# Robust Parameters for Speech Recognition Based on Subband Spectral Centroid Histograms

*Bojana Gajić*[1,*] *and Kuldip K. Paliwal*[2]

[1]Department of Telecommunications, Norwegian University of Science and Technology
O.S. Bragstads plass 2B, 7034 Trondheim, Norway
[2]School of Microelectronic Engineering, Griffith University, Brisbane, QLD 4111, Australia
gajic@tele.ntnu.no, K.Paliwal@me.gu.edu.au

## Abstract

In this paper we propose a new speech parameterization framework that efficiently combines frequency and magnitude information from the short-term power spectrum of speech. This is achieved through computation of subband spectral centroid histograms (SSCH). Relationship between the proposed method and auditory based speech parameterization methods is discussed. An experimental study on an automatic speech recognition task has shown that the proposed method outperforms the conventional speech front-ends in presence of different types of additive noise, while it performs comparably in the noise-free conditions. In the case of car noise, our method also outperforms the computationally expensive auditory based methods, while having simplicity and low computational cost similar to the conventional front-ends.

## 1. Introduction

The goal of speech parameterization for automatic speech recognition (ASR) is to extract the important discriminative information from the speech signal, while omitting any information irrelevant for speech recognition. In the context of robustness against background noise, this means that we desire parameters that are invariant to changes in the acoustic environment, but retain good discriminative ability for recognizing speech.

Conventional speech parameterization methods are based on extracting information from the short-term power spectrum estimates of the speech signal. However, they utilize only magnitude information provided by power spectrum, while frequency information is left unexplored. For example in mel-frequency cepstral coefficients (MFCC), we use only the total power in each subband, while we do not keep track of the dominant subband frequencies.

Speech power spectrum changes in the presence of additive background noise. However, the positions of spectral peaks (formants) remain unaffected. Some early ASR studies have shown that formant frequencies could be used successfully to discriminate between different speech sounds. However, due to the lack of a reliable method for estimating formant frequencies, they could not be efficiently utilized in speech recognition. It has been shown in [1] that subband spectral centroids (SSC) are closely related to speech formants. Several attempts have recently been made to incorporate frequency information from the power spectrum into speech parameter vectors through computation of SSCs, and using them as additional features in the MFCC-based front-end [1, 2, 3, 4, 5]. The aim of our study was to find an effective method of combining the frequency and magnitude information from the power spectrum. This is achieved through computation of subband spectral centroid histograms (SSCH). In our initial work on this topic [6], we showed the efficiency of our method in presence of white Gaussian noise and discussed the choices of the free parameters. In this paper, we further develop the idea, show that delta and delta-delta parameters can successfully be used with SSCHs, and we present an evaluation of our method for various types of background noise. We also discuss the relationship between the proposed method and auditory based methods, and compare their performances.

The paper is organized as follows. We start with a description of the proposed method in section 2, and discuss its relationship with the auditory based methods in section 3. In section 4, we present an experimental study aimed at evaluating the robustness of the proposed method. Finally, the major conclusions are summarized in section 5.

## 2. Description of the Method

Speech parameterization in the proposed method is done on the frame-by-frame basis in the same way as in the conventional speech parameterization techniques. Each frame is represented by a set of discrete cosine transform (DCT) coefficients derived from the corresponding subband spectral centroid histogram. In the following, we give a short description of all the steps involved in this algorithm.

**Power spectrum estimation:** We start by computing a short-term spectral estimate of a speech frame in the same way as in the conventional front-ends. We have a choice between using FFT-based unsmoothed spectral estimates and LP-based spectrum envelopes. Intuitively, spectral envelopes seem to be better starting points for computing SSCs. However, since LP based spectral envelope estimates become unreliable in the presence of noise, the FFT-based power spectrum was used throughout this study.

**Centroid computation:** The power spectrum is divided into a number of overlapping frequency bands by a set of band-pass filters. Filters with rectangular frequency responses were used, as any other shape would favor some of the frequencies during centroid computation. For each subband signal the first moment or centroid is then computed. The operations in this step are summarized as

[*]A part of this research was conducted while B. Gajić was a visiting researcher at the Griffith University, Brisbane, QLD 4111, Australia. It was funded by Australian Research Council grant.

