ABSTRACT

In this paper, a subband minimum classification error beamforming (S-MCEBEAM), instead of the subband likelihood maximizing beamforming (S-LIMABEAM) proposed by Seltzer, is investigated to closely integrate microphone array and speech recognizer for robust speech recognition in reverberant environments. The main idea behind this is to apply minimum classification error (MCE) criterion to directly match the goal of automatic speech recognition (ASR) and to simultaneously adjust both beamformer parameters and recognizer’s acoustic models. Experimental results on a Mandarin reverberation corpus created from Mandarin spontaneous speech corpus (TCC300) and RWCP’s sound scene database show S-MCEBEAM leads to better recognition results than S-LIMABEAM in reverberant environments.

Index Terms— microphone array, speech recognition

1. INTRODUCTION

While automatic speech recognition (ASR) systems perform well on recordings made by a close-talking microphone, distant microphone scenarios, such as meeting room or robot control tasks, still present a considerable challenge, especially for the reverberation issues. A proven method to improve recognition results in such a scenario is the use of microphone arrays [1].

Recent works by Seltzer [2-4] indicate that classical approaches to beamforming may not yield optimal results in terms of recognition error on speech recognition task. This may due to a mismatch between the training criterion of classical beamformers, i.e., maximizing output signal-to-noise ratio (SNR), and the target criterion, i.e., minimizing recognition error. Therefore, Seltzer proposed to use maximum likelihood (ML) criterion to approximate the target criterion more closely by adapting the beamformer parameters to maximize the likelihood of the filtered acoustic data evaluated by a recognizer’s acoustic models.

However, in Seltzer’s works, the beamformer parameters may still not be optimized with respect to recognition error. Therefore, the purpose of this work is to investigate the possible advantages of using minimum classification error (MCE) [5] criterion to directly match the goal of ASR and to adjust not only the beamformer parameters but also the underlying acoustic models.

This paper is organized as follows. Section 2 reviews the S-LIMABEAM [4] framework. Section 3 describes the proposed S-MCEBEAM approach. Section 4 reports the experimental results on a Mandarin reverberation corpus created using Mandarin spontaneous speech corpus (TCC300) [6] and RWCP sound scene database [7]. Some conclusions are given in the last section.
Therefore, \( \xi = \{ \mathcal{H}_p, m = 0 \sim M - 1, p = 0 \sim P - 1 \} \) is the set of beamformer parameters which has to be optimized.

### 2.2. MEL-FILTERBANK FEATURE EXTRACTION

On the center of the Fig. 1 is a typical mel-filterbank feature extraction frontend of ASR system.

If we define \( V'[k] \) as the value of the \( l \)-th mel triangle applied to subband \( k \), the \( l \)-th component of the mel-spectrum of frame \( i \), i.e., \( M_i' \), can be expressed as

\[
M_i' = \sum_{k=1}^{l} V'[k] Y[k] Y'[k]
\]  

(2)

where \( l \) and \( l \) are the discrete Fourier transform (DFT) bins corresponding to the left and right edges of the \( l \)-th mel filter, respectively. Outside of this range, the value of \( V'[k] \) is set to zero.

Finally, the \( l \)-th components of the mel-filterbank features \( Z_i' \) at frame \( i \) is as follows:

\[
Z_i' = \log(M_i')
\]  

(3)

Substituting Eq. (1) and (2) into (3) clearly reveals that a given mel-spectral component \( Z_i' \) is a function of the subband filter parameters of all microphones and all subbands in the frequency range spanned by its mel filter.

### 2.3. MAXIMUM LIKELIHOOD OPTIMIZATION

In ASR task, the recognizer usually chooses a hypothesis \( \hat{\omega} \) according to

\[
\hat{\omega} = \arg \max_\omega \log( P(\mathcal{Z}(\hat{\xi}) | \omega) P(\omega) )
\]  

(4)

where \( P(\omega) \) is the language score of hypothesis \( \omega \) according to, \( \hat{\xi} \) is the parameters of the microphone array, \( Z(\hat{\xi}) \) is the feature vector fed into the recognizer and \( P(\mathcal{Z}(\hat{\xi}) | \omega) \) is the likelihood given by the underlying hidden Markov models (HMMs).

