Non-Audible Murmur recognition based on fusion of audio and visual streams

Panikos Heracleous and Norihiro Hagita

ATR, Intelligent Robotics and Communication Laboratories, 2-2-2 Hikaridai Seika-cho, Soraku-gun, Kyoto-fu 619-0288, Japan
(panikos, hagita)@atr.jp

Abstract

Non-Audible Murmur (NAM) is an unvoiced speech signal that can be received through the body tissue with the use of special acoustic sensors (i.e., NAM microphones) attached behind the talker’s ear. In a NAM microphone, body transmission and loss of lip radiation act as a low-pass filter. Consequently, higher frequency components are attenuated in a NAM signal. Owing to such factors as spectral reduction, the unvoiced nature of NAM, and the type of articulation, the NAM sounds become similar, thereby causing a larger number of confusions in comparison to normal speech. In the present article, the visual information extracted from the talker’s facial movements is fused with NAM speech using three fusion methods, and phoneme classification experiments are conducted. The experimental results reveal a significant improvement when both fused NAM speech and facial information are used.

1. Introduction

Non-Audible Murmur (NAM) refers to a very softly uttered speech received through the body tissue. A special acoustic sensor (i.e., the NAM microphone) is attached behind the talker’s ear. This receives very soft sounds that are inaudible to other listeners who are in close proximity to the talked.

The attachment of the NAM microphone to the talker is shown in Figure 1. The first NAM microphone was based on stethoscopes used by medical doctors to examine patients, and was called the stethoscopic microphone [1]. Stethoscopic microphones were used for the automatic recognition of NAM speech [2]. The silicon NAM microphone is a more advanced version of the NAM microphone [3]. The silicon NAM microphone is a highly sensitive microphone wrapped in silicon; silicon is used because its impedance is similar to that of human skin. Silicon NAM microphones have been employed for automatic recognition of NAM speech as well as for NAM-to-speech conversion [6]. Similar approaches have been introduced for speech enhancement or speech recognition [4, 5]. Further, non-audible speech recognition has also been reported based on electromyographic (EMG) speech recognition, which processes electric signals caused by the articulatory muscles [7].

The speech received by a NAM microphone has different spectral characteristics in comparison to normal speech. In particular, the NAM speech shows limited high-frequency contents because of body transmission. Frequency components above the 3500-4000 Hz range are not included in NAM speech. The NAM microphone can also be used to receive audible speech directly from the body (Body Transmitted Ordinary Speech (BTOS)). This enables automatic speech recognition in a conventional way while taking advantage of the robustness of NAM

Figure 1: NAM microphone attached to the talker

against noise.

Previous studies have reported experiments for NAM speech recognition that produced very promising results. A word accuracy of 93.9% was achieved for a 20k Japanese vocabulary dictation task when a small amount of training data from a single speaker was used [2]. Moreover, experiments were conducted using simulated and real noisy test data with clean training models to investigate the role of the Lombard reflex [8, 9] in NAM recognition. The HMM distances of NAM sounds in comparison with the HMM distances of normal speech were also investigated, which indicated distance reduction when NAM sounds were concerned [10]. In the same study, preliminary results obtained using audio-visual data for NAM recognition based on concatenative feature fusion were also reported.

In the present study, audio-visual NAM recognition is further investigated by using the multistream HMM decision fusion and late fusion to integrate the audio and visual information. A statistical significance test was performed, and audio-visual NAM recognition in a noisy environment was also investigated.

2. Methodology

2.1. Corpus and HMM modeling

The corpus used in the experiment was 212 continuous Japanese utterances, containing 7518 phoneme realisations. A 3-state with no skip HMM topology was used. Forty-three monophones were trained using 5132 phonemes. For the purpose of testing, 2386 phonemes were used. The audio parameter vectors were of length 36 (12 MFCC, 12ΔMFCC, and 12 ΔΔMFCC). The HTK3.4 Toolkit was used for training and testing.
2.2. Extracting visual parameters

The face and profile views of the subject were filmed under conditions of good lighting. The system captured the 3-D positions of 112 colored beads glued on the speaker’s face at a sampling rate of 50 Hz (fig. 2), synchronized with the acoustic signal sampled at 16000 Hz. The collection of 30 lip points using a generic 3-D geometric model of the lips is shown in Figure 3 [12].

