Real-Time Speech-Driven Lip Synchronization

Kaihui Mu¹, Jianhua Tao¹, Jianfeng che², Minghao Yang¹
¹National Laboratory of Pattern Recognition (NLPR)  
²Hi-tech Innovation Center  
Institute of Automation, Chinese Academy of Sciences  
Beijing, China 
{khmu, jhtao}@nlpr.ia.ac.cn, {jianfeng.che, minghao.yang}@ia.ac.cn

Abstract—Speech-driven lip synchronization, an important part of facial animation, is to animate a face model to render lip movements that are synchronized with the acoustic speech signal. It has many applications in human-computer interaction. In this paper, we present a framework that systematically addresses multimodal database collection and processing and real-time speech-driven lip synchronization using collaborative filtering which is a data-driven approach used by many online retailers to recommend products. Mel-frequency cepstral coefficients (MFCCs) with their delta and acceleration coefficients and Facial Animation Parameters (FAPs) supported by MPEG-4 for the visual representation of speech are utilized as acoustic features and animation parameters respectively. The proposed system is speaker independent and real-time capable. The subjective experiments show that the proposed approach generates a natural facial animation.

Keywords—real-time speech-driven lip synchronization; FAP; MFCC; collaborative filtering

I. INTRODUCTION

Human speech is bimodal in nature: audio and visual [1]. Speech is produced by the vibration of the vocal cord and the configuration of the vocal tract that is composed of articulatory organs, including the nasal cavity, tongue and lips. Since some of these articulators are visible, there is an inherent relationship between the acoustic and visible speech, which is also demonstrated by the McGurk effect [2]. The goal of lip synchronization is to animate a face model to render lip movements that are synchronized with the acoustic speech signal. Lip synchronization, also known as visual speech synthesis, has been developed for improving human-computer interaction, computer-aided instruction, virtual announcer, video games, and multimedia telephony for hearing-impaired individuals and so on.

Two types of driven sources are usually utilized as input to a facial animation system: text and audio, i.e., text and speech-driven face animation [3]. Text-driven face animation is often applied to an automatic and intelligent dialogue system, in which the speech is usually rendered by a TTS system. Since human speakers can create more expressive speech due to their capability to adapt intonation, emphasis and pause easily to the semantics of the spoken words, more and more researchers pay attention to animate a face model from real speech.

In order to synthesize lip synchronization from speech, many algorithms have been developed, including linear prediction analysis [4], vector quantization [5], Gaussian mixture model [6], artificial neural networks (ANNs) [7], and hidden Markov models (HMMs) [10].

Vector quantization, which is applied to divide the acoustic training data into a number of classes, is a classification-based conversion approach. Each new input acoustic vector would be classified into one of these classes, which map to a corresponding visual output. This approach often leads to a discontinuous mapping result.

The Gaussian mixture approach models the joint probability distribution of the audio-visual vectors as a Gaussian mixture. Given an audio feature, each Gaussian mixture component generates a linear estimation for a visual feature. The estimations of all the mixture components are then weighted to produce the final estimation of the visual feature. This approach will produces smoother results than vector quantization method.

ANNs are another common method to convert acoustic features into animation parameters. In the training phase, back propagation algorithm is employed to training the network weights. The number of hidden layers and the number of nodes per layer may be experimentally determined.

HMMs are the most common method to transform audio features into a stream of animation parameters for a facial model due to its capability of handling voice contextual information and modeling coarticulation in synthetic visual speech. But HMMs based approaches have relative long time delay.

In this paper, we present a real-time speech-driven lip synchronization system using a data-driven approach based on k-nearest neighbors (kNN), which is called collaborative filtering [13]. An audiovisual database, which uses frames as units, is built firstly. Then, a new visual sequence is constructed by concatenating the appropriate weighted visual frames from the database selected by collaborative filtering when acoustic signals are input into this system. This approach creates lip synchronization in real-time using only one acoustic frame. It is real-time capable. The concatenation of the weighted and smoothed visual frames leads to a continuous mapping result. One of the advantages of this approach is to retain the original temporal and energy structure of the speech.

