ROBUST FEATURE EXTRACTION USING SUBBAND SPECTRAL CENTROID HISTOGRAMS

Bojana Gajić and Kuldip K. Paliwal*

School of Microelectronic Engineering, Griffith University, Brisbane, QLD 4111, Australia
E-mail: gajic@tele.ntnu.no, K.Paliwal@me.gu.edu.au

ABSTRACT

In this paper we propose a new framework for utilizing frequency information from the short-term power spectrum of speech. Feature extraction is based on the cepstral coefficients derived from the histograms of subband spectral centroids (SSC). Two new feature extraction algorithms are proposed, one based on frequency information alone, and the other which efficiently combines the frequency and amplitude information from the speech power spectrum. Experimental study on an automatic speech recognition task has shown that the proposed methods outperform the conventional speech front-ends in presence of additive white noise, while they perform comparably in the noise-free conditions.

1. INTRODUCTION

Signal parameterization techniques used for speech recognition are based on extracting information from the short-term power spectrum estimates of speech. However, they utilize only amplitude information provided by power spectrum, while the frequency information is left unexplored. For example in MFCC, we use only the information on the total power in each subband, but we do not keep track of the dominant subband frequencies.

Several attempts have recently been made to incorporate the frequency information from the power spectrum in the speech feature vectors [1, 2, 3, 4, 5]. They are based on computing subband spectral centroids (SSC) and using them as additional features in the MFCC-based front-end. It has been shown in [1] that SSCs are closely related to position of spectral peaks (formants) of speech sounds. Since spectral peak positions remain practically unaffected in presence of additive noise, it is expected that an SSC-based front-end would have a potential of improving the robustness of automatic speech recognition (ASR) systems.

The aim of this study was to find an effective way of utilizing the frequency information from the power spectrum, both alone and in combination with the amplitude information. It has been achieved through computing the histograms of the SSCs.

This paper is organized as follows. In Section 2 we define the SSCs, and propose two methods for their use in feature extraction. A discussion of the choice of free parameters is given in Section 3. Section 4 presents the results of an evaluation of the methods on an ASR task. Finally, Section 5 highlights some important aspects connected to the proposed methods and gives the main conclusions.

2. SSC HISTOGRAMS

Subband spectral centroids are found by applying a filterbank to the power spectrum of the signal, and then calculating the first moment (or centroid) for each subband. Thus, the SSC of the m-th subband can be computed as

$$C_m = \frac{\int_0^{F_s/2} f W_m(f) P^\gamma(f) df}{\int_0^{F_s/2} W_m(f) P^\gamma(f) df} \qquad (1)$$

where F_s is signal sampling frequency, $P(f)$ is short-term power spectrum, $W_m(f)$ is frequency response of the m-th bandpass filter, and γ is a constant used for controlling the dynamic range of the power spectrum.

In the following we propose two feature extraction algorithms based on accumulating the SSC values across all subbands of a given speech frame into a single histogram.

Method 1 This method, referred to as SSCH, consists of the following steps:

1. Estimate power spectrum.
2. Apply a filterbank to divide the power spectrum into a number of overlapping frequency bands.
3. Find the centroid for each subband.
4. Partition the entire signal frequency range into a number of bins.
5. Find the corresponding bin for each centroid and increase its count by one.
6. Calculate cepstral coefficients by computing the DCT of the histogram.

*This research was conducted while B. Gajić was on leave from Norwegian University of Science and Technology (NTNU), Trondheim, Norway. It was funded by Australian Research Council grant.

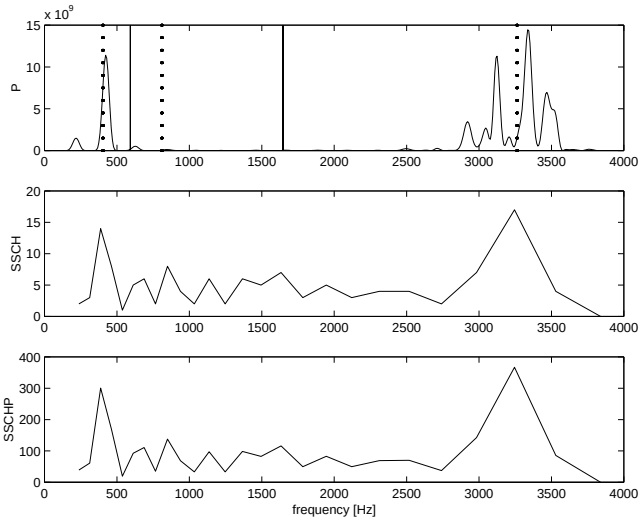

Fig. 1. Different representations of a 25 ms frame of sound /ee/: a) FFT-based power spectrum with three SSCs. SSCs are shown with dotted vertical lines, and subband boundaries with solid vertical lines. b) SSCH and c) SSCHP

Method 2: This method differs from the first one only in step 5. Instead of increasing bin counts by one, they are increased by $\log(1 + P_k/bw_k)$, where P_k is the power of the k-th subband signal given by the denominator of the expression in Eq. 1, bw_k is the bandwidth of the k-th bandpass filter in the filterbank, and the unity term is added to prevent from adding negative values to the bin counts. In this way power information is efficiently incorporated into SSC histograms. We refer to this algorithm as Power Weighted SSC Histogram (SSCHP).

