ROBUST SPEECH RECOGNITION USING FEATURES BASED ON ZERO CROSSINGS WITH PEAK AMPLITUDES

Bojana Gajić

Norwegian University of Science and Techn.
Department of Telecommunications
7491 Trondheim, Norway
gajic@tele.ntnu.no

Kuldip K. Paliwal

Griffith University
School of Microelectronic Engineering
Brisbane, QLD 4111, Australia
K.Paliwal@me.gu.edu.au

ABSTRACT

This paper presents an extensive study of zero crossings with peak amplitudes (ZCPA) features, that have earlier been shown to outperform both conventional and auditory-based features in presence of additive noise. The study starts by optimizing different parameters involved in ZCPA feature computation, followed by a comparison of ZCPA and MFCC features on two recognition tasks in different background conditions. The main differences between the two feature types were identified, and their individual effects on ASR performance were evaluated. The importance of a proper choice of analysis frame lengths and filter bandwidths in ZCPA feature extraction was demonstrated. Furthermore, the use of dominant frequency information in ZCPA features was found to be a major reason for increased robustness of ZCPA features compared to MFCC features.

1. INTRODUCTION

Features based on Zero Crossings with Peak Amplitudes (ZCPA) proposed by Kim at al. [1] evolved as a modification of the EIH auditory model [2]. They are computed by passing a speech frame through a subband filter bank, and finding all positive-going zero crossings for each subband signal. Then, for each pair of successive zero crossings the inverse interval length between the zero crossings is computed, as well as the peak signal value on the interval. Next, a single histogram of the inverse zero-crossing interval lengths is collected over all subband signals. However, instead of increasing the histogram bin counts by one, they are increased by the logarithm of the corresponding signal peak values. Finally, DCT is performed on the histogram for decorrelation purposes.

The dominant frequency principle [3] states that if there is a significantly dominant frequency in the signal spectrum, then the inverse zero-crossing interval lengths tend to take values in the vicinity of the dominant frequency. Thus, the inverse zero-crossing interval lengths of a subband signal can be seen as estimates of the dominant subband frequency. Furthermore, the peak signal value between subsequent zero crossings can be seen as a measure of signal power in the subband signal. Consequently, the construction of ZCPA histograms consists of assigning subband power estimates to frequency bins corresponding to dominant subband frequencies. Standard MFCC method, on the other hand, assigns subband power estimates to entire subbands, without taking into account the power distribution within subbands. Thus, the ZCPA representation can be seen as an alternative spectral representation of speech that emphasizes spectral peaks, while deemphasizing the information in spectral valleys, which is usually corrupted by noise.

In the study presented in [1], ZCPA features demonstrated greater robustness then LPCC, MFCC, PLP, SBCOR and EIH features in different background conditions. However, no attempt to optimize parameters involved in ZCPA computation was reported. Furthermore, the comparison between different feature types was done only on a small-vocabulary isolated-word database. In the study described in this paper, we investigated the influence of different parameter choices on the ZCPA performance. Furthermore, we compared the performance of ZCPA and MFCC methods on two different recognition tasks. Finally, we studied the individual effects of three main differences between the two feature types, in order to explain the difference in their overall performance.

2. RECOGNITION TASK

Two different recognition tasks were used for evaluating the ASR performance in this study. The first one is a small-vocabulary isolated-word task based on ISOLET Spoken Letter Database [4]. The second one is a medium-vocabulary continuous-speech task based on the speaker-independent part of DARPA Resource Management (RM) database [5]. In order to evaluate the robustness of ZCPA features against background noise, three different noise types were added to

the test data at several SNRs, namely, white Gaussian noise, factory noise and babble noise. Detailed description of the recognition systems, noise characteristics, and the noise addition algorithm can be found in [6].

3. OPTIMIZING PARAMETER VALUES

The computation of ZCPA features involves a number of free parameters. In this study, we investigated the influence of the choice of analysis frame lengths, subband filter bank and histogram bin allocation on the ASR performance of ZCPA features. The experimental study was performed on the ISOLET database, both on clean speech and in presence of white Gaussian noise added at different SNRs.