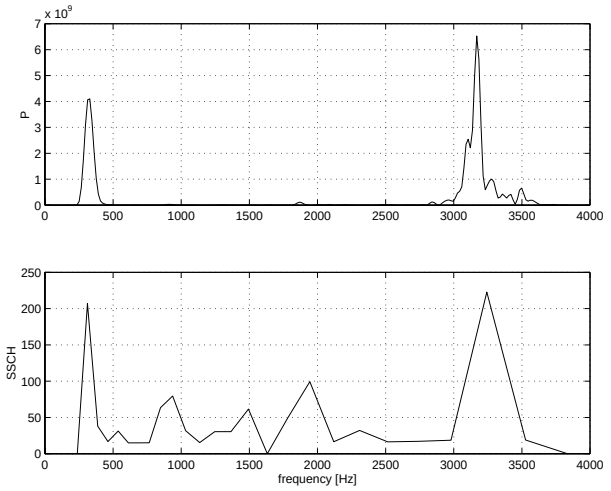


Figure 1: *FFT-based power spectrum estimate (P), and subband spectral centroid histogram (SSCH) for a 25 ms frame of sound /i/*

follows:

$$C_m = \frac{\sum_{k=0}^{N-1} k H_m(k) P^\gamma(k)}{\sum_{k=0}^{N-1} H_m(k) P^\gamma(k)} , \qquad (1)$$

where $P$ is the power spectrum estimate, $H_m$ is the frequency response of the m-th bandpass filter, $N$ is the available number of samples of the power spectrum in the interval $[0, \frac{F_s}{2}]$, and $\gamma$ is a parameter that decides the dynamic range of the spectrum used in the centroid computation. If $\gamma$ is too low (near 0), SSCs will be simply at the center of their subbands, and thus contain no information. If it is too large (near $\infty$), SSCs will correspond to locations of the peaks of the FFT-based power spectrum, and will thus be noisy estimates.

**Energy computation:** Next, we derive an energy measure associated with each centroid. One possibility is to compute the energy over the entire subband, in the same way as in the MFCC front-end. However, it could be advantageous to limit energy computation to the frequency range closer to the centroid, as that area is less prone to the addition of background noise than the low-energy parts of the spectrum. Both of the methods have been tested in our experimental study described in Section 4.

**Histogram construction:** The entire signal frequency range is divided into number of non-overlapping bins. Then, for each centroid, a corresponding bin is found, and its count is increased by the corresponding energy measure.

**Decorrelation:** The last step involve computation of DCT coefficients for the histogram. This is done in order to obtain a set of less correlated parameters, that are more suitable for use with hidden Markov model framework with diagonal correlation matrices.

In order to get a better intuitive understanding for the steps involved in the SSCH computation, Figure 1 illustrates the FFT-based power spectrum and the SSCH for a single frame of sound /i/.

## 3. Relationship to auditory methods

In this section, we discuss the relationship between the proposed parameterization method and the methods derived by mimicking the processes in the human auditory system. In particular we look at the steps involved in the Zero Crossings with Peak Amplitudes (ZCPA) method [7], that is a modification of the original auditory based method known as Ensemble Interval Histograms (EIH) [8]. A comparative study described in [7] has shown that ZCPA greatly outperformed all of the widely used feature extraction methods (LPCC, MFCC, SBCOR, PLP) for different types of background noise.

In the ZCPA method, the speech signal is passed through a filter-bank of bandpass filters, to generate a set of subband signals. The signals are then processed on the frame-by-frame bases with frame lengths inverse proportional to the center frequencies of the corresponding bandpass filters. For a given frame, all positive zero crossings are found and the interval between each pair of successive zero crossings is measured. Then, a histogram of the inverse of the intervals is collected over all the subband signals, with the histogram increments equal to the peak amplitudes between the corresponding pairs of zero crossings. At the end, DCT is performed on the resulting histogram.

