Mask Estimation Incorporating Time-Frequency Trajectories for a CASA-based ASR Front-end

Ji Hun Park, Jae Sam Yoon, and Hong Kook Kim

Department of Information and Communications
Gwangju Institute of Science and Technology, Gwangju 500-712, Korea

{jh.park, jsyoon, hongkook}@gist.ac.kr

Abstract
In this paper, we propose a mask estimation method for a computational auditory scene analysis (CASA) based speech recognition front-end using speech obtained from two microphones. The proposed mask estimation method incorporates the observation that the mask information should be correlated over contiguous analysis time frames and adjacent frequency channels. To this end, two different hidden Markov models (HMMs), time HMM and frequency HMM, representing the time and frequency trajectories respectively, are trained using features such as the interaural time difference and the interaural level difference of two-channel signals. A mask for the given time-frequency bin is estimated by combining the likelihoods estimated from the two HMMs, and used to separate the desired speech from noisy speech. To show the effectiveness of the proposed mask estimation, we first measure the root mean square error between the ideal mask and that estimated by the proposed method. Then, we compare the performance of a speech recognition system using the proposed mask estimation method to those using conventional methods. Consequently, the proposed method provides an average word error rate reduction of 63.2% and 3.1% when compared with the Gaussian kernel-based and time HMM-based mask estimation methods, respectively.

Index Terms: computational auditory scene analysis, mask estimation, hidden Markov model, speech recognition

1. Introduction
In the human auditory system, a desired signal can be localized and separated by the time and intensity difference of the arrival signals at each of the ears, differences referred to as the interaural time difference (ITD) and the interaural level difference (ILD), respectively. In a similar fashion, computational auditory scene analysis (CASA) attempts to separate sounds into a target signal and noise using ITDs and ILDs when sounds come from two or more microphones [1]. In general, mask information in each frequency bin for a given time frame is first estimated from ITDs and ILDs. Then, since the mask information indicates whether a particular frequency region for a given time frame includes target speech dominantly or noise dominantly, the target speech can be retrieved after applying the estimated mask to the sounds. Accordingly, correct estimation of the mask information is required to obtain high-quality separation performance, thus it is critical to improve the performance of an automatic speech recognition (ASR) system based on CASA.

There have been a number of research works investigating mask estimation in binaural environments reported. Of these, Roman et al. proposed a mask estimation method based on a supervised learning algorithm [2] under conditions assumed to be independent over each analysis frame. Conversely, to take the continuity of speech along the analysis frames into account, a hidden Markov model (HMM) based mask estimation method was proposed in [3]. Since this method considered only the time continuity of speech, there are still frequency discontinuities over contiguous frequency channels in mask patterns due to the time continuity of speech and noise.

In this paper, we propose a mask estimation method that incorporates the observation that mask information should be simultaneously correlated over contiguous analysis time frames and frequency channels. In other words, HMMs corresponding to time and frequency are used to estimate mask information that reflects the continuous properties of speech along both the time frame and frequency channels.

Following this introduction, Section 2 describes how a front-end for ASR can be constructed using CASA. In Section 3, we propose an HMM-based mask estimation method as a means of incorporating the time and frequency trajectory of speech. After that, in Section 4 we analyze the mask pattern estimated using the proposed method to measure how close the estimated mask is to the ideal mask. Next, the performance of the proposed HMM-based mask estimation method is evaluated with respect to speech recognition performance in Section 5, and we conclude our findings in Section 6.

2. CASA-based ASR front-end
In this section, we describe a schematic diagram of an ASR front-end based on CASA. As shown in Fig. 1, the front-end separates the desired speech from noise in a multiple microphone environment by employing mask information, and then extracts the speech recognition features.

2.1. Auditory periphery
Binaural input signals with a sampling rate of 16 kHz are decomposed into auditory spectral signals by employing a gammatone filterbank [4] with 32 channels, in which the center frequencies are linearly spaced on an ERB-scale [5] from 50 Hz to 8 kHz. Here, the auditory spectral signals are windowed using a rectangular window with a time resolution of 20 ms and a frame rate of 100 Hz. The envelopes of the left and right auditory spectral signals for the i-th frequency channel and the j-th frame, env\_L(i, j) and env\_R(i, j), are computed as follows.

\[
env\_L(i, j) = \sum_{n=0}^{N-1} x_{L}^{i,n}(n) \quad env\_R(i, j) = \sum_{n=0}^{N-1} x_{R}^{i,n}(n) \quad (1)
\]
where \( x_{L}^{i,j}(n) \) and \( x_{R}^{i,j}(n) \) represent the \( n \)-th left and right auditory spectral signals for the \( i \)-th frequency channel and the \( j \)-th frame, respectively. In addition, \( N \) is the number of speech samples of a frame, set to 320 in this paper.