3.1. Analysis frame lengths

Three different methods for allocating analysis frame lengths to the subband signals were compared. In the first method, the frame length of the k-th subband signal (given in ms) was computed as C/F_{c_k}, where F_{c_k} is the center frequency of the corresponding bandpass filter given in kHz, and C is a constant. Four different values of parameter C were tested: 10, 20, 30 and 40. Corresponding frame lengths and recognition results are presented in the first part of Table 1. Note that low values of parameter C lead to frame lengths for high-frequency subbands that are shorter than the average pitch period, which leads to unreliable frequency estimates. Higher values of parameter C, on the other hand, lead to too long frames for low-frequency subbands, that can cause obstruction of the stationarity assumption.

The goal of the second method was to increase the frame lengths at high frequencies without making the frames at low frequencies unreasonably long. The frame length of the k-th subband signal (given in ms) was computed as $C/\sqrt{F_{c_k}}$. Four different values of constant C were tested: 20, 40, 60 and 80. The corresponding frame lengths and recognition results are presented in the second part of Table 1.

In the third method, equal analysis frame lengths were used for all subband signals. The third part of Table 1 presents the recognition results for five different choices of frame lengths: 25 ms, 35 ms, 50 ms, 75 ms and 100 ms.

Note that in the last two methods high-frequency subband signals contribute to more histogram points than low-frequency subband signals. In order to avoid histogram biasing toward higher frequencies, histograms were normalized with respect to frequency.

The results in Table 1 show the importance of a proper choice of analysis frame lengths. The use of relatively long frames, especially in low-frequency subbands, led to a considerable increase in ASR performance in presence of noise. The best results were achieved using frame lengths determined by the second method with $C = 60$.

Table 1. ASR performance of ZCPA features for different choices of analysis frame lengths

Method	Frame length [ms]	Word accuracy [%]				
		No noise	SNR [dB]			
			25	20	15	10
1. C=10	3–50	78.3	70.5	61.8	55.6	40.6
1. C=20	6–100	81.2	73.8	67.6	61.4	48.0
1. C=30	9–150	80.5	75.5	72.6	65.0	52.2
1. C=40	12–200	79.1	75.9	72.4	65.3	51.8
2. C=20	11–45	81.1	73.8	68.2	57.8	44.7
2. C=40	22–89	82.9	77.8	72.7	65.1	51.5
2. C=60	33–134	82.2	78.1	74.1	68.1	54.7
2. C=80	43–179	81.2	78.2	74.5	68.1	54.3
3	25	80.8	71.9	64.9	55.8	41.2
3	35	81.2	73.8	68.5	59.2	44.7
3	50	81.4	75.4	71.2	62.1	49.1
3	75	80.5	75.9	72.4	65.1	52.8
3	100	80.5	76.9	72.9	66.8	55.6

3.2. Filter bank and histogram bin allocation

The filter bank used in this study consisted of 16 FIR Hamming filters of order 61 uniformly spaced on the Bark scale. This choice was motivated by the results in [1] which showed that ZCPA features based on a similar filter bank consistently outperformed the features based on carefully designed cochlear filters. Filter bandwidths should ideally be chosen such that each subband contains exactly one dominant spectral peak. In this case, the inverse zero-crossing interval lengths serve as good estimates of spectral peak locations.

Frequency resolution of ZCPA histograms is determined by frequency bin widths. In order to accurately locate dominant subband frequencies, bin widths should be small compared to subband bandwidths. On the other hand, too narrow bins would make ZCPA features too sensitive to random variations in spectral peak positions. In this study, histogram bins having equal lengths on the Bark scale were used. This provides better frequency resolution at low frequencies than at high frequencies, which is in agreement with human speech perception.

Table 2 shows the ASR performance obtained by several different choices of filter bandwidths and number of histogram bins. We observe that the choice of filter bandwidths had a significant influence on the ASR performance of ZCPA features, while it was not very sensitive to the particular choice of the number of histogram bins.

Finally, the influence of increased number of filters was tested by evaluating ZCPA features based on 20 filters. However no significant performance difference was observed.