The shape model is built using the Principal Component Analysis (PCA). Successive applications of PCA are performed on the selected subsets of the data, which generate the main directions. These directions are retained as linear predictors for the whole data set. The mobile points P of the face deviate from their average position B by a linear composition of the basic components M loaded by factors $\alpha$ (articulatory parameters) [13].

$$P = B + \alpha M \quad (1)$$

Only the first 5 parameters of the extracted 12 linear components M were used. These explained more than 90% of the data variance using the following iterative linear prediction on the data residual: the first component of the PCA on the lower teeth (LT) values leads to the “first jaw” predictor. The PCA on the residual lips values (without jaw influence) usually presented three pertinent lip predictors (i.e., lips protrusion, lips closing mainly required for bilabials, and lips raising mainly required for labiodental fricatives). The movements of the throat linked the underlying movements of the larynx and the hyoid bone, and served as the fifth one. The video parameters were interpolated at 200 Hz to synchronize with the audio analysis frame rate.

2.3. Fusion methods

This section introduces the fusion methods used for integrating the audio NAM signal with the visual parameters. In the present study, a feature fusion method and two decision methods, namely state-synchronous and state-asynchronous decision fusion methods, were used.

2.3.1. Concatenative feature fusion

The feature concatenation is the simplest state synchronous fusion method. It uses the concatenation of the synchronous audio NAM speech signal and visual signal as the joint feature vector:

$$O^{NV}_t = [O^{(N)}_t, O^{(V)}_t]^T \in R^D \quad (2)$$

where, $O^{NV}_t$ is the joint NAM-visual feature vector, $O^{(N)}_t$ is the NAM feature vector, $O^{(V)}_t$ is the visual feature vector, and $D$ is the dimension of the joint feature vector. In these experiments, the dimension of the NAM stream was 36 and the dimension of the visual stream was 15. Thus, the dimension $D$ of the joint NAM-visual feature vectors was 51.

2.3.2. Multi-stream HMM fusion

Multi-stream HMM fusion is a state synchronous decision fusion, which captures the reliability of each stream by combining the likelihoods of single-stream HMM classifiers [11]. The emission likelihood of the multi-stream HMM is the product of the emission likelihoods of the single-stream components, weighted appropriately by stream weights. Given the $O$ combined observation vector, that is, the NAM and visual elements, the emission probability of multi-stream HMM is given by:

$$b_j(O_t) = \prod_{s=1}^{S} \left( \sum_{m=1}^{M_s} c_{jsm} N(O_{st}; \mu_{jsm}, \Sigma_{jsm}) \right)^{\lambda_s} \quad (3)$$

where, $N(O; \mu, \Sigma)$ is the value in $O$ of a multivariate Gaussian with mean $\mu$ and covariance matrix $\Sigma$, and $S$ is the number of the streams. For each stream $s$, $M_s$ Gaussians in a mixture are used, each weighted with $c_{jsm}$. The contribution of each stream is weighted by $\lambda_s$. In the present study, it is assumed that the stream weights do not depend on state $j$ and time $t$. However, two constraints were applied, namely:

$$0 \leq \{\lambda_n, \lambda_v\} \leq 1, \text{ and } \lambda_n + \lambda_v = 1 \quad (4)$$

Table 1: Comparison of the fusion methods

<table>
<thead>
<tr>
<th>Fusion Method</th>
<th>Late</th>
<th>Multistream</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonemes</td>
<td>71.8</td>
<td>68.9</td>
<td>67.8</td>
</tr>
<tr>
<td>Vowels</td>
<td>86.2</td>
<td>83.7</td>
<td>83.5</td>
</tr>
<tr>
<td>Consonants</td>
<td>64.1</td>
<td>59.7</td>
<td>58.2</td>
</tr>
</tbody>
</table>
where $\lambda_n$ is the NAM stream weight, and $\lambda_v$ is the visual stream weight. In these experiments, the weights were experimentally adjusted to 0.6 and 0.4 values, respectively. The selected weights were obtained by maximizing the accuracy on several experiments.

### 2.3.3. Late fusion

A disadvantage of the previously described fusion methods is the assumption that there is a synchrony between the two streams. In the present study, late fusion was applied to enable asynchrony between the NAM stream and the visual stream. In the late fusion method, two single HMM-based classifiers were used for the NAM speech and the visual speech, respectively. For each test utterance (i.e., isolated phone), the two classifiers provided an output list, which included all the phone hypotheses with their likelihoods. Subsequently, all the separate monomodal hypotheses were combined into the bi-modal hypotheses using the weighted likelihoods, as given by:

$$
\log P_{NV}(h) = \lambda_n \log P_N(h|Q_N) + \lambda_v \log P_V(h|Q_V)
$$

where, $\log P_{NV}(h)$ is the score of the combined bi-modal hypothesis $h$, $\log P_N(h|Q_N)$ is the score of the $h$ provided by the NAM classifier, and $\log P_V(h|Q_V)$ is the score of the $h$ provided by the visual classifier. $\lambda_n$ and $\lambda_v$ are the stream weights with the same constraints applied in multi-stream HMM fusion.

