COMPLIMENTARY COMBINATION OF MICROPHONE ARRAY AND HMM COMPOSITION FOR NOISY SPEECH RECOGNITION

Takanobu Nishiura, Kazuhiro Miki, Satoshi Nakamura, and Kiyohiro Shikano

1 ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai Seika-cho Soraku-gun Kyoto, 619-0288 Japan
2 Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0101 Japan

1. ABSTRACT

Distant-talking speech recognition is very important in providing a natural interface for machines like self-moving robots. Distant-talking speech recognition systems must deal with noises and acoustic reverberations in real environments. A microphone array signal processing and model adaptation method are proposed for distant-talking speech recognition in noisy reverberant environments. However, speech sounds captured by a microphone array are distorted by the directional gain patterns of the microphone array and reverberations in the room. Furthermore, model adaptation would give better performance with high SNR. This paper proposes a method to combine microphone array signal processing with model adaptation methods. A speech recognition experiment in a real room showed that the proposed method provides better performance than conventional methods.

2. INTRODUCTION

It is very important for the natural interfaces of machines like self-moving robots to capture distant-talking speech with high quality and to recognize such speech accurately. However, background noise and room reverberations seriously degrade the sound capture quality in real acoustical environments. This, in turn, degrades automatic speech recognition (ASR) performance in such environments. To overcome this problem, we investigated a noise-reduction capture technique and a design technique for acoustical models that are robust against noise.

For a noise-reduction capture technique, a microphone array is an ideal candidate as an effective method for capturing distant-talking speech. The desired speech signals can be selectively acquired by precisely steering the microphone array in the desired speech direction [1, 2, 3, 4]. Thus, it is often used in the front-end processing of ASR [5, 6]. However, although distant-talking speech can be captured with high quality by a microphone array, it is difficult to recognize such speech accurately in highly reverberant noisy environments. This is because the background noise and room reverberations cannot be completely reduced (i.e., residual signals exist) by only using a microphone array. Also, if the microphone array is steered in the wrong direction, the captured speech includes the signal distortion by the directional gain pattern of the microphone array.

For designing noise-robust acoustical models, the Hidden Markov Model (HMM) decomposition and composition method [7] is proposed as a method to improve ASR performance in reverberant environments. However, although distant-talking speech can be recognized robustly by HMM decomposition and composition in moderately reverberant noisy environments, this method has difficulty in recognizing speech in highly reverberant noisy environments.

To overcome these problems, we propose combining a microphone array with the HMM decomposition and composition method for distant-talking speech recognition. In reverberant noisy environments, distant-talking speech is first enhanced by the microphone array. Then, the noise-reduced speech is recognized robustly by HMM decomposition and composition. At that time, even if the microphone array is steered in the wrong direction, the proposed method can recognize distant-talking speech robustly because the HMM decomposition and composition method can compensate for the effect of directional gain pattern. Also, although the HMM decomposition and composition method has only a slight effect in higher SNR (signal-to-noise ratio) environments, the proposed method can compensate for the performance degradation because the microphone array steering reduces the noise.

3. CONVENTIONAL METHOD

It is difficult to recognize distant-talking speech robustly in reverberant noisy environments. To overcome this problem, a noise-reduction capture technique and a design technique for noise-robust acoustical models are considered. This section describes some conventional methods used individually for these techniques.