This work is supported in part by National Natural Science Foundation of China under Grant 60873160, and in part by Project 863 under Grant 2009AA01Z320.
Thus the naturalness of rhythm in lip synchronization is guaranteed. Further benefit is that the synthesis system is speaker independent. Both male and female voices are able to drive any MPEG-4 compliant face model to render lip movements.

In the remaining of the paper, we first describe multimodal data collection and processing (Section II) and collaborative filtering algorithm with its similarity metrics (Section III). Next, the synthesized results and the subjective evaluation are presented in Section IV. Finally, the conclusions are drawn and future work is proposed in Section V.

II. DATA COLLECTION AND PROCESSING

A. Data Collection

A commercially available motion capture system (Motion Analysis) with 8 cameras (a) of Figure 1. is utilized to track 25 markers on the face of an actress, as shown in (b) of Figure 1., at a rate of 75 frames per second. These markers are placed according to the MPEG-4 feature point locations [14]. We obtain the 3D trajectories for each of the marker points as the output of the tracking system. The audio data is recorded at the same time with a sample rate of 16 KHz. Overall the database, 694 sentences are uttered by the performer in the neutral state during a series of motion capture sessions.

B. FAPs Extraction

The MPEG-4 Facial Animation specification [15] provides a set of FAPs for animating any MPEG-4 compliant face model. The Facial Definition Parameters (FDPs) are defined by the locations of the feature points and are used to customize a given face model to a particular face. Each FAP value is simply the displacement of a particular feature point from its neutral position expressed in terms of the Facial Animation Parameter Units (FAPUs). The FAPUs correspond to fractions of distances between key facial features. Thus, once we have the displacements of the feature points from the neutral position, it is easy to extract the FAPs corresponding to the given facial animation. Detailed description about MPEG-4 standard and FAPs can be found in [14].

In order to get the precise displacements of the feature points, we need to compensate the global head movements. Firstly, the first frame in each session is chosen to be the reference frame. Next, all the markers are translated so that the marker point placed at nose is at the local coordinate center in each frame of one session. Then, the Procrustes analysis [16] is employed to measure rotation matrix for each frame. Finally, the rotation matrix is applied to compensate the global head movements for each frame.

The distances between some key facial features are calculated as FAPUs in the reference frame according to MPEG-4 standard. Then each FAP value is measured by the displacement of a particular feature point in each frame. In this work, 25 FAPs in MPEG-4 (shown in TABLE I.) are extracted to represent the lip movements.

C. Acoustic Features Encoding

MFCCs, commonly used as features in speech recognition systems, are chosen as the acoustic features in this work. MFCCs are a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency.

The HTK tools [17] are utilized to compute 12 cepstral coefficients of MFCCs and the energy term. Their delta and acceleration components are also considered to model temporal dynamic changes in the speech signal. The acoustic signals are processed using a 20 ms Hamming window with overlapping frames of 10 ms.

The frame rate of FAPs is upsampled from 75 to 100 frames per second. Then, the synchronized FAPs and MFCCs with their delta and acceleration components are combined to form MFCC-FAP corpus.

III. COLLECTIVE FILTERING ALGORITHM

A. Collaborative Filtering

A collaborative filtering algorithm usually works by search a large group of samples to finding a smaller set similar to the input sample. It looks at features corresponding to each sample in this smaller set and combines them to create a ranked list of features. Then, these ranked features are weighted to form new features. In this work, MFCCs with their derivatives are used

