Submitted for publication
April 2002

IDIAF-RR 02-09
Herre Bouwland, t
Ian A. McGregor, t
Andrew Morris 1

TECHNIQUES
USING MISSING DATA
SMALL MICROPHONE ARRAYS
RECOGNITION PERFORMANCE OF
IMPROVING SPEECH
Improving Speech Recognition Performance of Small Microphone Arrays Using Missing Data Techniques

Submitted for Publication
April 2002

Liam A. McEwan
Andrew Worth
Herve Bourlard

NIHR Research Report 02-09
2.1 Overview

In automatic speech recognition (ASR), some of the observation data for the target signal can be masked by noise signals. One way to compensate for noise in ASR is to train the recognizer on data which has been corrupted by multiple different noise conditions. A limitation of this approach is that it is not as effective on clean speech at the test noise conditions than on clean speech training.

A second approach which avoids this restriction is to use models trained on clean speech. However, these models are not able to maintain a high level of accuracy on clean speech. This is due to the noise masking effect; the noise estimation can be used to improve recognition performance in two ways. One is to enhance the model trained on clean speech. The second technique is to use noise estimation of the unknown clean speech data. This approach has been shown to be effective on clean speech at the test noise conditions than on clean speech training.

The performance of the proposed ASR approach to robust speech recognition is equally applicable in stationary or non-stationary noise conditions. Section 2.2 describes the standard procedure for calculating an SNR-based missing data mask. Section 3 then details the use of the proposed technique for generating a missing data mask suitable for a microphone array. The performance of the proposed technique in missing data experiments using a small microphone array is investigated using a baseline single channel missing data system, as well as a baseline single channel missing data system.
3 Reliability Mask Estimation Using Microphone Arrays

The success of the missing data estimation approach depends critically on the calculation of the

where $a$ is the gain of the smoothing function.

\[
\frac{(p-g)(t)_{\text{true}} + 1}{I} = (q)^t_d
\]

To determine the missing data estimates, these values are fed into the range $0 \leq 1$

\[
g > (q)^t_{\text{true}} \quad : 0 \quad \text{and then the ratio of signal to noise energy is in the range is}
\]

\[
\log \left( \frac{(q)^{t+1}_{\text{true}} - (q)^{t+1}_{\text{true}}} {\text{signal}} \right) = (q)^{t+1}_d
\]

\[
\begin{align*}
\text{where $\text{SNR} = \text{SNR}_\text{true}$ for the missing data estimation of the wave function.}\n\end{align*}
\]

\[
\begin{align*}
\text{The right-hand side of the equation is then evaluated.}
\end{align*}
\]

\[
\begin{align*}
\text{where $\text{SNR}_\text{true}$ is the estimated SNR for the missing data estimation in the range.}
\end{align*}
\]

\[
\begin{align*}
\text{and the ratio of signal to noise energy is in the range is}
\end{align*}
\]

\[
\begin{align*}
\text{and then the ratio of signal to noise energy is in the range is}
\end{align*}
\]
3.3 Proposed SNR Mask Estimation Technique

The proposed technique is a combination of the methods developed in [9] and [10]. This combination allows us to estimate the post-filter output signal accurately and efficiently. The post-filter output signal is given by

\[ (f) \hat{y} = (f) z \]

(9)

\[ (f)x_{d}(f)w = (f)\theta \]

(10)

where \( f \) is the beamforming weight vector, \( \hat{y} \) is the output data vector, and \( \theta \) is the weight vector.

3.1 Microphone Array Speech Enhancement

In this section, we propose a technique using a microphone array to enhance speech. In this approach, the microphone array is used to capture the background noise and then apply a post-filter to the output signal. The output signal is then used to estimate the noise and subtract it from the original signal. The estimated noise is then used to enhance the speech signal.

Figure 1: Post-filter beamformer with post-filter

![Post-filter beamformer with post-filter](image-url)
4. Experiments and Results

For many elements, the desired speaker was simulated directly in front of the loudspeakers, and a

- frequency spectrum of 20,000 Hz was used in experiments consisted of 4 head-sized microphones, with their

4.1. Conjugation

The microphone array used in experiments consisted of 4 head-sized microphones, with

(a) md−real mask

(b) md−single mask

(c) md−array mask

Figure 2: Sample SNR Mixture (input SNR=0dB, mixture is ‘one’.

\[ (y)^*u - (y)^*z = (y)^*u \]

\[ \log \left( \frac{(y)^*u}{(y)^*z} \right) = \log (y)^*u \]

determine the mismatch can only partially be transmitted as noise. The noise that is not transmitted is estimated as

case for standard microphone array speech recognition systems, we instead propose using them to

\[ (y)^*u - (y)^*z = (y)^*u \]
For evaluating the possible fractional mask using the microphone array (indoor) is better than

4.2 Discussion

recognition results are given in Table 1 and plotted in Figure 4 in terms of the percentage word error.

The various recognition results were tested at different input SNR's of 0, 10 and 0 dB. The speech
interference and noise interference in a mask has a clear correlation to the real SNR
improvement of DB (dB) is shown in Figure 2. An average of 10 trials for each condition of SNR (dB) of 0, 10, and 20 and for different number of ga = 0.5% average of the three different signal generation at an
SNR threshold.

For all mismatched channel conditions, sets of mimicry were used according to Equation 1, with an SNR threshold.

7. mismatch detection on the noisy features with SNR mask generated from a prior knowledge.

4. mismatch detection on the noisy features with SNR mask generated from the microphone.

by comparing the noise spectrum over the first 10 input frames (indirect).

3. mismatch detection on the noisy features with SNR mask generated from a single channel.

2. standard recognition using the enhanced and original features, (indirect).

1. standard recognition using noisy input features from a single microphone (direct).

For each input to one of the microphone in the array.

The effect of different, giving a figure of frequency of the speech given on each rate of 0.5%
which are used to detect the frequency with best performance. Selecting the best of each combination of the order of the masks.

We calculated the speech input from each microphone. There input frequencies were used to calculate the multi-channel.

and the input frequency processing over the best performing frequency enhancement in magnitude. In

The SNR was measured in dB and the SNR was used to determine the best-performing microphone in dB,

Table 1: Speech recognition results

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>10 DB</th>
<th>0 DB</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
5 Conclusions

As a final point, we note that there exists room for some further improvement by tuning of the $g$ and $\lambda$ parameters of the mask calculation.

Instead of purely combinatorial, the strategy used in each case is to equip the learner with the knowledge only for multi-level estimation, thereby enabling the learner to achieve the best performance possible. The results of this strategy are shown in Table 1.

The second important trend we observe is the performance of the proposed learning (and) multichannel noise conditions.

In conclusion, we recommend that the reader refer to the full paper for a detailed analysis of the results presented in this section.


