Lip Reading Using a Dynamic Feature of Lip Images and Convolutional Neural Networks

Yiting Li, Yuki Takashima, Tetsuya Takiguchi, Yasuo Ariki
Kobe University, Japan 657–8501
Email: {liyiting, y.takashima}@me.cs.scitec.kobe-u.ac.jp, {takigu, ariki}@kobe-u.ac.jp

Abstract—In this paper, a lip-reading method using a novel dynamic feature of lip images is proposed. The dynamic feature of lip images is calculated as the first-order regression coefficients using a few neighboring frames (images). It constitutes a better representation of the time derivatives to the basic static image. The dynamic feature is processed by using convolution neural networks (CNNs), which are able to reduce the negative influence caused by shaking of the subject and face alignment blurring at the feature-extraction level. Its effectiveness has been confirmed by word-recognition experiments comparing the proposed method with the conventional static (original) image.

I. INTRODUCTION

Lip reading is the technique of understanding what someone is saying by interpreting or “reading” the movements of the speaker’s lips. When people with normal hearing are talking to each other, they do not only use voice information, but also pay attention to the movement of the other person’s face (especially the lips) under noisy conditions. Therefore, there is a great need for an approach that will help people understand one another when they cannot understand what others are saying simply from information obtained by hearing the speaker’s voice.

Most of lip-reading approaches, including an image-based approach or a model-based approach, employ a continuous image sequence [1], [2], [3], [5]. Y. Pei, et al. [1] proposed a technique to handle lip reading using random forest manifold alignment, where the small patches around the lips are extracted using AAM, resulting in patch trajectories. H. L. Bear, et al. [2] introduced a phoneme-clustering method to form new phoneme-to-viseme maps for both individual and multiple speakers, where a lip-only feature is extracted by using AAM from full-face images. K. Noda, et al. [3] applied higher-level and model-based features extracted using ASM and AAM, which are derived from the shape and appearance of mouth area images. Faridah F., et al. [16] proposed lip image feature extraction based on a “snake”, which is a contour model. In our previous work [5], the continuous lip images are used for multimodal speech recognition.

In this paper, we propose a novel approach using a dynamic feature that is distinct from the original continuous image, where a dynamic feature represents a difference image. A difference image was often used in object recognition, tracking and detection [14], [15]. In this way, the moving object is extracted clearly because the shade and color of the moving object is distinct from the background.

Feature extraction is an indispensable part of any lip-reading approach. Various approaches for visual feature extraction have been proposed. Discrete Cosine Transform (DCT) feature extraction has been used in [9]. Principle Component Analysis (PCA) [10], Discrete Wavelet Transform (DWT) [10], and Linear Discriminant Analysis (LDA) [9] have also been used for lip reading. In recent years, neural networks have been increasingly used in image classification, image recognition, speech recognition, lipreading [4], [12], [13], etc. Convolutional Neural Networks (CNNs) have also been used in lip reading [3], [5]. CNNs have much fewer connections and parameters than traditional neural networks. Therefore, it is much easier to train the network. K. Noda, et al. proposed a seven-layer CNN for audio-visual speech recognition [3].

This paper proposes a lip-reading approach using a dynamic feature and convolutional neural networks. An original image sequence of lips, which is commonly used in lipreading research, is replaced with a dynamic feature image. For feature extraction, convolutional neural networks were used to reduce the negative influence caused by shaking of the subject and blurring of camera or face alignment.

The rest of this paper is organized as follows: In Section II, lip image extraction is described. In Section III, the dynamic feature is described. In Section IV, the CNN architecture is described. In Section V, the experimental data are evaluated, and the final section is devoted to our conclusions.

II. LIP IMAGE EXTRACTION

Extracting a lip area from the whole face is an essential part of lipreading process. By face alignment, each face part (lip, eyes, etc.) will be located. In this paper, a Constrained Local Model (CLM) was used, which we applied in our previous approach [5].