The procedure described in this study finally resulted in a combined N-best list in which the top hypothesis was selected as the correct bi-modal output. A similar method was also introduced in [11].

### 3. Results

#### 3.1. Experiments in a clean environment

In this section, the experimental results of the phoneme classification in a clean environment are presented.

Table 2: Confusion matrix of Japanese plosives using NAM speech.

<table>
<thead>
<tr>
<th></th>
<th>/p/</th>
<th>/b/</th>
<th>/t/</th>
<th>/d/</th>
<th>/k/</th>
<th>/g/</th>
</tr>
</thead>
<tbody>
<tr>
<td>/p/</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>/b/</td>
<td>0</td>
<td>8</td>
<td>5</td>
<td>3</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>/t/</td>
<td>2</td>
<td>0</td>
<td>36</td>
<td>13</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>/d/</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>14</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>/k/</td>
<td>1</td>
<td>10</td>
<td>8</td>
<td>0</td>
<td>45</td>
<td>6</td>
</tr>
<tr>
<td>/g/</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>20</td>
</tr>
</tbody>
</table>

Table 3: Confusion matrix of Japanese plosives using NAM-visual speech.

<table>
<thead>
<tr>
<th></th>
<th>/p/</th>
<th>/b/</th>
<th>/t/</th>
<th>/d/</th>
<th>/k/</th>
<th>/g/</th>
</tr>
</thead>
<tbody>
<tr>
<td>/p/</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>/b/</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>/t/</td>
<td>0</td>
<td>0</td>
<td>39</td>
<td>1</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>/d/</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>17</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>/k/</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>50</td>
<td>3</td>
</tr>
<tr>
<td>/g/</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>22</td>
</tr>
</tbody>
</table>

#### 3.1.1. Comparison of fusion methods

A comparison of the three classification methods used in the present study is shown in Table 1. As seen in the table, the highest classification accuracies are achieved when late fusion is used. The second best classification accuracies are achieved when using multi-stream HMM decision fusion. Finally, the lowest accuracies are observed when using feature fusion. Specifically, when using late fusion, an accuracy of 71.8% is achieved for phoneme classification, 86.2% accuracy for vowel classification, and 64.1% accuracy for consonant classification. The highest accuracies, when using late fusion, might be an evidence of asynchrony between the NAM speech and the visual stream. In the following experiments late fusion is used to integrate the NAM audio speech with the visual data.

#### 3.1.2. Phoneme classification using audio, visual, and audio-visual information

The results obtained when using visual data, NAM data, and visual-NAM data are shown in Figure 4. The results indicate that the classification accuracy is very low when only visual data is used. As many sounds appear to be similar on the lips/face, the sole use of visual parameters cannot distinguish these sounds. In the case of NAM data, the accuracies are higher in comparison to visual data. Specifically, an accuracy of 79.2% was achieved for vowel recognition, 49.8% accuracy for consonant recognition, and 59.7% accuracy for phoneme recognition. It is observed that the accuracy is considerably lower for consonant recognition in comparison to vowel recognition. However, because of the unvoiced nature of NAM, both voiced and unvoiced sounds articulated at the same place become similar, resulting in a larger number of confusions between consonants. The significant improvements in accuracy, when visual data were fused with NAM speech, are shown in Figure 4. Specifically, a relative improvement of 33% was achieved for vowel recognition, 28% for consonant recognition, and 30% for phoneme recognition.

Table 2 and Table 3 show the confusion matrices of the
Figure 5: Phoneme classification in noisy environment.

Figure 5: Phoneme classification in noisy environment.

In this study, the experimental results for NAM recognition in both clean and noisy environments are presented. Owing to the unvoiced nature of NAM and the place of articulation, NAM sounds articulated in the same place become more similar. This results in a larger number of confusions between the sounds, which consequently leads to a lower accuracy in comparison to normal speech. However, by integrating the visual information of the lips/face, these sounds can be better discriminated. As a result, the accuracy drastically increases. The experimental results obtained using audio-visual NAM data indicate about 30\% relative improvement in accuracy in the case of a clean environment, and to a 15\% absolute increase in a noisy environment.

5. References