with peak amplitudes," in *Proc. ICASSP*, Apr. 2003.

[14] S. Tsuge, T. Fukada, and H. Singer, "Speaker normalized spectral subband parameters for noise robust speech recognition," in *Proc. ICASSP*, May 1999.

[15] 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, Sep. 1999, pp. 1503–1506.

[16] R. De Mori, D. Albesano, R. Gemello, and F. Mana, "Ear-model derived features for automatic speech recognition," in *Proc. ICASSP*, vol. 3, 2000, pp. 1603–1606.

[17] B. Gajić, "Feature extraction for automatic seech recognition in noisy acoustic environments," Ph.D. dissertation, Norwegian Univ. Sci. Technol., Trondheim, 2002.

[18] B. Gajić and K. K. Paliwal, "Robust feature extraction using subband spectral centroid histograms," in *Proc. ICASSP*, vol. 1, May 2001, pp. 85–88.

[19] ——, "Robust parameters for speech recognition based on subband spectral centroid histograms," in *Proc. EUROSPEECH*, vol. 1, Sep. 2001, pp. 591–594.

[20] K. K. Paliwal and B. S. Atal, "Representing frequencies in speech," AT&T Res. Labs., Florham Park, NJ, 2000.

[21] ——, "Frequency-related representation of speech," in *Proc. EUROSPEECH*, Sep. 2003, pp. 65–68.

[22] M. Blomberg, R. Carlson, K. Elenius, and B. Granström, "Auditory models in isolated word recognition," in *Proc. ICASSP*, Mar. 1984.

[23] R. A. Cole, Y. K. Muthusamy, and M. Fanty, "The ISOLET Spoken Letter Database," Oregon Graduate Inst. Sci. Technol., Beverton, Tech. Rep. CSE 90-004, Mar. 1990.

[24] P. Price, W. M. Fisher, J. Bernstein, and D. S. Pallett, "The DARPA 1000-word resource management database for continuous speech recognition," in *Proc. ICASSP*, vol. 1, Apr. 1988, pp. 651–654.

[25] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. Redmond, WA: Microsoft, 2000.

[26] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, pp. 247–251, 1993.


**Bojana Gajić** (M'99) received the Siv.Ing. (M.Sc.) and Dr.Ing. (Ph.D.) degrees in electrical engineering from the Norwegian University of Science and Technology, Trondheim, in 1996 and 2002, respectively.

From September 1995 to March 1996, she was a Visiting Scholar at Furui Laboratory, NTT, Tokyo, Japan, where she conducted the work on her master thesis in the field of speech detection in noise. From September 1999 to April 2000, she was a Visiting Research Scientist at AT&T Labs-Research, Florham Park, NJ, working on noise robust automatic speech recognition. During fall 2000, she was a visiting Research Fellow at Griffith University, Brisbane, Australia, working on feature extraction for noise robust speech recognition. She joined the Department of Telecommunications, Norwegian University of Science and Technology as a Postdoctoral Fellow in May 2002, and as an Associate Professor in January 2004. Her current research interest is in the field of robust automatic speech recognition.

**Kuldip K. Paliwal** (M'89) received the B.S. degree from Agra University, Agra, India, in 1969, the M.S. degree from Aligarh Muslim University, Aligarh, India, in 1971, and the Ph.D. degree from Bombay University, Bombay, India, in 1978.

He has been carrying out research in the area of speech processing since 1972. He has worked at a number of organizations including the Tata Institute of Fundamental Research, Bombay, the Norwegian Institute of Technology, Trondheim, the University of Keele, U.K., AT&T Bell Laboratories, Murray Hill, NJ, AT&T Shannon Laboratories, Florham Park, NJ, and the Advanced Telecommunication Research Laboratories, Kyoto, Japan. He joined Griffith University, Brisbane, Australia, in July 1993 as a Professor (Chair, Communication/Information Engineering) in the School of Microelectronic Engineering, where he is currently teaching subjects related to digital signal processing. His current research interests include speech recognition, speech coding, speaker recognition, speech enhancement, face recognition, image coding, pattern recognition, and artificial neural networks. He has published more than 200 papers in these research areas. He has coedited two books: *Speech Coding and Synthesis* (New York: Elsevier) and *Speech and Speaker Recognition: Advanced Topics* (Norwell, MA: Kluwer). He is currently the Editor-in-Chief of *Speech Communication*.

Dr. Paliwal received the IEEE Signal Processing Society's Best (Senior) Paper Award in 1995 for his paper on LPC quantization. He served the IEEE Signal Processing Society's Neural Networks Technical Committee as a founding member from 1991 to 1995 and the Speech Processing Technical Committee from 1999 to 2003. He was an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1994–1997 and 2003–2004). He also served as Associate Editor of the IEEE SIGNAL PROCESSING LETTERS (1997–2000). He was the General Co-Chair of the 10th IEEE Workshop on Neural Networks for Signal Processing (NNSP2000).