If we compare the steps involved in ZCPA and SSCH, we can notice the following similarities between the two methods:

- The inverse of an interval between successive positive zero crossings can be interpreted as the instantaneous dominant frequency in the subband signal. Similarly, an SSC is an estimate of the dominant subband frequency derived from the frequency domain.
- The peak amplitude between successive zero crossings, is a measure of the instantaneous energy in the subband signal. This is similar to the energy measure associated with each centroid in the SSCH method.
- Thus, the ZCPA histograms have their maxima at the dominant signal frequencies, in the same way as the SSC histograms.

We conclude, that both the methods utilize the same conceptual information from the speech signal, but the algorithms used to extract this information are different. The major difference are summarized in the following:

- ZCPA operates entirely in the time domain, while all the computations in SSCH are done in frequency domain.
- ZCPA operates with instantaneous values, while SSCH uses averaged values over a speech frame.
- ZCPA utilizes frequency dependent window lengths, while SSCH uses equal frame length for all of the subband signals.
- The major problem with using ZCPA in practical applications is its prohibitively high computation cost compared to the conventional algorithms. This is due to its operation in the time domain, and the need for heavy interpolation of the high-frequent subband signals that is necessary for proper operation of the algorithm. On the other hand, computational cost associated with the SSCH method is in the same order as for the conventional (MFCC-type) methods.

# 4. Experimental Study

## 4.1. Task and database

The proposed method was evaluated on the ISOLET Spoken Letter Database [9] down-sampled to 8 kHz. The database consists of English letters spoken in isolation recorded in a quiet room. Two repetitions of each word were recorded for each speaker. Utterances from 90 speakers (subsets ISOLET-1, ISOLET-2 and ISOLET-3) were used for training, while utterances from 30 speakers (subset ISOLET-5) were used for evaluation. Although the vocabulary consisting of 26 English letters is rather small, this is not a simple recognition task, since the vocabulary words are very short and highly confusable.

For the purpose of evaluating the robustness of the proposed algorithm against different types of background noise, four different noise types were added to the test set at four different signal-to-noise ratios (SNR). Those are: white Gaussian noise, factory noise, car noise and background speech. The last three noise types were taken from the NOISEX database, where they were referred to as factory1, volvo and babble noise respectively. A segment of the noise file equal to the length of the speech file was randomly extracted and added to the speech file at the required SNR. SNR was computed as the ratio between the maximal frame energy of the speech file, and the average energy of the noise segment. This way of computation makes SNR independent of the duration of the surrounding silence in the speech files.

## 4.2. Choice of parameters

Speech parameterization methods depend on a number of free parameters. This section describes the parameter choices made for all the methods evaluated in our experimental study (SSCH, MFCC and ZCPA). They all operate on the frame-by-frame basis, with the frame shift equal to 10ms, which results in computation of 100 parameter vectors per second of speech file.

### *4.2.1. MFCC*

MFCC parameters were extracted using the HTK tool HCopy. Frame length was set to 25 ms. A first order preemphasis filter and the Hamming window were applied to each frame before the DFT computation. The resulting spectrum was passed through a mel-spaced filter-bank with overlapping triangular filters, and energy was computed for each subband signal. Finally, 12 cepstrum coefficients were derived from the set of the subband energies.

### *4.2.2. SSCH*

**Power spectrum estimation:** The initial steps in SSCH computation are identical to those in the MFCC computation. The 25 ms long speech frames were passed through a preemphasis filter, and multiplied with the Hamming window before the DFT computation. Samples of the power spectral estimate for the given frame were obtained by squaring the DFT coefficients.

**Dynamic range:** We experimented with several values of parameter $\gamma$ in Eq. (1), that controls the dynamic range of the power spectrum used in the centroid computation. Table 1 shows the recognition performance for different values of parameter $\gamma$ in the presence of white Gaussian noise. The best results were obtained using $\gamma = 1$, and we used this value in the rest of our study.