Let \( \omega_i \) be the correct transcript for the input signal. The optimal set of beamformer parameters \( \hat{\xi} \) should be selected to maximize the likelihood of the features \( \mathcal{Z}(\hat{\xi}) \) under the hypothesis \( \omega_i \), i.e.,

\[
\hat{\xi} = \arg \max_{\xi} \log( P(\mathcal{Z}(\hat{\xi}) | \omega_i) )
\]  

(5)

To find the optimal \( \hat{\xi} \), an Expectation-maximization (EM) algorithm which iteratively applies Eq. (6) and (7) is usually used where \( \hat{s} \) is the optimal state sequence with respective to \( \omega_i \).

\[
\hat{\xi} = \arg \max_{\xi} \sum_i \log \{ P(\mathcal{Z}_i(\hat{\xi}) | \hat{s}) P(\hat{s} | s_{i \ldots}, \omega_i) \}
\]  

(6)

\[
\hat{s} = \arg \max_{\hat{s}} \sum_i \log \{ P(\mathcal{Z}_i(\hat{\xi}) | \hat{s}) P(\hat{s} | s_{i \ldots}, \omega_i) \}
\]  

(7)

### 3. S-MCEBEAM

To improve recognition performance in distant microphone scenario, we try to further close the gap between the optimization criterion and actual optimization target by replacing the ML in S-LIMABEAM with minimum classification error (MCE) criterion and operating in the cepstral domain [8]. Therefore, the right-hand side of the Fig. 1 is modified by adding a discrete cosine transform (DCT) and a MCE training module.

#### 3.1. MFCC FEATURES

First, the \( l \)-th components of MFCC, \( C_i^l \) at frame \( i \), is calculated as follows:

\[
C_i^l = \sum_n D_{ln} Z_i^l(\xi)
\]  

(8)

where \( D_{ln} \) is the DCT matrix, indexed by DCT band \( l \) and mel band \( n \).

Substituting Eq. (1), (2) and (3) into (8) also shows that a given MFCC component \( C_i^l \) is again a function of beamformer parameter \( \xi \).

#### 3.2. MCE CRITERION

In cepstral domain, the recognizer then should choose a hypothesis \( \hat{\omega} \) according to Eq. (9).

\[
\hat{\omega} = \arg \max_{\omega} \log( P(C(\xi) | \omega) P(\omega) )
\]  

(9)

Following the MCE framework, a score function \( g(C(\xi); \omega, \Lambda) \) for the beamformer parameter \( \xi \), hypothesis \( \omega \), HMMs \( \Lambda \) and state sequence \( s \), is defined as:

\[
g(C(\xi); \omega, \Lambda) = \sum_i \log \{ P(C_i(\xi) | s_i) P(s_i | s_{i \ldots}, \omega) \}
\]  

(10)

Then a miss classification function \( d(C(\xi); \omega, \Lambda) \) for a training utterance is defined as:

\[
D(C(\xi); \omega, \Lambda) = -g(C(\xi); \omega, \Lambda) + \log \left( \frac{1}{N - 1} \sum_{\omega \neq \omega_i} \exp \left[ g(C(\xi); \omega, \Lambda) - g(C(\xi); \omega_i, \Lambda) \right] \right)^{\frac{1}{N - 1}}
\]  

(11)

where \( \omega_i \) is correct hypothesis of the input utterance and \( \omega, \omega_i \) is its top \( N \) competitors. Let \( \eta \to \infty \), Eq. (11) could be simplified as:

\[
D(C(\xi); \omega, \Lambda) = -g(C(\xi); \omega, \Lambda) + g(C(\xi); \omega_i, \Lambda)
\]  

(12)
where $\omega_i$ is the most competitive hypothesis.

Finally the recognition error is approximated with a zero-one sigmoid loss function:

$$L(C(\xi); \omega_i, \Lambda) = \frac{1}{1 + \exp(-\alpha D(C(\xi); \omega_i, \Lambda) + \beta)} \quad (13)$$

where $\alpha$ and $\beta$ are the slope and offset of the sigmoid function.

In this way, the MCE loss function is not only a function of the beamformer parameters $\xi$ but also the underlying HMMs $\Lambda$.

### 3.3. MCE OPTIMIZATION

To minimize the MCE loss function, generalized probabilistic decent (GPD) [9] or second order optimization methods could be used to optimize both $\xi$ and $\Lambda$ as long as we have the corresponding gradient function.

