Zero-Crossing-Based Channel Attentive Weighting of Cepstral Features for Robust Speech Recognition: The ETRI 2011 CHiME Challenge System

Young-Ik Kim, Hoon-Young Cho, Sang-Hun Kim

Electronics and Telecommunications Research Institute
161 Gajeong-Dong, Yuseong-Gu, Daejon, 305-700, Republic of Korea
{youngik, hycho, ksh}@etri.re.kr

Abstract

We present a practical and noise-robust speech recognition system which estimates a target-to-interferers power ratio using a zero-crossing-based binaural model and applies the power ratio to a channel attentive missing feature decoder in the cepstral domain. In a natural multisource environment, our binaural model extracts spatial cues at each zero-crossing of a filterbank output signal to localize multiple sound sources and estimates a ratio mask reliably which segregates target speech from interfering noises. Our system uses gammatone filterbank cepstral coefficients (GFCCs) for the recognition and the channel attentive decoder utilizes the ratio mask on weighting the cepstral features when calculating the output probability in the Viterbi decoding. On the experiments of CHiME final testset, our channel attentive GFCC system improves the baseline recognition result 12.2% on average, and with noisy training condition, the average improvement amounts to 18.8%.

Index Terms: robust speech recognition, binaural processing, zero-crossing, channel attentive decoding

1. Introduction

The human auditory system can select and segregate a specific sound source among multiple interfering noises using various cues such as spatial location, pitch continuity, and speaking rate. This capability of handling the cocktail party problem has been one of the great barriers to overcome for the contemporary automatic speech recognition system. Among many research activities, the computational auditory scene analysis (CASA) focuses on developing automatic sound separation systems based on human hearing. And recently many research groups in CASA successfully adopted binaural processing techniques for solving the cocktail party problem. In our previous works, effective sound source localization and segregation techniques were also developed using zero-crossing-based binaural extraction of spatial cues [1, 2].

The missing data techniques (MDT) are potentially good choices when some regions of the time-frequency domain are partially corrupted by interfering noises [3, 4]. However, because they use nonorthogonal features of the log spectral domain, their performance are lower than ASR systems using orthogonal features such as cepstral coefficients. A recent work of M. Segbroeck et.al. [5] applied MDT to any other feature domain that is a linear transform of a log-spectrum. In the channel attentive missing feature decoder [6], the reliability of each subband in a mel filterbank analysis is expressed as a weighting factor of each filterbank output and then a combination of the cosine transform and a channel weighting matrix is directly applied to both input cepstrum features and the mean vectors of an acoustic model.

As a successor of the 1st monaural speech separation challenge in 2006 [7], the CHiME (Computational Hearing in Multisource Environments) is a binaural challenge that aims to tackle the speech separation and recognition in more typical everyday listening conditions. In this paper we try to attack the CHiME challenge with our zero-crossing-based binaural frontend [1, 2] and the channel attentive missing feature decoder [6]. The proposed system extends our previous works in several aspects: 1) The nonlinear head-related transfer functions (HRTFs) are approximated into two monotonic increasing functions so that every frequency channel uses the same mappings between spatial cues and azimuth angle. This approximation greatly reduced the computational complexity in a binaural processing and did not drop system performance in the experiments. 2) We developed a simple decision rule for segregating target speech from interfering noises in the azimuth plane. 3) To improve the accuracy in speech recognition, we adopt the channel attentive missing feature decoder which utilizes cepstral domain features instead of the spectral features. 4) The proposed system was tested in a more natural multisource environment of CHiME corpus and proved to be effective in all signal-to-noise conditions.

A block diagram of the proposed system is presented in figure 1. In the following sections we will describe each block in more detail. The zero-crossing-based binaural frontend which extracts the GFCC feature and masking information is explained in section 2. The missing data recognizer with channel attentive weighting is shown in section 3. And our experiments on CHiME corpus and conclusions follow in section 4 and 5, respectively.