The face alignment model that we employed in this paper is based on Point Distribution Model (PDM). There are two steps to capture the lip area of the image: face point detection and parameter estimation.

PDM models a facial image by 2-dimensional shape vectors. The position vector can be defined as matrices composed by position of each point of PDM. \( X_i \) denotes the \( i \)-th point of PDM and \( M \) indicates the number of points of PDM.

\[
X = (X_1^T, \ldots, X_M^T)^T \tag{1}
\]

\[
X = \bar{X} + \Phi q \tag{2}
\]
Here, $\Phi$ indicates the principle vectors extracted by Principal Component Analysis (PAC). $q$ and $\bar{X}$ denote the parameter vector and the mean vector of the shape vector, respectively. By using PDM, the $i$-th point on the image, $X_i(p)$ is represented as follows:

$$X_i(p) = sR(\bar{X}_i + \Phi_i q) + t$$  \hspace{1cm} (3)

where $p = \{s, R, t, q\}$ indicates the parameter set. $s$ denotes a scale, and $R$ denotes a rotation which consists of pitch $\alpha$, yaw $\beta$, roll $\gamma$. $t$, $q$, and $\Phi_i$ denote the shift vector, the parameter vector and the $i$-th principal vector, respectively.

$$Q(p) = \sum_{i=1}^{M} \left\| \bar{X}_i + X_i(p) \right\|^2 + R(p)$$  \hspace{1cm} (4)

The parameter of PDM is estimated by using CLM. The model parameter $p$ is estimated from the $i$-th detected feature point $\hat{X}_i$ by minimizing the $Q(p)$, where $R(p)$ is a regularization term to avoid over fitting.

### III. Dynamic Feature

A difference image is frequently used in the case of detecting an object in motion. In this paper, a novel dynamic feature, which is obtained by calculating the first-order regression coefficients, is used for lip reading. As shown in Fig. 1, the first order regression coefficients are calculated using the current frame and $\pm 3$ frames (in this example). Examples of original images and dynamic images are shown in Figs. 2 and 3, respectively.

![Fig. 2. Original image of lips](image)

![Fig. 3. Dynamic image of lips](image)

We consider, using a data sequence $y$, to calculate the first order regression coefficient based on the least squares method. The mean square sum of the error, $E$, is obtained by approximating the data sequence to a straight line.

$$E = \frac{1}{n} \sum_{i=1}^{n} \{(ai + b) - y_i\}$$  \hspace{1cm} (5)

Now, in order to determine $a$ and $b$ so that $E$ is minimized, partial differential of $E$ for $a$ and $b$ is calculated.

$$\frac{\partial E}{\partial a} = \frac{2}{n} \sum_{i=1}^{n} \{(ai + b) - y_i\} \cdot i = 0$$  \hspace{1cm} (6)

From (6) and (7), the formulas will be obtained.

$$\frac{\partial E}{\partial b} = \frac{2}{n} \sum_{i=1}^{n} \{(ai + b) - y_i\} = 0$$  \hspace{1cm} (7)

$$\frac{1}{n} \sum_{i=1}^{n} i^2 \cdot a + \frac{1}{n} \sum_{i=1}^{n} i \cdot b = \frac{1}{n} \sum_{i=1}^{n} (i \cdot y_i)$$  \hspace{1cm} (8)

$$\frac{1}{n} \sum_{i=1}^{n} i^2 \cdot a + \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{n} \sum_{i=1}^{n} y_i$$  \hspace{1cm} (9)

By eliminating $b$ by using (8) and (9), the gradient of the straight line (regression coefficient) can be obtained as follows:

$$a = \frac{\frac{1}{n} \sum_{i=1}^{n} i \cdot y_i - \frac{1}{n} \sum_{i=1}^{n} i}{\frac{1}{n} \sum_{i=1}^{n} i^2 - \frac{1}{n} \sum_{i=1}^{n} i^2} \sum_{i=1}^{n} y_i$$  \hspace{1cm} (10)