<table>
<thead>
<tr>
<th>FAP No.</th>
<th>FAP name</th>
<th>FAP No.</th>
<th>FAP name</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>open_jaw</td>
<td>16</td>
<td>Push_b_lip</td>
</tr>
<tr>
<td>4</td>
<td>lower_t_midlip</td>
<td>17</td>
<td>Push_t_lip</td>
</tr>
<tr>
<td>5</td>
<td>raise_b_midlip</td>
<td>51</td>
<td>lower_t_midlip_o</td>
</tr>
<tr>
<td>6</td>
<td>stretch_l_cornerlip</td>
<td>52</td>
<td>raise_b_midlip_o</td>
</tr>
<tr>
<td>7</td>
<td>Stretch_r_cornerlip</td>
<td>53</td>
<td>stretch_l_cornerlip_o</td>
</tr>
<tr>
<td>8</td>
<td>lower_t_lip_lm</td>
<td>54</td>
<td>stretch_r_cornerlip_o</td>
</tr>
<tr>
<td>9</td>
<td>lower_t_lip_rm</td>
<td>55</td>
<td>lower_t_lip_lm_o</td>
</tr>
<tr>
<td>10</td>
<td>raise_b_lip_lm</td>
<td>56</td>
<td>lower_t_lip_rm_o</td>
</tr>
<tr>
<td>11</td>
<td>raise_b_lip_rm</td>
<td>57</td>
<td>raise_b_lip_lm_o</td>
</tr>
<tr>
<td>12</td>
<td>raise_l_cornerlip</td>
<td>58</td>
<td>raise_b_lip_rm_o</td>
</tr>
<tr>
<td>13</td>
<td>raise_r_cornerlip</td>
<td>59</td>
<td>raise_l_cornerlip_o</td>
</tr>
<tr>
<td>14</td>
<td>thrust_jaw</td>
<td>60</td>
<td>raise_r_cornerlip_o</td>
</tr>
<tr>
<td>15</td>
<td>shift_jaw</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
as samples and the corresponding FAPs are utilized as the features of those samples.

The algorithm is described as follows:

- The samples and the corresponding features are collected.
- The similar samples are found using some similarity metric as to a new input sample.
- The similar sample set is ranked according to the similarity values to form the first k items of the sorted results.
- The features of each sample in the ranked sample set are scored by multiplying the similarity value for each ranked sample.
- Then, the total of these features is divided by the sum of all the similarities. The mathematical expression shows below:

\[
    f_{ij}^k = \frac{\sum_{j=1}^{k} w_{ij} f_{ij}}{\sum_{j=1}^{k} w_{ij}}
\]

Here, \( f_{ij} \) and \( w_{ij} \) are the feature vector and the similarity value of the \( j \)th sample in the \( i \)th sample set. \( f_{ij} \) is the generated feature vector of the \( i \)th new input sample.

### B. Similarity Metric

To determine how similar two samples are, two similarity metrics, Euclidean distance and Pearson correlation, are usually utilized to compute similarity scores. Their mathematical expressions are given below:

\[
    r = \frac{1}{1+\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}}
\]

\[
    r = \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 \right) \left(\sum_{i=1}^{n} y_i^2\right) - n \left(\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \right)}}
\]

Here, \( x_i \) and \( y_i \) are the items in two vectors that both have the same dimension. \( r \) is the similarity score between these two vectors. Higher value of \( r \) indicates being more similar. The Pearson correlation coefficient is a measure of how well two sets of data fit on a straight line, which tends to give better results in situations where the data are not well normalized. These two similarity metrics are utilized for collaborative filtering algorithm in the experiments.

### IV. Experiments and Subjective Evaluation

In the experiments, 4272 frames of MFCC-FAP pairs with silence removed are selected from the MFCC-FAP corpus to synthesize the new FAPs. When new audio signals input, the MFCCs and their first and second order derivatives are encoded as the new sample input into the collaborative filtering. Then, new FAPs are synthesized as the features corresponding to the new samples.