2. Zero-Crossing-Based Binaural Frontend

In the frontend stage of the system, Gammatone filters [8] are used to simulate the cochlea frequency analysis of binaural input signals. For a practical speech recognition frontend, a bank of 32 overlapping bandpass filters with center frequencies between 200 and 5000 Hz are spaced uniformly on the equivalent rectangular bandwidth (ERB) scale. Here we denote the channel output signals of a filterbank as $x_i^L(t)$ and $x_i^R(t)$ for the left-hand and the right-hand sensors, respectively. Hereafter the symbol L and R represent binaural channels and i represent the frequency channel index.

2.1. Extraction of spatial cues

To extract spatial cues, such as interaural time difference (ITD) and interaural intensity difference (IID), our system first detects upward zero-crossing events at each channel output sig-

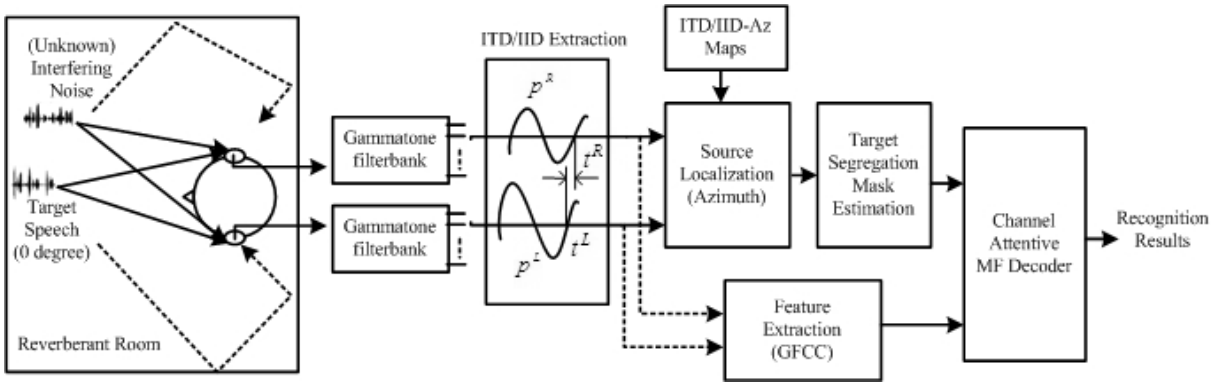


Figure 1: An overview of the proposed speech recognition system for CHiME challenge

nal. When we denote the upward zero-crossing times of left and right sensor as $t_i^L(n)$ and $t_i^R(m)$, the ITD is defined as

$$ITD = \Delta t_i(n,m) = t_i^L(n) - t_i^R(m), \tag{1}$$

where n and m are zero-crossing time indexes. The IID is determined from the powers of zero-crossing intervals. If we denote the powers of zero-crossing intervals measured at zero-crossing time indexes as $p_i^L(n)$ and $p_i^R(m)$, the IID is defined as

$$IID = \Delta p_i(n,m) = 10log_{10}\frac{p_i^L(n)}{p_i^R(m)}. \tag{2}$$

In our zero-crossing-based binaural model, the ITD and the IID can be measured for an arbitrary pair of zero-crossing samples which are generated from the binaural channel signals. To resolve the ambiguity of selecting true ITD and IID pair from many candidates, we use following coincidence principle. Since the measured ITD and IID convey the same directional information of sound sources, they should be coincident with each other in direction. When we project the ITD and the IID on the azimuth angle plane with maps of sound directions and the binaural cues, we can compare the distance in the azimuth plane to select the best coincident ITD and IID pair. In practice, we select the coincident ITD-IID pair as Equation 3. For the nth zero-crossing in left-hand channel, the mth coincident zero-crossing is selected in right-hand channel as follows.

$$m = \underset{k}{argmin} \mid \theta_{ITD}(n,k) - \theta_{IID}(n,k) \mid \tag{3}$$