Table 2. ASR performance of ZCPA features for different choices of filter bandwidths and number of histogram bins

Filter bw [Bark]/ # bins	Word accuracy [%]				
	No noise	SNR [dB]			
		25	20	15	10
1/30	75.5	69.8	67.0	60.3	49.4
1/60	74.7	69.5	67.7	60.4	48.1
2/30	81.2	76.2	71.0	63.9	50.5
2/45	80.7	75.1	71.3	63.4	52.2
2/60	80.5	75.5	72.6	65.0	52.2
2/75	79.8	74.7	70.8	63.7	51.3
3/20	82.6	76.8	71.3	61.9	45.0
3/30	82.5	77.4	72.7	62.6	47.4
3/40	82.0	77.2	71.4	62.5	47.9

4. COMPARING PERFORMANCE OF ZCPA AND MFCC FEATURES

In this section the performance of ZCPA and MFCC features is compared on ISOLET and RM databases in different background conditions. MFCC were computed in the standard way, using 25 ms frame length and 20 triangular bandpass filters, while ZCPA features were computed using the best parameter choices found in Section 3. Frame lengths were set to $60/\sqrt{F_{c_k}}$, filter bank consisted of 16 Hamming FIR filters of order 61, with center frequencies uniformly distributed on the Bark scale between 200 Hz and 3400 Hz. The bandwidths of the ideal prototype filters were equal to 2 Bark. Frequency range between 0 and 4000 Hz was partitioned into 60 histogram bins uniformly distributed on the Bark scale. Both feature vectors consisted of 12 static features and corresponding delta and delta-delta features.

Figure 1 shows the absolute difference in word accuracy between ZCPA and MFCC features as a function of SNR for the three different noise types on ISOLET and RM databases respectively. We observe that MFCC features outperformed ZCPA features at high SNRs, while ZCPA features were better at low SNRs. The advantage of ZCPA features generally increased with reduced SNR. This is in agreement with earlier results [1] that showed greater robustness of ZCPA features. The advantage of ZCPA features in noisy conditions was largest on ISOLET task and in presence of white noise.

ZCPA and MFCC features used in this study differ in three main aspects: they are based on different subband filter banks, they are derived in time domain rather than frequency domain, and they combine dominant subband frequency information with subband power information, rather than using subband power information alone.

In order to determine the effect of each of the three aspects on the ASR performance, two intermediate MFCC-like feature types were evaluated. The first one, referred to as frequency-domain derived Bark-frequency cepstral coefficients (BFCCF), differs from the MFCC only in the subband filter bank, which was chosen to closely correspond to the filter bank used in the ZCPA computation. The second one, referred to as time-domain derived Bark-frequency cepstral coefficients (BFCCT), is derived from the subband power estimates computed in the time domain. The subband filter bank and analysis frame lengths were identical to those used in the ZCPA method. Thus, the effect of using different filter banks in ZCPA and MFCC methods was determined by comparing the ASR performance of BFCCF and MFCC features, while the effect of using time-domain processing instead of frequency-domain processing was determined by comparing the performance of BFCCT and BFCCF features. Finally, the effect of incorporating dominant subband frequency information into speech features was determined by comparing ZCPA and BFCCT features.

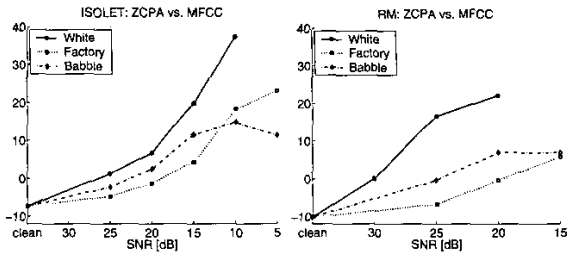

Fig. 1. Difference in word accuracy between ZCPA and MFCC features on ISOLET and RM databases

4.1. Effect of different filter banks

Figure 2 shows the absolute difference in word accuracy between BFCCF and MFCC features as a function of SNR on ISOLET and RM databases respectively. We observe that the difference in performance due to the use of different filter banks is relatively small.

