Speech Pattern Discovery using Audio-Visual Fusion and Canonical Correlation Analysis

Lei Xie¹, Yinqing Xu¹, Lilei Zheng¹, Qiang Huang² and Bingfeng Li¹

¹School of Computer Science, Northwestern Polytechnical University, Xi’an, China
²School of Computing Sciences, University of East Anglia, Norwich, United Kingdom

lxie@nwpu.edu.cn, xyzqzki@hotmail.com, llzheng@nwpu-aslp.org, h.qiang@uea.ac.uk

Abstract

In this paper, we address the problem of automatic discovery of speech patterns using audio-visual information fusion. Unlike those previous studies based on single audio modality, our work not only uses the acoustic information, but also takes into account the visual features extracted from the mouth region. To improve the effectiveness of the use of multimodal information, several audio-visual fusion strategies, including feature concatenation, similarity weighting and decision fusion, are utilized. Specifically, our decision fusion approach retains the reliable patterns discovered in the audio and visual modalities. Moreover, we use canonical correlation analysis (CCA) to address the issue of temporal asynchrony between audio and visual speech modalities and unbounded dynamic time warping (UDTW) is adopted to search for the speech patterns through audio and visual similarity matrices calculated on the aligned audio and visual sequence. Experiments on an audio-visual corpus show that, for the first time, speech pattern discovery can be improved by the use of visual information. The decision fusion approach shows superior performance compared with standard feature concatenation and similarity weighting. CCA-based audio-visual synchronization plays an important role in the performance improvement.

Index Terms: Speech pattern discovery, canonical correlation analysis, audio-visual speech processing, dynamic time warping

1. Introduction

Discovery of term repetitions is the basic paradigm of many spoken document processing tasks, e.g., topic classification, spoken document retrieval, spoken term detection and speech summarization. Typically, term repetitions are discovered from automatic speech recognition (ASR) transcripts, either at word or subword levels [1]. However, building an ASR requires considerable linguistic resources and a tremendous amount of annotated training data to learn acoustic and language models.

Moreover, distinctive terms in spoken documents often lie outside the lexicon of an off-the-shelf speech recognizer and recognition errors are also inevitable. Therefore, researchers have been interested in discovering recurrent terms using unsupervised, low-resource methods without building a speech recognizer [2, 3, 4, 5, 6, 7].

Unsupervised speech pattern discovery [2] has recently drawn much interest in the literature. It aims to identify automatically long, reliably repeated patterns in the acoustic signal. A more powerful version of dynamic time warping, segmental dynamic time warping (SDTW), has been used to search for acoustic repetitions. Crucially, the discovered acoustic repetitions often correspond to meaningful terms, e.g., words and phrases. Moreover, this method does require neither a phonetically-interpretable acoustic model nor any knowledge of the target language. Speech pattern discovery approaches have been successfully adopted in document clustering and classification [5], story segmentation [8] and spoken term detection [6, 9]. However, finding acoustic repetitions using SDTW is a time-consuming task. Recent approaches have been focusing on how to improve its running efficiency [7, 3, 4]. Anguera et al. [7] have proposed an unbounded DTW (UDTW) algorithm that is able to improve the accuracy by over 9% while running almost 10 times faster, as compared with SDTW.

In this paper, we present a pilot study on speech pattern discovery using audio-visual speech information. As we know, speech production and perception are bimodal in nature. Plenty of previous work has shown that integration of bimodal speech information achieves performance improvement in tasks such as speech recognition and speaker identification [10]. Therefore, when video of speakers’ faces is available, it is natural to integrate audio and visual information in speech pattern discovery. In this paper, we have tested several audio-visual fusion strategies, including feature concatenation, similarity weighting and decision fusion. Specifically, the decision fusion method first searches for potential patterns using UDTW on the individual inter-utterance audio and visual similarity matrices and then retains the reliable ones as the discovered speech patterns. It is a well known fact that audio and visual speech activities are asynchronous [10]. To account for asynchrony, we use canonical correlation analysis (CCA) [11] as a preprocessing step to align audio and visual features and to maximize the audio-visual correlation before pattern discovery. We demonstrate, for the first time, that the performance of speech pattern discovery can be improved by the use of visual information. Specifically, the decision fusion approach shows superior performance and CCA-based audio-visual synchronization plays an important role in this performance gain.