3.1. Microphone array steering for noise reduction

A microphone array requires steering (beamforming) to capture distant-talking speech with high quality. Distant-talking speech can be selectively acquired by precisely beamforming in the desired speech direction. Thus, it can be used in the front-end pro-
Therefore, the delay-and-sum beamformer can form directivity to
in Equation
is the sound propagation 5戸ed. Output signal
ent from that of the desired sound signal. Thus, the directivity of
ple transducers are added after synchronizing them. On the other
hand, no other sound signal is
high quality by a m.icrophone aπay, it is diffìcult to recognize that
speech accurately in highly reverberant noisy environments. This
sound signal because the directions of th巴other signals are differ­
the delay-and-sum beamformer can only be formed in direction
the desired talker direction.
In Equation (2), the desired sound signal from direction
is em­
where  \( m \) is the number of transducers and \( c \) is the sound propagation speed. Output signal \( y(t) \) of the delay­
and-sum beamformer is shown in Equation (2).
\[
y(t) = \sum_{m=1}^{M} x_m (t + (m - 1)\tau), \quad \tau = \frac{d \cos \theta}{c},
\]
In Equation (2), the desired sound signal from direction \( \theta \) is em­
phazized \( M \) times because the sound signals captured with multi­
ple transducers are added after synchronizing them. On the other
hand, no other sound signal is \( M \) times as large as the desired sound signal because the directions of the other signals are differ­
ten from that of the desired sound signal. Thus, the directivity of
the delay-and-sum beamformer can only be formed in direction \( \theta \). Therefore, the delay-and-sum beamformer can form directivity to
the desired talker direction.
However, although distant talking speech can be captured with high quality by a microphone array, it is difficult to recognize that
speech accurately in highly reverberant noisy environments. This
is because the background noise and room reverberations can not
be completely reduced (i.e., residual signals exist) by only using
a microphone array. Also, if the microphone array is steered in
the wrong direction, the captured speech may include the signal
distortion by the microphone array’s directional gain pattern.

3.2. HMM decomposition and composition for robust ASR
When using clean speech HMM, ASR performance becomes de­
graded in reverberant noisy environments. To overcome this prob­
lem, a HMM decomposition and composition method has been pro­
banced [7]. Figure 2 shows an overview of the HMM decompo­
sition and composition method, where \( \lambda \) represents the model pa­
rameter (the mean and the deviation for the distribution of output
probability) and \( \lambda_D, \lambda_S, \lambda_H, \) and \( \lambda_N \) represent observed HMM, clean speech HMM, convolutional noise HMM, and additive noise

\[
x_m(t) = x_1 (t-(m-1)\tau), \quad \tau = \frac{d \cos \theta}{c},
\]
where \( m (m = 1, 2, \cdots, M) \) is the number of transducers and \( c \)
is the sound propagation speed. Output signal \( y(t) \) of the delay­
and-sum beamformer is shown in Equation (2).
\[
y(t) = \sum_{m=1}^{M} x_m (t + (m - 1)\tau), \quad \tau = \frac{d \cos \theta}{c}.
\]

<table>
<thead>
<tr>
<th>Table 1: Characteristics of conventional methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delay-and-sum beamformer</td>
</tr>
<tr>
<td>Convolutional noise is generated by directional gain pattern and characteristics of transducers</td>
</tr>
<tr>
<td>HMM decom. &amp; comp.</td>
</tr>
<tr>
<td>Absorbs each noise by statistical model</td>
</tr>
<tr>
<td>CMN</td>
</tr>
<tr>
<td>Does not effect to additive noise</td>
</tr>
<tr>
<td>SS</td>
</tr>
<tr>
<td>HMM, respectively. The HMM decomposition and composition method is combined with the clean speech HMMs after estimat­ing additive noise (e.g., directional noise or ambient noise) model parameters and convolutional noise (e.g., reverberation) model parameters by HMM decomposition. Thus, few adaptation data in the desired environments are needed for estimating the additive noise model parameters and the convolutional noise model parameters. However, reverberant noisy speech can be recognized ro­bustly by composing the clean speech HMM, the additive noise HMM, and the convolutional noise HMM.</td>
</tr>
</tbody>
</table>

3.3. Cepstral mean normalization (CMN) for convolutional noise reduction
Cepstral mean normalization (CMN) [8] is the most common me­
thod for convolutional noisy speech recognition. CMN can reduce
the effect of convolutional noise by removing the cepstrum’s long
time average from the currently observed cepstrum, subject to an
environment with stationary convolutional noise and no additive
noise. Thus, convolutional noisy speech is recognized robustly
with CMN in reverberant environments. However, CMN only has a slight effect in high-additive noise environments.

3.4. Spectral subtraction (SS) for additive noise reduction
Spectral subtraction (SS) [9] is an effective method for additive
noise reduction. SS can reduce the stationary additive noise by
subtracting the long time average of additive noise spectrum from
the currently observed spectrum. However, it is not an effective
method at all for noise reduction in convolutional noise or non­
stationary additive noise environments.