Next, we consider a dynamic feature of images, as shown in Fig. 1. By using (10), the $i$-th regression coefficient at the current frame $n$, $\Delta C_i(n)$, is given by
\[ \Delta c_i(n) = \left( \frac{1}{2\delta+1} \right) \sum_{k=-\delta}^{\delta} (n+k) \cdot c_i(n+k) - \left( \frac{1}{2\delta+1} \right) \sum_{k=-\delta}^{\delta} (n+k) \left( \frac{1}{2\delta+1} \right) \sum_{k=-\delta}^{\delta} c_i(n+k) \] 
\[ \quad \quad / \left( \frac{1}{2\delta+1} \right) \sum_{k=-\delta}^{\delta} (n+k)^2 \] 
\[ \quad \quad - \left( \frac{1}{2\delta+1} \right) \sum_{k=-\delta}^{\delta} c_i(n+k)^2 \] 

(11)

where \( c_i(n) \) represents the \( d \)-th pixel value at the current frame, \( n \), and \( \delta \) is the number of the frames before and after the current frame. The \( d \)-th regression coefficient is calculated by using the \( d \)-th pixel values from \( (n-\delta) \) to \( (n+\delta) \). In this method, the dynamic feature of lips is obtained by calculating the approximate gradient (regression coefficient) between frames.

Because of \( \sum_{k=-\delta}^{\delta} k = 0 \), we have (12) and (13).

\[ \frac{1}{2\delta+1} \sum_{k=-\delta}^{\delta} (n+k) = \frac{1}{2\delta+1} \sum_{k=-\delta}^{\delta} n + \frac{1}{2\delta+1} \sum_{k=-\delta}^{\delta} k = n \] 

(12)

\[ \frac{1}{2\delta+1} \sum_{k=-\delta}^{\delta} (n+k)^2 \] 

(13)

The overview of this procedure can be expressed as Fig. 4. The vertical axis denotes the frame number, while the horizontal denotes the value of pixel \( i \) at each frame. By approximating the value of these points to a straight line, the \( i \)-th regression coefficient at the current frame \( n \), can be equivalent to the slope of this line.

IV. CONVOLUTIONAL NEURAL NETWORKS

A. Architecture of CNNs

Convolutional neural networks (CNNs) with bottleneck architecture was used for lipreading in our previous work [5]. The CNN in our approach consists of an input layer, a convolution layer, a pooling layer, a fully-connected Multi-Layer Perceptron (MLP) with a bottleneck structure and an output layer as shown in Fig. 5. The CNN which contain these sort of architecture are also called convolutive bottleneck network (CBN). C1, P1, and m denote a convolution layer, a pooling layer, and MLPs, respectively. The MLP shown in Fig. 5 consists of three layers (m1, m2, m3), and the number
of units in the layer \( m2 \), is reduced as “bottleneck feature”. The number of units in each layer will be discussed in Section V. As the bottleneck layer has reduced the number of units for the adjacent layer, we can anticipate that every single unit in the bottleneck layer aggregates information and behaves as a compact feature descriptor which represents an input with a small number of bases. For the output of the CBN, phoneme labels are used. The bottleneck (BN) feature in the trained CBN is used for training HMMs for lipreading.

**B. Bottleneck Feature Extraction**

First, visual CBNs are trained by using the images as the input data, where phoneme labels are used as the output units of the CBN. For example, if we have the utterance /i/ in the data, the phoneme label can indicate where is the pronunci /i/ by only set the unit /i/ 1 and the other set to 0. The label data are obtained by forced alignment using HMMs from the speech data.

For the visual features, because of the depression of the sampling rate of the video, we adopt spline interpolation to the images to fill the sampling rate gap.

The parameters of the CBNs are trained by back-propagation with stochastic gradient descent, where the parameter value is initialized to random numbers. The bottleneck (BN) features obtained by the trained CBNs will be used in the training of HMMs for recognition. In the test stage, the features extracted by using the CBNs, which tries to produce the appropriate phoneme labels in the output layer. At the recognition stage, the BN features in the middle layer are used, where it is considered that information in the input data is aggregated. Finally, the extracted bottleneck visual feature is used as the input feature of visual HMMs.