First, the gradient of the loss function with respective to $\xi$ is:

$$\frac{\partial L(C(\xi); \omega_i, \Lambda)}{\partial \xi} = (1 - \chi(C(\xi); \omega_i, \Lambda) - L(C(\xi); \omega_i, \Lambda)$$

$$\cdot \frac{\partial D(C(\xi); \omega_i, \Lambda)}{\partial \xi} \quad (14)$$

$$\frac{\partial D(C(\xi); \omega_i, \Lambda)}{\partial \xi} =$$

$$- \frac{\partial g(C(\xi); \omega_i, \Lambda)}{\partial \xi} + \frac{\partial g(C(\xi); \omega_i, \Lambda)}{\partial \xi} \quad (15)$$

To further simplify the computation, HMMs with partition Gaussians models could be adopted, i.e., let $P(C_s(\xi) \mid s_i) = \mathcal{N}(C_s(\xi); \mu_{s_i, \xi}, \sigma_{s_i, \xi})$ for state $s_i$ and its corresponding strongest mixture $j$ at frame $i$. Then, the gradient function becomes:

$$\frac{\partial g(C(\xi); \omega_i, \Lambda)}{\partial \xi} = - \sum_s \left( \frac{C_s(\xi) - \mu_{s_i, \xi}}{\sigma_{s_i, \xi}^2} \right) \frac{\partial C_s(\xi)}{\partial \xi}$$

$$\frac{\partial C_s(\xi)}{\partial \xi} = \sum_s D_{s,a} \frac{\partial Z_s'(\xi)}{\partial \xi} \quad (16)$$

where the derivation of the last term could be found in [4].

On the other hand, the parameters $\Lambda$ of the HMMs could also be adjusted at the same time. If we only consider the $i$-th frame of an utterance, its corresponding state $s_i$ and strongest mixture $j$, the following formulation could be used:

$$\frac{\partial g(C_i(\xi); \omega_s, \Lambda)}{\partial \mu_{s_i, j}} = \frac{\left( C_i(\xi) - \mu_{s_i, j} \right)}{\sigma_{s_i, j}^2} \quad (18)$$

The detail formulation on MCE training of HMMs could be found in [5].

### 4. EXPERIMENTS

To evaluate the proposed S-MCEBEAM method, a Mandarin reverberation corpus created from Mandarin spontaneous speech corpus (TCC300) [6] and RWCP’s sound scene database [7] was used.

Moreover, in order to focus only on array signal processing and discriminative acoustic model training backend, free-syllable decoding task is adopted here. In other words, no language model is utilized in the decoding procedure.

#### 4.1. TCC300 REVERBERANT CORPUS

The TCC300 reverberant corpus for speech recognition experiments was created by convolving the utterances from the TCC300 test set with the room impulse responses recorded by a seven-element linear microphone array, with an inter-microphone spacing of 5.66 cm. The user was directly in front of the array at a distance of 2 m.

This corpus consists of three separate test sets, each corresponding to a different reverberation time. We refer to these test sets as TCC600, where “T60” indicates the 60-dB reverberation time of the room. For example, TCC0.3 represents the test set from a room with a reverberation time of 0.3 second. Each test set consisted of 29 speakers with 5 utterances for array calibration and/or HMM adaptation and 10 utterances for recognition test experiment per speaker.

#### 4.2. EXPERIMENTAL SETTING

Mandarin is a tonal language. Each character is pronounced as a mono-syllable. Because there are only about 408 toneless monosyllables in Mandarin, sub-syllable HMMs were chosen to build the Mandarin recognizer in this paper.

There are 100 right-context-dependent (RCD) initials and 40 context-independent (CI) final HMMs in the Mandarin recognizer. Each initial and final HMM has 3 and 5 states, respectively, and each state has at most 32 mixtures. Beside, 39-dim MFCCs were utilized with frame size 32 ms and frame shift 10 ms. According to our preliminary experiments on TCC300 corpus, this sub-syllable-based recognizer has comparable performance with a tri-phone-based one and achieves 33.03% syllable error rate (SER).

Five different approaches were evaluated including (1) delay-and-sum, (2) S-LIMABEAM, (3) S-MCEBEAM-1, (4) S-MCEBEAM-3 and (5) S-MCEBEAM-3.

The differences between the three S-MCEBEAM variants are:

- S-MCEBEAM-1 uses the beamformer parameters found by S-LIMABEAM as a starting point of MCE training
- S-MCEBEAM-2 skips the S-LIMABEAM step and directly adjusts the beamformer parameters using MCE criterion
- S-MCEBEAM-3 adjusts not only the filter parameters but also the underlying HMMs

For all the following experiments, the number of taps in array filter was set to two. The parameters of these filters were initialized to an impulse function, i.e., $[1, 0, 0]$, and then iteratively optimized.