2. UDTW for Speech Pattern Discovery

We use UDTW [7] to discover the pattern repetitions between two utterances. Given two utterances, X and Y, we can represent them as two time series of feature vectors, (x₁, ⋯, xₙ) and (y₁, ⋯, yₘ), respectively. n and m denote the number of frames in X and Y. We use a similarity matrix S to measure the similarities between the frame vectors of the two utterances, in which element (i, j) is defined as:

$$s_{ij} = \cos(x_i, y_j) = \frac{x_i^T y_j}{||x_i|| ||y_j||}.$$  \hspace{1cm} (1)

Here, s_{ij} is the cosine similarity between the i'th frame of X and the j'th frame of Y. Figure 1(a) shows a dotplot of a similarity matrix.
matrix calculated between the MFCC vectors of two utterances. Pattern repetitions are then searched on matrix S.

UDTW aims to find multiple alignment paths between two utterances. Firstly, diagonal lines with inter-line horizontal distance $R$ are selected in the similarity matrix $S$ and points on these lines are used as start points for path searching, as shown in Figure 1(a). Secondly, from each start point, we look for path extension forward and backward respectively, as described in [7]. A point $d$ is added to a current considered path $p$ and generates a new path $p'$ when the following conditions are met: (1) the average similarity score $\bar{\sigma}_p$ of path $p'$ is bigger than any previous paths across the point $d$; (2) the average similarity score $\bar{\sigma}_p$ is bigger than a threshold $\theta$. If there is no previous path across $d$, only condition (2) needs to be met. Finally, we reserve only paths with length greater than a threshold $L$. More details for the UDTW algorithm can be found in [7]. Figure 1 (b) shows the reserved paths after path searching in Figure 1 (a). Projections of each path on the two utterances are regarded as the discovered speech patterns in the two utterances.

3. Audio-Visual Synchronization using CCA

Audio and visual speech activities are temporally asynchronous [10]. For example, when we start to talk, our mouth first opens and then sounds come out (visual precedes audio); when we end a talk, sounds first vanish and then the mouth closes (visual follows audio). Previous studies of audio-visual fusion has demonstrated apparent performance improvement when accounting for this asynchrony [10]. In this paper, we use canonical correlation analysis (CCA) as a preprocessing step to align audio and visual feature frames and maximize the bimodal correlation before pattern discovery.

CCA is a linear statistical analysis technique that measures how much and in what directions two given multidimensional variables are correlated [11]. Given a time series of acoustic feature vectors $X_a = (x_a^1, \ldots, x_a^N)$ and a time series of visual feature vectors $X_v = (x_v^1, \ldots, x_v^N)$, we use CCA to calculate an overall audio-visual correlation measure $\gamma_{av}$ between them. Suppose $X_a$ and $X_v$ can be treated as two Gaussian variables with dimensions $N_a$ and $N_v$, respectively. CCA seeks two linear transformations $H_a$ and $H_v$ to maximize the mutual information between the transformed variables $X_a'$ and $X_v'$:

$$X_a' = H_aX_a, \quad X_v' = H_vX_v$$

where the transformations $H_a$ and $H_v$ are matrices with dimensions $N \times N_a$ and $N \times N_v$ respectively and $N \leq \min(N_a, N_v)$.

The rows of the projections $X_a(i)$ and $X_v(j)$ ($1 \leq i,j \leq N$) are defined as pairs of canonical components and the canonical correlations between each pair are calculated [11]:

$$\gamma_i = E(X_a'(i)X_v'(j))$$

where $E(\cdot)$ is the expected value function. The overall correlation measure $\gamma_{av}$ between $X_a$ and $X_v$ is:

$$\gamma_{av} = \sum_{i=1}^{N} \gamma_i^2$$

For a speech segment, assume that there is a time gap $\tau$ between the audio and visual features, i.e., the $k$th visual feature vector $x_v^k$ corresponds to the $(k+\tau)$th audio feature vector $x_a^{k+\tau}$. We apply CCA to $x_a^{k+\tau}$ and $x_v^k$ with varying values of $\tau$, and calculate the value of the correlation measure $\gamma_{av}(\tau)$ for each $\tau$. Finally, we select the optimal temporal gap $\tau$ which maximizes $\gamma_{av}(\tau)$, i.e., the audio-visual correlations.