Figure 1 shows four different representations of a 25 ms frame of sound /ee/. In the first plot, the FFT-based power spectrum estimate is shown together with the three SSCs (dotted vertical lines) obtained using a rectangular Bark-spaced filterbank (filter cut-off frequencies are shown as solid vertical lines). The second and third plots show the SSCH and SSCHP, respectively. We observe a close relationship between the histogram representations and the power spectrum. This is especially interesting for the SSCH, since it does not explicitly use any power information.

3. CHOICE OF PARAMETERS

The algorithms proposed in Section 2 depend on a number of free parameters. In this section, we present a discussion on the choice of the parameters, while Section 4.2 presents the results of an experimental study of the effect of different parameter choices on the recognition performance.

Power spectrum estimation method: We have a choice between using FFT-based unsmoothed spectrum 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.

Dynamic range: The dynamic range of the spectrum used in the computation of SSCs is controlled by parameter γ (as shown in Equation 1). If γ 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 ∞), SSCs will correspond to locations of the peaks of the FFT-based power spectrum, and will thus be noisy estimates.

Filterbank: We decided to use filters with rectangular frequency responses in this study. Any other shape (such as triangular) would favor some frequencies within the subband more than the others, and thus give an biased SSC estimate.

Filter bandwidths have to be large enough to suppress appearance of local peaks in the histogram. However, they should be small enough to prevent inclusion of more than one formant into a single subband.

Number of filters should be chosen sufficiently large to provide enough points in the histogram. Too low bin counts lead to unreliable histograms. On the other hand, the computational cost increases proportionally with number of filters.

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.

Filter and bin placement: Filters can be distributed linearly along the Hertz scale or along a perceptually based frequency scale such as Bark or Mel. In any case, it is important that bin distribution is in accordance with the filter distribution in order to obtain unbiased histograms.

4. EXPERIMENTAL STUDY

This section is divided into three parts. First we describe the recognition task used for performance evaluation of the pro-

posed algorithms. Then, we present the results of an experimental study of the effect of different parameter choices on recognition performance. Finally, we compare the recognition performance of the SSC-based feature extraction methods with the conventional front-ends, both in clean and noisy environments.

4.1. Task and database

The proposed methods were evaluated on an isolated word, speaker independent task, with the vocabulary consisting of 26 letters from English alphabet. Two repetitions of each word were recorded for each speaker. Speakers were divided into two sets, 90 for training and 30 for testing. One hidden Markov model (HMM) with 5 states and 5 Gaussian mixtures was used to model each vocabulary word. Both training and testing were performed using the speech recognition toolkit HTK. Although the vocabulary is relatively small, this is a rather difficult task as all vocabulary words are very short and highly confusable. The baseline performance for MFCC and LPCC front-ends, measured as word accuracy (WAC), is given in Table 1.

Table 1. Baseline performance

Method	WAC	Parameters
MFCC	77.76	12 cep. coeff
LPCC	74.49	12 cep. coeff.

4.2. Experimenting with parameter values

In the following we present the results of an experimental study aimed at finding the effect of different parameter values on the recognition performance. All the experiments were done using the SSCH method, but it is reasonable to expect that they would generalize to SSCHP method too.

First, we investigated the importance of choosing the bin bandwidths sufficiently small compared to the filter bandwidths. This was achieved by varying the total number of bins, while keeping filter bandwidths constant. Recognition performance for different choices of the ratio between filter and bin bandwidths is shown in Table 2. We observe that choosing the ratio close to 4 gives the best performance. Too small values degrade the performance considerably, which is in agreement with the discussion in Section 3. We repeated the experiment for different choices of filter bandwidths, and found out that choosing the ratio between 3 and 5 always maximized the performance.

Table 2. Performance for different choices of ratio between filter and bin bandwidths

Filter BW [Bark]	# Bins	Filt BW/Bin BW	WAC
2	16	2.1	69.87
2	21	2.8	71.92
2	30	4	74.29
2	50	6.2	73.78

Next, we investigated the influence of filter bandwidths to the performance of SSC histograms. The results are summarized in Table 3. We conclude that the choice of filter bandwidths is not critical as long as the number of bins is adjusted to achieve an appropriate ratio between filter and bin bandwidths. Similar results were obtained when filters were uniformly spaced along the Hertz scale. We observed little change in performance for filter bandwidths between 200 Hz and 400 Hz, with gradual decrease in performance with further increase of the bandwidths.

Table 3. Performance for different filter bandwidths

Filter BW [Bark]	# Filters	FiltBW/BinBW	WAC
1	200	3	73.46
2	150	4	74.29
3	100	4	73.72

Next, we investigated the importance of parameter γ in Equation 1, that determines the dynamic range of the power spectrum used in SSC computation. Setting γ to 0.5, 1 and 1.5 led to the similar recognition performance, with the case of $\gamma = 1$ being slightly better than the other two. Thus, $\gamma = 1$ was used in all further experiments.