4. PROPOSED METHOD
Table 1 shows the characteristics of conventional methods. Ac­
cording to Table 1, if the CMN or HMM decomposition and com­
position method is used for noisy speech recognition, higher SNR
signals are needed for effective performance due to the demands
of ASR. On the other hand, the delay-and-sum beamformer and SS
can replace lower SNR signals with higher SNR signals because
they can reduce the additive noise. Therefore, we propose combining a microphone array with the
HMM decomposition and composition method for distant-talking
speech recognition. Figure 3 shows an overview of the proposed
system. First, distant-talking speech is enhanced by a delay-and­
sum beamformer in additive and convolutional noisy environments. At that time, the enhanced speech may distort as convolutional
noise due to direction errors or the characteristics of transducers in
delay-and-sum beamforming. Speech enhanced by only us­
ing a delay-and-sum beamformer also includes residual signals
as additive or convolutional noise. To overcome this problem, the proposed method compensates for those residual signals by using the HMM decomposition and composition method. Moreover, although ASR performance with the HMM decomposition and composition method degrades as SNR becomes lower, ASR performance with the proposed system improves because SNR is enhanced by the delay-and-sum beamformer as the pre-process of the HMM decomposition and composition method.

Therefore, the proposed system is designed to improve the ASR performance of distant-talking speech by combining microphone array processing with the HMM decomposition and composition method.

5. EVALUATION EXPERIMENTS

We conducted evaluation experiments in a real acoustic room and evaluated ASR performances to compare the conventional and proposed methods.

5.1. Experimental conditions

We evaluated ASR performance with various conventional methods and the proposed method. In this paper, we conducted evaluation experiments under the conditions described below.

1. Only convolutional noise environment
2. Convolutional noise and stationary additive noise environment
3. Convolutional noise and non-stationary additive noise environment

We conducted the evaluation experiments in the real acoustic room shown in Figure 4. Table 2 shows the experimental conditions. The MFCC parameter in the HMM decomposition and composition method is composed of that used for clean speech HMMs. The ΔMFCC and Δpower parameters in this method are also those used for clean speech HMMs.

5.2. Experimental results: only convolutional noise environment

We evaluated ASR performance in an environment with only convolutional noise. Ambient noise level was 22 dBA in this experimental environment. In this environment, not only reverberation exists as convolutional noise but convolutional noise was also generated by steering the microphone array in the wrong direction. Figure 5 shows these results, where DS represents delay-and-sum beamformer processing. As a result, ASR performance was significantly degraded by steering the microphone array in the wrong direction. However, ASR performance can be improved by using the HMM decomposition and composition method, although it is degraded by using clean speech HMM. Next, when the CMN and HMM decomposition and composition methods are compared, the HMM decomposition and composition method was about 4~8% more effective than CMN.

5.3. Experimental results: convolutional noise and stationary additive noise environment

We evaluated ASR performance in an environment with convolutional noise and stationary additive noise. In this environment, the microphone array is steered in the known desired speech direction. However, ASR performance can be improved by using the HMM decomposition and composition method, although it is degraded by using clean speech HMM. Next, when the CMN and HMM decomposition and composition methods are compared, the HMM decomposition and composition method was about 4~8% more effective than CMN.
5.4. Experimental results: convolutional noise and non-stationary additive noise environment

Finally, we evaluated ASR performance in an environment with convolutional noise and non-stationary additive noise. In this environment, the microphone array was also steered in the known desired speech direction. Figure 7 shows these results. As a result, in an SNR = 10 dB environment, speech recognition rate (i.e., Word Recognition Rate (WRR)) was 71.4% with the delay-and-sum beamformer. However, WRR improved to 80.6% with both delay-and-sum beamformer and SS+CMS. In addition, we confirmed that the proposed method using the HMM decomposition and composition method instead of SS+CMS, can also improve WRR from 80.6% to 84.5%. Thus, we confirmed that the proposed method is more effective than the conventional methods.

6. CONCLUSIONS

In this paper, we proposed a complimentary combination of a microphone array and the HMM decomposition and composition method for distant-talking speech recognition in real environments. To evaluate the proposed system, evaluation experiments were conducted in a real acoustic room. As a result, we could confirm that ASR performance of the proposed method is higher than that of the conventional methods. Thus, we confirmed that the proposed method can improve ASR performance by combining a microphone array with the HMM decomposition and composition method.

REFERENCES