Table 1: Comparison of recognition rates for different values of the dynamic range parameter $\gamma$ in presence of white Gaussian noise

| $\gamma$ | SNR [dB] | | | | |
|---|---|---|---|---|---|
| | clean | 25 | 20 | 15 | 10 |
| 0.5 | 88.78 | 78.97 | 69.29 | 55.83 | 31.60 |
| 1 | 88.01 | 78.08 | 70.32 | 57.31 | 36.67 |
| 2 | 86.79 | 75.45 | 67.63 | 55.83 | 37.18 |
| 4 | 84.49 | 73.01 | 66.09 | 53.65 | 34.10 |

**Filterbank:** Filter bandwidths and placements have been investigated in detail in our earlier study [6], where we concluded that choice of the filter bandwidths was not very critical. We also showed that linear filter spacing on the hertz scale performed better than the linear spacing on the bark scale, but was also computationally more expensive. As a compromise, we decided to use linear spacing on the hertz scale with 300 Hz bandwidths in the law frequency range, and continue with the linear spacing on the bark scale with the bandwidths equal to 2 Bark. Number of filters should be chosen sufficiently large to provide enough points in the histogram. On the other hand, the computational cost increases proportionally with the number of filters. We found that average bin count of 2.5 was a reasonable choice. This results in 65 filters when number of bins is set to 26. (The reasoning for this choice is given bellow.)

**Frequency bins:** In order for centroids to provide any useful information, each filter must stretch over several frequency bins. Thus, it is of crucial importance that the ratio between filter and bin bandwidths is chosen sufficiently high. For given filter bandwidths, this is achieved by increasing the total number of frequency bins. However, too small bins might cause histograms to become too sensitive to small fluctuations of the spectral peak positions. Our earlier investigation of this issue [6] has shown that the best performance is obtained by choosing the ratio between filter and bin bandwidths to be close to 4. This results in 26 frequency bins, and this number was chosen for the rest of our study. The final parameter vectors consist of the 12 DCT coefficients extracted from the 26 bin counts.

**Energy computation:** Table 2 compares recognition performances for the two energy computation methods discussed in Section 2 in presence of background speech (babble noise). The first row is for the case of energies computed over the entire subband, while the second row is for the case of energies computed over one half of the critical bandwidth around the subband centroid. Since the energy measure computed over the narrow frequency range around centroids is more robust to additive noise, it was used in the rest of our experimental study.

Table 2: Comparison of recognition rates for different energy computation methods in presence of background speech

| Frequency range | SNR [dB] | | | | |
|---|---|---|---|---|---|
| | clean | 20 | 15 | 10 | 5 |
| Entire subband | 88.01 | 72.56 | 58.08 | 39.04 | 22.76 |
| 0.5 * critical bw | 87.24 | 73.46 | 60.38 | 40.51 | 23.01 |

### *4.2.3. ZCPA*

For computation of ZCPA, the speech signal was first passed through a filter-bank consisting of 20 bandpass FIR filters lin-

early spaced on the bark-scale, with bandwidths equal to two times critical bandwidth. The filters had order 61, and were designed using the windowing method. The number of frequency bins was set to 26. Frequency dependent frame lengths equal to $20/f_c$ were used, where $f_c$ is the center frequency of the corresponding bandpass filter. We used both positive and negative zero-crossings, since this gave better performance than using only positive zero-crossings. Number of DCT coefficients extracted from the histogram bin counts was 12.

### 4.3. Comparison with other front-ends

In this section we present the results of a comparative study between the SSCH-based method, the most commonly used parameterization method, MFCC, and the auditory based method, ZCPA.

In all the methods, we used the speech recognition toolkit HTK [10] to augment delta and delta-delta parameters to the original parameter vectors. This resulted in a 36 dimensional parameter vectors for all the three methods. One hidden Markov model (HMM) with 5 states and 5 Gaussian mixtures per state was trained for each vocabulary word. Both training and test were performed using HTK. Table 3 shows the results of an evaluation of MFCC, SSCH and ZCPA in presence of white noise, car noise, factory noise and background speech respectively.