At the end, we compared the performance for uniform filter spacing along the Hertz and Bark scales. The first two rows of Table 4 show the best performances achieved using Bark and Hertz scales respectively. In an attempt to retain the good performance of the Hertz scale, and the low computational cost of the Bark scale, we combined the two scales by applying the Hertz scale in the low frequencies and the Bark scale in the high frequencies. The boundary between the two scales is chosen to provide the smooth transition. The recognition performance for the combined approach is shown in the last row of Table 4. As it gives the best compromise between recognition performance and computational cost, it has been used in all further experiments.

Table 4. Performance for different filter spacings

Scale	FiltBW	# Filters	# Bins	WAC
Bark	2 Bark	150	30	74.29
Hertz	308 Hz	247	50	76.15
Hz/Bark	300 Hz/2 Bark	130	26	75.83

4.3. Comparison with other front-ends

In this section we present the results of a comparative study between conventional and SSC based speech recognition front-ends both in clean and noisy environments. Noisy speech was produced by adding samples of white Gaussian noise to the clean speech at given signal-to-noise ratios (SNR). SNR was computed by dividing the total power of the clean utterance by the noise variance. Recognition performance was compared for the following front-ends:

- 12 LPCCs derived from 12 LP coefficients.
- 12 MFCCs derived from 24 Mel-filterbank log-energies.
- 3 SSCs derived from the FFT-based power spectrum using three rectangular non-overlapping Bark-spaced filters, as described in [1].
- 12 cepstrum coefficients derived from SSCH with parameter values given in the last row of Table 4.
- 12 cepstrum coefficients derived from SSCHP with parameters same as above.

Table 5 summarizes the recognition performances of the five frontends both on clean speech, and for four different noise levels. Comparing the two conventional front-ends, we see

Table 5. Performance comparison of different recognition front-ends on clean and noisy speech

	SNR [dB]				
Method	clean	20	15	10	5
LPCC	74.49	60.32	46.41	27.50	14.04
MFCC	77.76	66.28	54.29	34.94	16.92
SSC	59.10	40.00	31.09	24.17	15.45
SSCH	75.83	65.77	57.56	41.35	23.91
SSCHP	76.06	67.96	61.22	47.37	28.46

that MFCC outperforms LPCC in all test condition. The difference is especially pronounced in moderate noisy conditions. Further, it was interesting to see a surprisingly good performance obtained using only three SSCs. However, it was still much poorer than that of the conventional front-ends. On the other hand, SSC histogram based methods were proven to be much more efficient in utilizing SSC information than using SSC as features directly. Both, SSCH and SSCHP were shown to be more robust than MFCC in presence of additive white noise, while exhibiting only a slight decrease in performance in the noise-free conditions. SSCHP consistently outperformed SSCH in all test conditions. This is not surprising, since SSCHP incorporates both frequency and amplitude information from the power spectrum. The relative error reduction obtained by SSCHP with respect to MFCC was up to 19%.

5. DISCUSSION AND CONCLUSIONS

In this paper we proposed a new framework for efficient utilization of frequency information from the power spectrum in the speech feature extraction. It is achieved through computation of subband spectral centroid histograms. Two different methods were proposed. The first one, SSCH, uses frequency information alone, while the second one, SSCHP, incorporates amplitude information in addition to the frequency information into the histograms.

In an evaluation on an ASR task the proposed methods outperformed the conventional feature extraction methods in presence of additive white noise. This was particularly pronounced for SSCHP.

It should be noted that the conventional front-ends (MFCC and LPCC) utilize only amplitude information from the speech power spectrum, while the proposed front-ends (SSCH and SSCHP) utilize frequency information derived from the power spectrum. The SSCH front-end uses only frequency information and has performed as well as the MFCC and LPCC front-ends for clean speech. For noisy speech, it has given better results. This is very satisfying as it derives the frequency information from the power spectrum which is already corrupted due to additive noise distortion in the speech signal. If we had used a robust estimation method to derive the frequency information directly from the speech signal [6, 7], this front-end would have resulted much better performance for noisy speech. We are currently investigating the use of such robust frequency estimation methods.

6. REFERENCES

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

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

[3] Dario Albesano, Renato De Mori, Roberto Gemello, and Franco Mana, "A study of the effect of adding new dimensions to trajectories in the acoustic space," in *Proc. EUROSPEECH*, September 1999, vol. 4, pp. 1503–1506.

[4] Renato De Mori, Dario Albesano, Roberto Gemello, and Franco Mana, "Ear-model derived features for automatic speech recognition," in *Proc. ICASSP*, 2000.

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

[6] Oded Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 1, pp. 115–132, January 1994.

[7] Doh-Suk Kim, Soo-Young Lee, and Rhee 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, no. 1, pp. 55–69, January 1999.