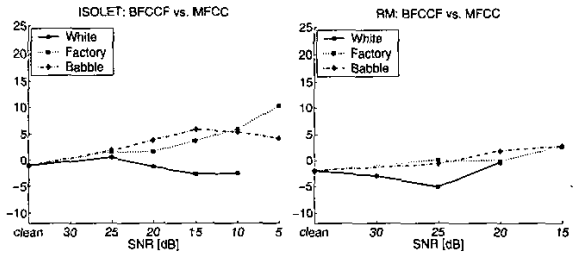

Fig. 2. Difference in word accuracy between BFCCF and MFCC features on ISOLET and RM databases

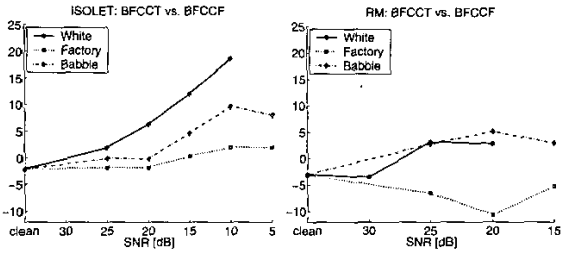

Fig. 3. Difference in word accuracy between BFCCT and BFCCF features on ISOLET and RM databases

4.2. Effect of time-domain processing

Figure 3 shows the absolute difference in word accuracy between BFCCT and BFCCF features as a function of SNR on ISOLET and RM databases respectively. We observe that time-domain processing gave largest improvement for stationary white noise on ISOLET database. One major difference between BFCCT and BFCCF features is the use of frequency-dependent frame lengths that ranged between 33 ms and 134 ms in BFCCT, rather than a constant frame length of 25 ms used in BFCCF. Longer frames lead to more reliable power estimates. However, too long frames violate the stationarity assumption. This might explain why the use of time-domain processing was least beneficial in the case of highly unstationary factory noise, and why the improvements were reduced when tested on continuous speech, that is characterized by shorter stationary intervals.

4.3. Effect of dominant subband frequencies

Figure 4 shows the absolute difference in word accuracy between ZCPA and BFCCT features as a function of SNR on ISOLET and RM databases respectively. We see that the use of dominant subband frequency information led to improved performance in presence of white and factory noise at sufficiently low SNRs, while no improvement was observed in presence of babble noise. This indicates that dominant frequency information has the largest positive effect

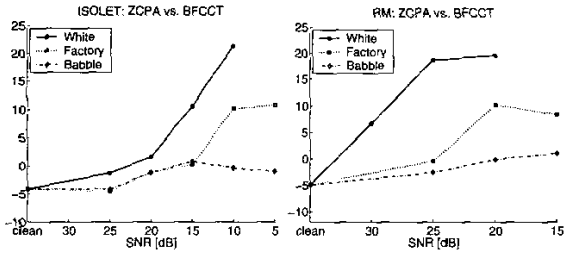

Fig. 4. Difference in word accuracy between ZCPA and BFCCT features on ISOLET and RM databases

when additive noise has relatively flat spectrum. Similar improvements were achieved on both databases. At high SNRs, BFCCT features performed better than ZCPA features. This can be explained by the fact that ZCPA features do not provide reliable information about spectral valleys. This information becomes unreliable in presence of additive noise, in which case its exclusion from speech features can be advantageous. However, at high SNRs, this information contributes to better discrimination between different speech units.

5. CONCLUSIONS

The major conclusions from this study are summarized in the following. Proper choice of analysis frame lengths and filter bandwidths was important for the good performance of ZCPA features, while the performance was not very sensitive to the choice of the number of filters and frequency bins. ZCPA features were shown to be more robust than MFCC features in presence of additive noise. The use of dominant frequency information was shown to have a considerable positive influence on robustness on both databases, especially for noise types with relatively flat spectral characteristics.

6. REFERENCES

[1] 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. Speech Audio Processing*, vol. 7, no. 1, pp. 55–69, Jan. 1999.

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

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

[4] Ron A. Cole, Yeshwant K. Muthusamy, and Mark Fanty, "The ISOLET spoken letter database," Technical report CSE 90-004, Oregon Graduate Institute of Science and Technology, Beverton, OR, USA, Mar. 1990.

[5] Patti Price, William M. Fisher, Jared Bernstein, and David S. Pallett, "The DARPA 1000-word resource management database for continuous speech recognition," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing (ICASSP)*, New York, USA, Apr. 1988, vol. 1, pp. 651–654.

[6] Bojana Gajić, *Feature Extraction for Automatic Speech Recognition in Noisy Acoustic Environments*, Ph.D. thesis, Norwegian University of Science and Technology, Trondheim, Norway, 2002.