### 2.2. Binaural cue extraction

In order to estimate the mask patterns, we extract an ITD and an ILD for each time-frequency (T-F) bin. To achieve this, we first compute the normalized cross-correlation between the auditory spectral signals from the left channel \( \{ x_{L}^{i,j}(n) \} \) and the right channel \( \{ x_{R}^{i,j}(n) \} \) defined as

\[
CC^{i,j}(\tau) = \frac{\sum_{n=0}^{N-1}[x_{L}^{i,j}(n)]^{*}[x_{R}^{i,j}(n-\tau)]}{\sqrt{\sum_{n=0}^{N-1}|x_{L}^{i,j}(n)|^{2}}\sqrt{\sum_{n=0}^{N-1}|x_{R}^{i,j}(n)|^{2}}}
\]

(2)

where * indicates a complex conjugate, and \( \tau \) ranges from -16 to 16, corresponding to the range from -1ms to 1 ms at a sampling rate of 16 kHz. Then, we estimate the ITD for each T-F bin as a time lag where the normalized cross-correlation is maximized. In other words,

\[
ITD(i,j) = \arg \max_{\tau} CC^{i,j}(\tau).
\]

(3)

Next, the ILD for each T-F bin is computed as the ratio of auditory envelopes obtained from the left and right channel signals using

\[
ILD(i,j) = 20 \log_{10} \left( \frac{env_{L}(i,j)}{env_{R}(i,j)} \right).
\]

(4)

In order to determine which channel the speech source is coming from, we compute the pooled cross-correlation by taking the sum of the cross-correlations in (2) over all frequency channels and all time frames. Then, the time lag, \( \tau_{\text{max}} \), where the pooled cross-correlation is maximized can be estimated as

\[
\tau_{\text{max}} = \arg \max_{\tau} \sum_{i,j} CC^{i,j}(\tau).
\]

(5)

Note that the left channel is selected as the speech channel if \( \tau_{\text{max}} \) is negative; otherwise, the right channel is selected. Finally, the envelope extracted from the corresponding channel is used for the speech recognition front-end, further described in Section 2.4.

### 2.3. Mask estimation

The mask information is extracted from the estimated ITDs and ILDs, and used to separate the desired speech from noisy speech. To this end, several mask estimation methods have been developed, including the Gaussian kernel-based method [2], and the time HMM-based method [3]. In this paper, we propose a T-F HMM-based mask estimation method that incorporates the time and frequency trajectories of speech, further described in Section 3.

### 2.4. Feature extraction for speech recognition

In order to construct a front-end for ASR, a mask pattern is first applied to noisy speech. Let us define the estimated mask pattern for the \( i \)-th frequency channel and the \( j \)-th frame as \( m(i,j) \). This mask pattern is used to estimate the envelope of the speech component only, \( env_{m}(i,j) \), by multiplying \( m(i,j) \) to the envelope of the speech channel selected by the value of \( \tau_{\text{max}} \) as described in Section 2.2. That is,

\[
env_{m}(i,j) = \begin{cases} 
env_{L}(i,j) \cdot m(i,j) & \text{if } \tau_{\text{max}} < 0 \\
env_{R}(i,j) \cdot m(i,j) & \text{otherwise}
\end{cases}
\]

(6)

As speech recognition features, 13 cepstral coefficients for each frame are extracted from the separated envelopes, \( env_{m}(i,j) \), by applying a discrete cosine transform (DCT) defined as

\[
c(j,k) = \sum_{i=1}^{I} \log \left( env_{m}(i,j) \right) \cos \left( \frac{k\pi}{I} (i - 0.5) \right)
\]

(7)

where \( I \) is the total number of the gammatone filters and set to 32 in this paper, and \( k \) indicates the index number of cepstral coefficients ranging from 0 to 12.

### 3. HMM-based mask estimation

The mask information must simultaneously incorporate the time and frequency trajectory of speech to accurately reflect the time and frequency continuity of a speech signal. For this purpose, an T-F HMM-based mask estimation method is proposed.

#### 3.1. Mask estimation

Fig. 2 shows the proposed mask estimation process. To reflect the time and frequency continuity of speech, we first train two HMM-based mask models such as time HMM and frequency HMM models. The mask patterns are then estimated using a Viterbi search [6] with the trained time and frequency HMMs. As a result, we can obtain the probabilities of speech and noise for each T-F bin, the ratio of which is calculated as follows.

\[
R_{i,j} = \frac{P_{n}(i,j)}{P_{s}(i,j) + P_{n}(i,j)}
\]

(8)

where \( P_{s}(i,j) \) and \( P_{n}(i,j) \) represent the speech and noise probability, respectively, for the \( i \)-th frequency channel and \( j \)-th frame. In particular, let us define the probability ratio obtained by the time HMM and the frequency HMM as \( R_{c}(i,j) \) and \( R_{r}(i,j) \), respectively. Then, the two ratios can be combined as

\[
R_{c}^{\text{mu}}(i,j) = \mu R_{c}(i,j) + (1 - \mu) R_{r}(i,j)
\]

(9)

where \( \mu \) indicates the weighting factor. Here, \( \mu \) is determined using a least square error estimator. The error is defined as the squared sum of the difference between the ideal mask and the
The RMS error of the proposed mask estimation method is the smallest from among the methods listed. Fig. 3 shows