Table 3: Comparison of recognition rates for different feature extraction methods in various noisy environments

a) White Gaussian noise

| Parameterization method | SNR [dB] | | | | |
|---|---|---|---|---|---|
| | clean | 25 | 20 | 15 | 10 |
| MFCC | 89.55 | 76.86 | 67.44 | 48.33 | 17.44 |
| SSCH | 87.24 | 78.91 | 70.58 | 57.69 | 38.21 |
| ZCPA | 85.19 | 76.68 | 71.28 | 62.37 | 48.08 |

b) Car noise

| Parameterization method | SNR [dB] | | | | |
|---|---|---|---|---|---|
| | clean | 20 | 10 | 0 | -5 |
| MFCC | 89.55 | 81.15 | 69.87 | 46.54 | 22.37 |
| SSCH | 87.24 | 86.92 | 86.15 | 81.35 | 72.69 |
| ZCPA | 85.19 | 85.19 | 82.31 | 73.27 | 61.47 |

c) Factory noise

| Parameterization method | SNR [dB] | | | | |
|---|---|---|---|---|---|
| | clean | 20 | 15 | 10 | 5 |
| MFCC | 89.55 | 78.78 | 66.99 | 46.35 | 21.09 |
| SSCH | 87.24 | 79.36 | 71.79 | 54.36 | 35.96 |
| ZCPA | 85.19 | 78.65 | 71.67 | 59.87 | 37.31 |

d) Background speech

| Parameterization method | SNR [dB] | | | | |
|---|---|---|---|---|---|
| | clean | 20 | 15 | 10 | 5 |
| MFCC | 89.55 | 73.14 | 57.56 | 39.04 | 22.18 |
| SSCH | 87.24 | 73.46 | 60.38 | 40.51 | 23.01 |
| ZCPA | 85.19 | 76.22 | 67.44 | 50.19 | 30.32 |

We see that SSCH is significantly more robust than MFCC for all the noise types. The improvements are largest for the car noise, and smallest for the background speech. The relatively poor performance for the case of background speech is probably due to the existence of speech-like spectral peaks in the background signal.

SSCH even outperforms the ZCPA in the case of car noise, while ZCPA is more robust in presence of the other noise types. This might be due to the differences between the two algorithms listed in Section 3. However, it is important to note that ZCPA could not be used in place for SSCH in practical applications, due to its prohibitively high computation cost.

## 5. Conclusions

In this paper we proposed a new speech parameterization framework that efficiently combines magnitude and frequency information from the power spectrum. This was achieved through computation of subband spectral centroid histograms. We have shown that the conceptual information used in the proposed method is similar to that of the auditory based methods, while the computational complexity is similar to that of the conventional (MFCC-type) front-ends.

In an evaluation on an ASR task, the proposed method greatly outperformed the conventional speech parameterization methods in presence of different types additive noise. We have also shown that the concept of delta and delta-delta coefficients can be used efficiently with this framework.

## 6. References


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

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

[3] D. Albesano, R. D. 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, pp. 1503–1506, September 1999.

[4] R. D. Mori, D. Albesano, R. Gemello, and F. Mana, "Ear-model derived features for automatic speech recognition," in *Proc. ICASSP*, 2000.

[5] E. Gjelsvik, "Modification of front-end processing for robust speech recognition." Diploma thesis, Norwegian University of Science and Technology, June 1999.

[6] B. Gajić and K. K. Paliwal, "Robust feature extraction using subband spectral centroid histograms," in *Proc. ICASSP*, May 2001.

[7] 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. on Speech and Audio Processing*, vol. 7, pp. 55–69, January 1999.

[8] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 115–132, January 1994.

[9] R. A. Cole, Y. K. Muthusamy, and M. Fanty, "The ISOLET spoken letter database," Technical report CSE 90-004, Oregon Graduate Institute of Science and Technology, Beverton, OR, USA, March 1990.

[10] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*. Entropic, 1999.