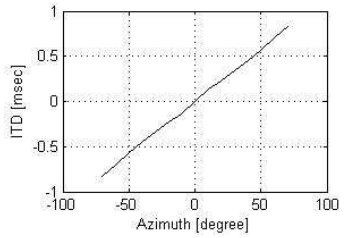
2.2. ITD/IID to azimuth mapping

The map $f_T(\theta)$ that transforms the azimuth angles measured from the frontal axis to ITD values and the map $f_I(\theta)$ that transforms the angles onto IID values are constructed from the binaural room impulse responses (BRIR) data. These mappings are a kind of head-related transfer functions (HRTF) and play an important role in binaural selection of consistent ITD and IID cues. Usually the HRTFs vary over frequency channels but in our approximated maps the variations are ignored. To estimate the sound direction from measured ITD and IID values, we can use the inverse mappings of ITD and IID to azimuth angle.

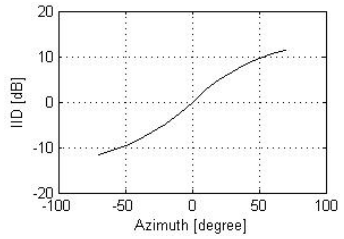
$$\theta_{ITD}(n,m) = f_T^{-1}(\Delta t_i(n,m)) \tag{4}$$

$$\theta_{IID}(n,m) = f_I^{-1}(\Delta p_i(n,m)) \tag{5}$$

These inverse mappings are possible since we approximated the maps as monotonically increasing functions.



(a)



(b)

Figure 2: Approximate HRTF maps calculated from CHiME binaural room impulse responses. (a) azimuth to ITD map and (b) azimuth to IID map.

For CHiME experiment, we build two approximate maps from the BRIR data which is recorded in the head and torso simulator in CHiME environment. The two maps are shown in Figure 2. And detailed descriptions about the BRIR data can be found at [9]. In these maps, the mapping data are exist in the range between -70 to 70 azimuth degrees because the BRIR recordings are conducted only those range of directions.

2.3. Localization

One of the useful characteristics in zero-crossing-based binaural processing is that the local SNR at frequency channel i and zero-crossing index n can be approximated from the variance of neighboring ITD samples as the following Equation 6:

$$SNR \approx 10log_{10}\frac{1}{f_i^2 Var(\Delta t_i(n,m))} \tag{6}$$

where f_i is the center frequency of the bandpass filtered signal and $Var(\Delta t_i(n,m))$ is the variance of ITD samples. For more detailed description about the approximation, please refer to [1].

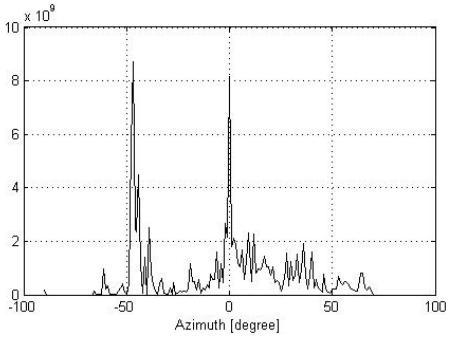


Figure 3: An illustration of the weighted angle histogram. The speech sample is extracted from the CHiME corpus in -3dB mixture case. Here the target is located in 0 degree and a dominant noise is detected at near -50 degrees.

To localize multiple sound sources effectively, the SNR values are computed at each zero-crossing times over all frequency channels. And during the entire speech periods, the SNR value is weighted by the zero-crossing's interval power and accumulated in the corresponding azimuth angle histogram bin. The resultant SNR and power weighted azimuth angle histogram is used to localize multiple sound sources and to determine a decision boundary that separates the target speech from the interfering noises in azimuth plane.

An illustration of the weighted angle histogram is shown in Figure 3. In this illustration, we used a speech sample (s1_sgib8n.wav) extracted from the CHiME development corpus in SNR -3dB set. From the histogram, we can clearly locate two dominant sources. Since the target is located at 0 degree in the CHiME corpus, we can assume that a dominant noise is located at -50 degrees in the histogram. The low level peaks in the histogram correspond to other background noises and reverberation effects.