#### A. Selection of Similarity Metric

As introduced in III.B, two metrics are employed to measure the similarity score between two samples. To get accurate results, the MFCCs with their first and second order derivatives are normalized by the maximum value of each dimension. The mean square error (MSE) and the time are utilized to evaluate these two metrics. MSE is applied to measure the error between new generated FAPs and the original FAPs corresponding to the same MFCCs. The time to create new FAPs is related to the frame number of MFCC-FAP pairs used in synthesis stage, the similarity metric and the \( k \) value using in collaborative filtering. The \( k \) value is set 5 in this experiment to balance the time consuming and the accuracy of FAPs. A new acoustic signal with 226 frames is input into the system. This experiment is done in Matlab in a computer with AMD 4000+ CPU and 1GB RAM. The results are shown in TABLE II.

From the results, the Euclidean distance has smaller error and shorter time-consuming, which is employed in this work. The Pearson correlation has bigger error because the normalization of MFCC is made before weakening the advantage of the Pearson correlation. In real system, this algorithm is implemented in C language and it can generate real-time lip movements rendered in any MPEG-4 compliant head model when real-time acoustic signal is recorded.

#### B. Smoothness of the Synthesized FAPs

The discontinuous results in generated FAP sequence can be smoothed across several adjacent frames. We smooth the past 2 frames, the present frame and the future 2 frames to get the current FAPs to drive face model. The energy in acoustic signal is also employed to weight the FAPs in each frame to get the final FAPs, which retains the energy structure for lip movements. The original and final created trajectories of FAP 3 (open_jaw) for a sentence are shown in Figure 2. Several frames of lip movements rendered in a 2-D head model using the created FAPs are also shown in Figure 3.

<table>
<thead>
<tr>
<th>Similarity metric</th>
<th>MSE</th>
<th>Time-consuming (second)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean distance</td>
<td>20925.75</td>
<td>62.7</td>
</tr>
<tr>
<td>Pearson correlation</td>
<td>21955.98</td>
<td>133.6</td>
</tr>
</tbody>
</table>
C. Subjective Evaluation

In the subjective experiment, each of 8 sentences is uttered by 8 different people (6 women and 2 men). Then, all of lip movements rendered in a 3-D head model for each of these 64 sentences are evaluated by another 8 people on a five level mean opinion score (MOS) scale. The score is account to one decimal place. The average value of the 8 scores given by the 8 testers is shown in TABLE III.

In TABLE III, the MOS scores in each row except the last two rows are the scores for the same sentence uttered by 8 different people (F for women and M for men). The minimum and maximum scores in each row are emphasized by the italic font. The scores in each column are for eight sentences uttered by the same person, where the minimum and maximum scores are highlighted by the bold font. The mean and variance values of the eight utterances given by the same person are exhibited in the last two rows.

From the results, it is noted that F1 has the minimum MOS score (2.9) and mean score (3.43) and the maximum variance (0.09). M1 has the maximum MOS score (4.6) and mean score (4.39) and the minimum variance (0.02). These differences are mainly because that there is different speed rate and pitch between different people. The F1 has the smaller MOS scores due to its bigger tone difference from the acoustic signal in the multimodal database recorded by another woman. However, overall the evaluation, the MOS scores show that our system can generate acceptable and natural lip movements.

V. DISCUSSION AND CONCLUSION

In this paper, a data-driven approach based on collaborative filtering is utilized to create FAPs in real time for any MPEG-4 compliant head model from a bimodal database. This system is able to retain the original temporal and energy structure of the speech. It is also speaker independent. The synthesis experiment and subjective evaluation show that our system can generate real-time lip synchronization and natural lip movements in any MPEG-4 compliant head model.

There are also some issues that need to study further. Firstly, the noise reduction should be researched deeply to make the system available for various noisy environments. The MFCC-MMSE [18] estimator is a good solution to this problem. Next, the real-time features are guaranteed well, but the size of the corpus is very large that needs to be cut further for computing effectively. Finally, the co-articulatory is not well modeled in this work, which limits the expression of this system. We should study further for the co-articulatory modeling.
ACKNOWLEDGMENT

Thanks to Mi Zhou for collecting multimodal database and some processing. Also thanks to Yi Zhang as an actress for performing all the sessions. Final thanks to people who take participate in the evaluations in our work.