**V. EXPERIMENTS**

**A. Experimental Conditions**

Our proposed method was evaluated on word-recognition tasks for a male. We recorded 2,620 words included in the ATR Japanese speech database A-set, which were used as training data, and 216 words included in the ATR Japanese speech database B-set, which were used as test data. The sampling rate of video was 60 fps. We compared our method with a Discrete Cosine Transform (DCT) feature and a feature obtained from the CNN using original (static) images. In each method, 2,620 words from the ATR B-set database were used for training, 216 words from B-set of ATR database were used for test. For recognition, we employed the monophone-HMMs with 5 states and 8 mixtures of Gaussians. The number of phonemes was 54.

The grey-scale image of lips is extracted from the image of frontal face by using the CLM approach as illustrated in section II. Then, the image is up-sampled by spline interpolation due to the low frame rate of the video. As shown in Fig. 5, the input data of the network are resized to 12×24. Moreover, the number of dimensions for the feature is 30. Two types of CNN architecture were tested. Each feature map and the number of units in each layer for the two CNNs are shown in Tables I and II, respectively. As shown in Tables I and II, CNN1 has a pair consisting of the convolution layer and pooling layer, while CNN2 consists of the architecture of two convolution layers and two pooling layers. Both CNN1 and CNN2 have a fully-connected MLP with a bottleneck structure.

**TABLE I**

<table>
<thead>
<tr>
<th>Input</th>
<th>C1</th>
<th>P1</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 12×24</td>
<td>13, 8×20</td>
<td>13, 4×10</td>
<td>108</td>
<td>30</td>
<td>108</td>
<td>54</td>
</tr>
</tbody>
</table>

**TABLE II**

<table>
<thead>
<tr>
<th>Input</th>
<th>C1</th>
<th>P1</th>
<th>C2</th>
<th>C2</th>
<th>P2</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 12×24</td>
<td>13, 10×22</td>
<td>13, 5×11</td>
<td>27, 3×9</td>
<td>27, 1×13</td>
<td>108</td>
<td>30</td>
<td>108</td>
<td>54</td>
<td></td>
</tr>
</tbody>
</table>

974
Fig. 6. Word-recognition rates using two types of CNNs for the static feature and the dynamic feature. The dynamic feature was obtained using CNNs.

B. Results and Discussions

The results of word recognition are shown in Fig. 6, where the static feature and dynamic feature were recognized by using two kinds of architectures. As shown in this figure, the result for CNN1 is a slightly better than that of CNN2. Thus, the CNN1 architecture has been applied to the remaining experiments.

Fig. 7 shows the recognition results for the dynamic features, the static feature, and the DCT feature, where the vertical axis represents word-recognition accuracies and the horizontal axis shows $\delta$ in (14). The static feature used in Fig. 7 was obtained using the same CNN architecture used for the dynamic feature, but the input data of the CNNs are the original image sequence. The number of dimensions of static feature and DCT feature are 30, which is the same as dynamic feature. The $\delta$ of the dynamic feature used here is changed from 2 to 9.

As shown in Fig. 7, our proposed method obtained a higher score than the static feature and the DCT feature. The best accuracy 71.76% was obtained when $\delta$ was set to 3, which is 12.82% higher than the accuracy obtained by the static feature and 19.91% higher than that of the DCT feature.

Fig. 8 shows the word-recognition rates of the dynamic features and the combined features, where the dynamic feature and the static feature are combined. Both of the features are 30-dimensional, so the number of dimensions of the combined features is 60. These results demonstrate that when we combine these two kinds of features, the recognition rates decrease.

VI. Conclusion

This paper proposed a novel lip-reading method using the dynamic image feature of lips and CNNs. Experimental results indicated that the proposed method is more effective than two other methods: the conventional static feature and the DCT feature. In future work, we will apply our method to multi-speaker word recognition and images from any angles.

REFERENCES