2.4. Segregation mask estimation

The peak values are identified as sound sources in the weighted angle histogram. But to generate a mask which segregates the target from interfering noises, we need to determine a decision boundary in azimuth plane. Since the target is always located at front in the CHiME environment, the decision can be made relatively easy. The procedure we have taken is as follows: 1) Locate the target at front. The target peak value $peak_T$ is searched within [-5, 5] degrees. 2) Search dominant peaks from 0 degree point to left-hand side and right-hand side. A dominant peak need to have a value greater than $peak_T/3$. 3) Set the azimuth angle separation boundary as the middle value between the target angle and neighboring peak angle. Here we set the limit of the boundary values as [-15, 15] since there are many background noises and reverberations in the CHiME environment which cannot be detected as a dominant peak.

Using zero-crossing-based spatial cues and the azimuth angle decision boundary, we can reliably estimate the target-to-interferers power ratio in a time-frequency region. To estimate the power ratio $r(\tau, i)$ for a region with time index τ and frequency index i, we collect the zero-crossings which belong to the region and separate them into two groups of the target and interferers based on the zero-crossing's azimuth angle estimate. Then the power sum of target zero-crossings $p_T(\tau, i)$ and that of interferers $p_I(\tau, i)$ are calculated. The power ratio is calculated as Equation 7 and it is used as a weighting factor of the channel attentive missing feature decoder during the speech recognition experiments.

$$r(\tau, i) = \frac{p_T(\tau, i)}{p_T(\tau, i) + p_I(\tau, i)} \tag{7}$$

3. Channel Attentive Decoding with Cepstral Features

Conventional cepstral feature extraction methods based on filter bank analysis apply a discrete cosine transform to a logarithmic filterbank energy vector to remove redundant information as well as to orthogonalize the features. Denoting a Q-dimensional log filterbank energy by $\mathbf{x}^l = (x_1^l, \cdots, x_Q^l)$, a N-dimensional cepstral coefficient is derived by $\mathbf{x}^c = \mathbf{C} \cdot \mathbf{x}^l = (x_0^c, x_1^c, \cdots, x_{N-1}^c)$, where $\mathbf{C}$ is a discrete cosine transform matrix. A ceptral mean vector, $\bar{\mathbf{x}}^c$, is obtained from the entire temporal interval of an utterance and subtracted from $\mathbf{x}^c$ to remove a channel distortion and a final cepstral feature vector, $\tilde{\mathbf{x}}^c$ is extracted.

Assume that the degree of reliability for each filter bank output is somehow obtained and expressed as $\mathbf{W} = diag(w_1, \cdots, w_Q)$, where the $w_i (0 < w_i \leq 1)$ is the reliability value for the i-th sub-band. The previously studied channel attentive Mel frequency cepstral coefficient (CAMFCC) method utilizes the channel reliability information by using it as a weighting factor for each element of logarithmic filter bank energy vector. Thus, a CAMFCC vector is obtained by $\hat{\mathbf{x}}^c = \mathbf{C} \cdot \mathbf{W} \cdot \mathbf{C}^{-1} \cdot \mathbf{x}^c = \mathbf{V} \cdot \mathbf{x}^c$. In the CAMFCC method, the same weight matrix $\mathbf{W}$ is applied equally to the mean vectors of hidden Markov model at the run time. If we denote a mean vector of an HMM state s by μ^c, $\mathbf{W}$ is applied to the HMM mean vectors as,

$$\mu^l = \mathbf{C}^{-1} \mu^c \tag{8}$$

$$\hat{\mu}^l = \mathbf{W} \mu^l \tag{9}$$

$$\hat{\mu}^c = \mathbf{C}\, \hat{\mu}^l \tag{10}$$

The logarithmic output probability of an observation vector $\hat{\mathbf{x}}^c$ given the state s with the modified mean vector $\hat{\mu}^c$ is now expressed as,