We align the audio and visual feature sequences as follows. We firstly divide the audio feature vector sequence $X_a$ into several segments by detecting short silences. Secondly, we search the visual feature vector sequence $X_v$ via CCA to find the optimal temporal gap $\tau$ between the corresponding audio and visual segments. Finally we concatenate the visual segments to achieve an ’audio-synchronized’ visual feature vector $X_v'$. Figure 2 shows the diagram of the alignment process. Audio-visual fusion strategies can thus be implemented on the synchronized audio and visual feature sequences, $X_a$ and $X_v'$.

4. Audio-Visual Fusion Strategies

4.1. Feature Concatenation

A straightforward fusion strategy is to concatenate audio and visual features, $x_a^k$ and $x_v^k$, into a new feature vector $[x_a^k, x_v^k]$.
Pattern discovery is then performed on the similarity matrix $S^{av}$ calculated using equation (1) on the concatenated feature vectors.

### 4.2. Similarity Weighting

Another natural idea is to weight the audio and visual similarity matrices before pattern discovery. Since cosine similarity is bounded between 0 and 1, we use the following feature weighting strategy:

$$s_{ij}^{av} = w \times s_{ij}^a + (1 - w) \times s_{ij}^v$$

where $w$ denotes the weighting factor that is tuned empirically. Subsequently, pattern discovery is performed on the similarity matrix $S^{av}$.

### 4.3. Decision Fusion

We can also carry out audio-visual fusion on the paths discovered using UDTW in both audio and visual modalities, and view this kind of fusion as decision fusion. For path $p_i^m$ in modality $m$ ($m \in \{a, v\}$), we achieve a block $B_i^m$ that contains $p_i^m$, as shown in Figure 3 (a) and (b). For block $B_i^m$, we then check in the counterpart modality whether the corresponding block area also contains a path. If paths within a same block in both audio and visual modalities can be found, we will regard it as a valid speech pattern discovery. Figure 3 shows an example of decision fusion that helps to remove a false alarm. As can be seen in Figure 3, since "white" and "five" have similar pronunciations, a false alarm occurs in audio-only. The two words have quite different lip movements, so the false alarm does not occur in visual-only. Finally, decision fusion is able to remove this false alarm.

![Figure 3: Repeated patterns (color lines) discovered by UDTW. Since “white” and “five” have similar pronunciations, a false alarm occurs in audio-only. The two words have quite different lip movements, so the false alarm does not occur in visual-only. Finally decision fusion is able to remove this false alarm.](image)

### 5. Experiments

#### 5.1. Experimental Setup

We carried out experiments on the GRID audiovisual corpus\footnote{http://www.dcs.shef.ac.uk/spandh/gridcorpus} that contains facial recordings of speakers reading short sentences. Each recording lasts for about 3 seconds. The original audio sampling rate (44.1KHz) was down-sampled to 16KHz and the video frame rate was 25 fps. In this paper, we selected 696 videos of the 23rd speaker as the experimental data. A total of 696 videos of the 23rd speaker as the experimental data. A total of 696 videos of the 23rd speaker as the experimental data.