The RMS errors for a mask estimation method are then calculated as

\[
\epsilon_{static}^2 = \frac{1}{(I-1) \cdot (J-1)} \sum_{i=1}^{I-1} \sum_{j=1}^{J-1} d_i(i,j)^2
\]

\[
\epsilon_t^2 = \frac{1}{(I-1) \cdot (J-1)} \sum_{i=1}^{I-1} \sum_{j=1}^{J-1} d_i(i,j)^2
\]

\[
\epsilon_f^2 = \frac{1}{(I-1) \cdot (J-1)} \sum_{i=1}^{I-1} \sum_{j=1}^{J-1} d_i(i,j)^2
\]

where \(d_i(i,j) = m_{\text{ideal}}(i,j) - m(i,j)\), \(d_i(i,j) = \Delta_{\text{ideal},f}(i,j) - \Delta_f(i,j)\), and \(d_f(i,j) = \Delta_{\text{ideal},f}(i,j) - \Delta_f(i,j)\). In addition, \(I\) and \(J\) are the total numbers of frames and of frequencies for a speech signal, respectively. In (14), (15), and (16), we do not take into the mask values at the boundary account for mask pattern analysis due to the error of delta value taken at the boundary.

Table 1 shows the RMS errors of the mask pattern estimated from the different mask estimation methods. As shown in the table, the RMS error of the proposed mask estimation method is the smallest from among the methods listed. Fig. 3 shows
the ideal mask pattern and the mask patterns estimated by using different mask estimation methods, including the proposed method. As shown in the figure, when only the time HMM is used to estimate mask information, the target speech region, represented by the white color, is corrupted by noise due to time continuity of the noise signal. Conversely, when the mask pattern is estimated using the proposed T-F HMM, the effect of noise continuity is reduced, resulting in the target speech regions being more refined. Consequently, it is shown from this figure that the T-F HMM-based mask pattern reflects well the time and frequency continuity of speech.

5. Speech recognition experiments

In this section, we evaluated the performance of the proposed T-F HMM-based mask estimation method in terms of speech recognition ability. In other words, the six different front-ends described in Section 2 were constructed after applying an ideal mask, a uniform mask, a Gaussian kernel-based, time HMM, frequency HMM, and T-F HMM-based masks. MFCCs were also used to provide a reference performance by being extracted from the noisy speech of the left channel which target speech source was coming from.

For the speech recognition experiments, a binaural database was artificially constructed using a Korean speech corpus [10], following the procedure described in Section 3.2. Subsequently, 18,240 utterances of the corpus were used to train the acoustic model, and 570 utterances were used as the target speech data. Then, each target speech utterance was transformed into a bin-aural signal and mixed with female speech and music noise. The aural signal and mixed with female speech and music noise. Then, each target speech utterance was transformed into a bin-aural signal and mixed with female speech and music noise. The acoustic models were based on left-to-right triphone HMMs, and trained using the HMM toolkit [7]. The number of Gaussian mixtures was 4 per state. For the language model, the lexicon size was 2,250 words and a unigram was employed.

Table 2 shows the word error rates (WERs) according to different mask estimation methods. It was shown from the table that the proposed T-F HMM-based mask provided the smallest WER for all the noise conditions except for the ideal mask. As a result, relative WER reductions of 63.2%, 3.1% and 33.4% were achieved by the proposed method, when compared with the Gaussian kernel-based mask, time HMM-based mask, and frequency HMM-based mask estimation method, respectively.

6. Conclusion

In this paper, we proposed an HMM-based mask estimation method for an ASR front-end based on computational auditory scene analysis. In order to incorporate the observation that the mask information should be correlated over contiguous time frames and adjacent frequency channels, spatial parame-

---

Table 2: Comparison of WERs(%) according to different mask estimation methods under a female speech noise condition and a background music condition.

<table>
<thead>
<tr>
<th>Mask</th>
<th>Noise Condition</th>
<th>Female Speech</th>
<th>Music</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10°</td>
<td>20°</td>
<td>40°</td>
</tr>
<tr>
<td>MFCC</td>
<td>47.3</td>
<td>43.8</td>
<td>55.4</td>
</tr>
<tr>
<td>Ideal</td>
<td>9.7</td>
<td>9.8</td>
<td>10.1</td>
</tr>
<tr>
<td>Uniform</td>
<td>41.3</td>
<td>36.8</td>
<td>30.9</td>
</tr>
<tr>
<td>Gaussian Kernel</td>
<td>39.8</td>
<td>45.1</td>
<td>44.6</td>
</tr>
<tr>
<td>Time HMM</td>
<td>16.7</td>
<td>14.7</td>
<td>13.6</td>
</tr>
<tr>
<td>Freq. HMM</td>
<td>26.4</td>
<td>21.4</td>
<td>20.5</td>
</tr>
<tr>
<td>T-F HMM</td>
<td>16.3</td>
<td>13.9</td>
<td>13.2</td>
</tr>
</tbody>
</table>

---

7. Acknowledgement

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2007-314-D00245) and partially by the Ministry of Knowledge Economy, Korea, under the Information Technology Research Center support program supervised by the Institute of Information Technology Advancement (IITA-2008-C1090-0801-0017).

8. References