#### 4.3. EXPERIMENTAL RESULTS

Fig. 2 shows four spectrographic displays of 40-dimensional log mel-spectral feature vectors for a segment of one of the utterances in the TCC0.3 test set. The figure compares the log mel-spectra...
extracted from (a) a close-talking recording, (b) a single microphone from the array, (c) output of a delay-and-sum beamformer and (d) output of the S-MCEBEAM-2 algorithm. As the figure shows, delay-and-sum processing did little to reduce the temporal smearing and equalize the frequency selective response caused by room reverberation. In fact, the delay-and-sum spectrogram is very similar to that of the single microphone. Compared to the close-talking log mel-spectra, some distinctions between high- and low-energy regions across time, especially for mel-band 20–35, have been lost. On the other hand, the features generated by the S-MCEBEAM-2 algorithm look sharper and the low-energy regions between speech segments, especially for mel-band 20–35, have been restored. Beside, the frequency selective response has been significantly equalized.

Table 1 shows the performance comparison on SER between different array processing algorithms for the TCC300 reverberation corpus. As the table shows, S-LIMABEAM and three S-MCEBEAM variants all dramatically improved the performance. Beside, the performances of three S-MCEBEAM variants are all better than the S-LIMABEAM method. It is also worth noting that S-MCEBEAM-2 performed better than S-MCEBEAM-1.

Comparing with the delay-and-sum method, S-MCEBEAM-3 achieved 33.8%, 35.4% and 29.1% relative error reduction (RER) on TCC0.3, TCC0.47 and TCC1.3 test set, respectively. On the other hand, comparing with the more advance S-LIMABEAM approach, S-MCEBEAM-3 still gained 6.4%, 7.8% and 7.8% improvement in RER for TCC0.3, TCC0.47 and TCC1.3 test set, respectively. Therefore, S-MCEBEAM approaches are superior to S-LIMABEAM.

5. CONCLUSION

In this paper, a discriminative training-based S-MCEBEAM approach is proposed to adjust not only the beamformer parameters but also the underlying HMMs. Since, MCE criterion directly matches the goal of minimizing classification error; it therefore leads to better recognition results in reverberation environments. The experiments results on a Mandarin reverberation corpus have also showed the proposed S-MCEBEAM leads to better recognition results than S-LIMABEAM in reverberant environments. S-MCEBEAM is therefore a promising approach.

6. ACKNOWLEDGEMENT

This work was partially supported by the National Science Council, Taiwan, under the project with contract NSC 97-2628-E-027-003-MY3 and 98-2221-E-027-081-MY3.

7. REFERENCES


Table 1: Mandarin syllable error rate (SER in %) obtained using 5 different array processing algorithms on the three test sets of TCC300 reverberation corpus.

<table>
<thead>
<tr>
<th>Test Set</th>
<th>TCC0.3</th>
<th>TCC0.47</th>
<th>TCC1.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Close-talking mic.</td>
<td>33.03</td>
<td>33.03</td>
<td>33.03</td>
</tr>
<tr>
<td>Distant mic.</td>
<td>65.28</td>
<td>113.28</td>
<td>124.34</td>
</tr>
<tr>
<td>Delay-and-sum</td>
<td>64.67</td>
<td>110.46</td>
<td>120.34</td>
</tr>
<tr>
<td>S-LIMABEAM</td>
<td>46.16</td>
<td>79.39</td>
<td>95.54</td>
</tr>
<tr>
<td>S-MCEBEAM-1</td>
<td>45.41</td>
<td>78.02</td>
<td>92.41</td>
</tr>
<tr>
<td>S-MCEBEAM-2</td>
<td>44.88</td>
<td>75.43</td>
<td>90.26</td>
</tr>
<tr>
<td>S-MCEBEAM-3</td>
<td>43.20</td>
<td>73.22</td>
<td>88.13</td>
</tr>
</tbody>
</table>

Figure 2: Log mel-spectrograms of a segment of an utterance from the TCC0.3 test set obtained from (a) a close-talking microphone signal, (b) a single channel in the array, (c) delay-and-sum beamformer, (d) S-MCEBEAM algorithm with two taps per filter.