We evaluated eight approaches: audio only (AO), visual only (VO), audio-visual feature concatenation (AVFC), audio-visual similarity weighting (AVSW), audio-visual decision fusion (AVDF) and their revised versions with CCA preprocessing (AVFC+CCA, AVSW+CCA and AVDF+CCA). We first conducted empirical parameter tuning on the development set to obtain the optimal parameter setting that achieved the best performance of speech pattern discovery, and carried out evaluation on the test set using the best-tuned parameters. Note that the parameters for AO, VO and AVFC are diagonal band width...
Table 1: Experimental results of speech pattern discovery.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
<th>F1 Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>AO</td>
<td>53.28%</td>
<td>52.94%</td>
<td>53.11%</td>
<td>-</td>
</tr>
<tr>
<td>VO</td>
<td>54.43%</td>
<td>53.01%</td>
<td>53.62%</td>
<td>40.96%</td>
</tr>
<tr>
<td>AVFC</td>
<td>54.25%</td>
<td>53.27%</td>
<td>53.03%</td>
<td>3.62%</td>
</tr>
<tr>
<td>AVFC+CCA</td>
<td>54.80%</td>
<td>55.16%</td>
<td>54.77%</td>
<td>5.01%</td>
</tr>
<tr>
<td>AVSW</td>
<td>62.79%</td>
<td>50.16%</td>
<td>55.77%</td>
<td>+5.01%</td>
</tr>
<tr>
<td>AVSW+CCA</td>
<td>57.24%</td>
<td>55.11%</td>
<td>56.15%</td>
<td>+5.72%</td>
</tr>
<tr>
<td>AVDF</td>
<td>52.34%</td>
<td>64.77%</td>
<td>57.90%</td>
<td>+9.02%</td>
</tr>
<tr>
<td>AVDF+CCA</td>
<td>53.02%</td>
<td>65.13%</td>
<td>58.45%</td>
<td>+10.05%</td>
</tr>
</tbody>
</table>

AO: Audio Only; VO: Visual Only; AVFC: A-V Feature concatenation; AVSW: A-V Similarity Weighting; AVDF: A-V Decision Fusion; CCA: Canonical Correlation Analysis; F1 Gain: Relative F1 gain as compared to AO.

R, similarity threshold θ and the length threshold of alignment path L. For AVSW and AVSW+CCA, w is an additional parameter. In AVDF and AVDF+CCA, we tuned two sets of parameters, i.e., \( \{R_A, \theta_A, L_A\} \) and \( \{R_V, \theta_V, L_V\} \).

5.2. Results

Experimental results are summarized in Table 1. Some important observations are summarized as follows. The use of only visual information (VO) shows inferior speech pattern discovery performance. The recall score for this technique is comparable to the audio-only (AO) approach but the precision score is low. This is because the discrimination power of mouth movements is relatively low. However, we can clearly see that all audio-visual information integration approaches outperform the audio-only and visual-only approaches. Especially, decision fusion demonstrates superior performance when compared with feature concatenation and similarity weighting. Furthermore, we can see that CCA also leads to improved performance. Figure 5 gives an example that demonstrates the effectiveness of CCA. After CCA, the discovered path is correctly synchronized with the acoustic boundaries. We notice that the decision fusion approach with CCA preprocessing (AVDF+CCA) achieves the highest F1-score with a relative performance gain of 10.05% as compared with the audio-only approach (AO).

6. Conclusions

In this paper, motivated by previous success in audio-visual speech processing, we use visual speech information to improve the performance of speech pattern discovery. Firstly, we align the audio and visual feature sequences using canonical correlation analysis to account for the temporal asynchrony between the two modalities. Secondly, we use unbounded dynamic time warping to search for repeated speech patterns in the inter-sentence audio and visual similarity matrices. Finally, we integrate audio and visual information for pattern discovery, using feature concatenation, similarity weighting and decision fusion. Experiments on an audio-visual corpus demonstrate that the performance of speech pattern discovery can be considerably improved by the use of visual information. Decision fusion achieves superior performance as compared with feature concatenation and similarity weighting. Crucially, CCA-based audio-visual synchronization greatly contributes to the performance improvement. In the future, we will extend our work to a large speech corpus and study audio-visual pattern discovery in speech from multiple speakers.

7. Acknowledgements

This work is supported by the National Natural Science Foundation of China (61175018), the Natural Science Basic Research Plan of Shaanxi Province (2011JM8009), the Key Science and Technology Program of Shaanxi Province (2011KJXX29) and the Fok Ying Tung Education Foundation (131059).

8. References


Figure 5: The effectiveness of CCA. After CCA, the discovered path is synchronized with acoustic boundaries.