$$\log Pr(\hat{\mathbf{x}}^c | s) = -0.5(\hat{\mathbf{x}}^c - \hat{\mu}^c)^t \mathbf{\Sigma} (\hat{\mathbf{x}}^c - \hat{\mu}^c) + K \tag{11}$$

$$= -0.5(\mathbf{C}\mathbf{W}\mathbf{x}^l - \mathbf{C}\mathbf{W}\mu^l)^t \times \mathbf{\Sigma}\, (\mathbf{C}\mathbf{W}\mathbf{x}^l - \mathbf{C}\mathbf{W}\mu^l) + K \tag{12}$$

$$= -0.5 \sum_{i=0}^{N-1} \left[\sum_{j=1}^{Q} \frac{c_{ij} w_j (x_j^l - \mu_j^l)}{\sigma_i} \right]^2 + K \tag{13}$$

In 13, if the jth channel has very low reliability, then its weight value will be closer to zero, $w_j \simeq 0$, therefore the corresponding channel is excluded from the output probability calculation. In other words, the jth channel is controlled to contribute little to the output probability score.

In this work, we slightly extended the original CAMFCC by considering the cepstral mean subtraction together. In this case, $\tilde{\mathbf{x}}^c$ is used instead of $\hat{\mathbf{x}}^c$ in 11. In addition, the weight matrix $\mathbf{W}$ is extended to be time-varying and denoted by $\mathbf{W}_t = diag(r(t, 1), \cdots, r(t, Q))$, where $r(t, i)$ is from equation 7.

$$\log Pr(\tilde{\mathbf{x}}^c|s) = -0.5(\tilde{\mathbf{x}}^c - \tilde{\mu}^c)^t \mathbf{\Sigma}(\tilde{\mathbf{x}}^c - \tilde{\mu}^c) + K \quad (14)$$

$$= -0.5(\hat{\mathbf{x}}^c - \bar{\mathbf{x}}^c - \tilde{\mu}^c)^t \times \mathbf{\Sigma}(\hat{\mathbf{x}}^c - \bar{\mathbf{x}}^c - \tilde{\mu}^c) + K \quad (15)$$

Since the whole process of domain transforms and the channel weighting is a series of linear transforms as suggested by 8 $\sim$ 10, and the cepstral mean subtraction is also a linear operation, both of $\tilde{\mathbf{x}}^c$ and $\tilde{\mu}^c$ can substitute the $\hat{\mathbf{x}}^c$ and $\hat{\mu}^c$, respectively, in 11 $\sim$ 13.

4. Experiments

The CHiME corpus which is designed for noise-robust speech processing research contains binaural recordings from reverberant domestic environments with many simultaneous and unpredictable sound sources [9]. Using the BRIR data the binaural signals are mixed with the target speech taken from the Grid corpus [7]. When mixing the signal, the target speech is always located in a frontal direction and the task is to recognize keywords of the Grid utterance (a letter and a digit) from the mixture signal. As a baseline, they provide a standard 39-dimensional Mel frequency cepstral coefficients (MFCCs) frontend applied with cepstral mean normalization (CMN). The baseline also employs speaker dependent training of HMMs using reverberant signals without any additional noise.

Table 1: Results for baseline training condition

	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB
MFCC	30.33	35.42	49.50	62.92	75.00	82.42
GFCC	35.33	40.42	53.17	65.42	77.08	85.50
GFCC+CA	49.50	52.58	63.58	73.00	81.92	88.00

Using the baseline training condition in the CHiME, Table 1 shows the results of the baseline and the proposed systems for 6 different SNR conditions. Compared to the baseline MFCC system, the proposed system utilizes GFCCs (Gammatone frequency cepstral coefficients) extracted from the output of a Gammatone filterbank. The feature also contains deltas and accelerations and apply same technique of CMN as the baseline system. As with the baseline, monaural signals are used for the GFCC feature extraction by averaging binaural data in waveforms. As shown in Table 1, the GFCC system has an average performance gain of 3.6% compared to the MFCC system. The channel attentive missing feature decoder was implemented by extending the HTK code. We simply modified the model likelihood computation of the HTK routine to adopt segregation weight vector and applied the channel attentive weighting frame by frame. As shown in the table, when we applied the segregation weighting to our channel attentive missing feature decoder, the performance gain was amount to 12.2% on average.

Table 2: Results for noisy-condition training

	-6 dB	-3 dB	0 dB	3 dB	6 dB	9 dB
GFCC	49.50	54.17	66.50	77.83	86.08	90.17
GFCC+CA	57.00	60.83	72.17	80.75	87.08	90.50

Given that the noisy test condition is known in advance, by incorporating similar noisy data in the training stage, the overall system performance can be greatly enhanced. To generate a noisy acoustic model set, we used the background noise signals supplied by the CHiME corpus. A total of 57 background signals were used to generate a set of noisy training data. For each training speech signal, one of the background noise signals was randomly selected and then was added to the utterance with SNR 5dB. Since the noise signal is much longer than training speech, noise was added from an arbitrary starting point. We assumed that the entire period of each training sample is speech period in the calculation of SNR. Both of the original and the noisy training data were used to generate a final noisy acoustic model. Table 2 shows the recognition results of the proposed systems when we employ the noisy- condition training. In this experiment, compared to the baseline MFCC system, the average performance gain of our GFCC system was 14.8% and in the case of GFCC with channel attention, the average gain was amount to 18.8%.

5. Conclusion

In this paper, we extended our previous works of zero-crossing-based binaural speech segregation and channel attentive missing data recognition techniques in several ways. The proposed system showed impressive results on the test of the CHiME robust speech recognition challenge and practical enough to be implemented in many contemporary real-time applications. We believe that the zero-crossing-based analysis of other kinds of auditory features such as pitch continuity or onset/offset detection is also useful. For a future research, we plan to incorporate them in our binaural speech recognition system.

6. References

[1] Y. Kim and R. M Kil, "Estimation of inter-aural time differences based on zero-crossings in noisy multi-source environments", *IEEE Transactions on Audio, Speech, and Language Processing*, 15(2), pp. 734-743, 2007.

[2] S. J. An, R. M. Kil and Y. Kim, "Zero-crossing-based speech segregation and recognition for humanoid robots", *IEEE Transactions on Consumer Electronics*, 55(4), pp. 2341-2348, 2009.

[3] J. Barker, N. Ma, A. Coy and M. Cooke, "Speech fragment decoding techniques for simultaneous speaker identification and speech recognition," *Computer Speech and Language*, 24(1), pp. 94-111, 2010.

[4] N. Ma, P. Green, J. Barker, and A. Coy, "Exploiting correlogram structure for robust speech recognition with multiple speech sources," *Speech Communication*, 49, pp. 874-891, 2007.

[5] M. V. Segbroeck and H. V. hamme, "Advances in missing feature techniques for robust large vocabulary continuous speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, 19(1), pp. 123-137, 2011.

[6] H. Cho and Y. Oh, "On the use of channel-attentive MFCC for robust recognition of partially corrupted speech", *IEEE Signal Processing Letters*, 11(6), pp. 581-584, 2004.

[7] M. Cooke, J. R. Hershey and S. J. Rennie, "Monaural speech separation and recognition challenge," *Computer Speech and Language*, 24, pp. 1-15, 2010.

[8] B. Glasberg and B. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, 47, pp. 103-138, 1990.

[9] H. Christensen, J. Barker, N. Ma, and P. Green, "The CHiME corpus: a resource and a challenge for Computational Hearing in Multisource Environments," *Interspeech 2010*, Makuhari, Japan, 2010.