Photo-Realistic Visual Speech Synthesis Based on AAM Features and an Articulatory DBN Model with Constrained Asynchrony

Peng Wu 1,2, Dongmei Jiang 1,2, He Zhang 1,2, Hichem Sahli 3,4

VUB-NPU Joint Research Group on AVSP
1 Northwestern Polytechnic University, Xi’an 710072, China
2 Shaanxi Provincial Key Laboratory on Speech, Image and Information Processing
3 Vrije Universiteit Brussel (VUB) - AVSP, Department ETRO,
4 Interuniversity Microelectronics Centre – IMEC,
VUB-ETRO Pleinlaan 2, 1050 Brussels, Belgium

Abstract

This paper presents a photo realistic visual speech synthesis method based on an audio visual articulatory dynamic Bayesian network model (AF_AVDBN) in which the maximum asynchronies between the articulatory features, such as lips, tongue and glottis/velum, can be controlled. Perceptual linear prediction (PLP) features from the audio speech and active appearance model (AAM) features from mouth images of the visual speech are adopted to train the AF_AVDBN model for continuous speech. An EM-based optimal visual feature learning algorithm is deduced given the input auditory speech and the trained AF_AVDBN parameters. Finally, photo realistic mouth images are synthesized from the learned AAM features. In the experiments, mouth animations are synthesized for 30 connected digit audio speech sentences. Subjective evaluation results show that by considering the asynchronies between articulatory features in the AF_AVDBN (as well between audio and visual states in the SA_DBN), the synchronization between the audio speech and mouth animations are well obtained. Moreover, since AF_AVDBN captures the dynamic movements of articulatory features and model the pronunciation process more precisely, the accuracy of the mouth animations from the AF_AVDBN is much higher than those from the SA_DBN and SS_DBN models, very accurate, clear, and natural mouth animations can be obtained through the AF_AVDBN model and AAM features.

Index Terms: visual speech synthesis, AF_AVDBN, asymynchrony, AAM features

1. Introduction

Speech driven talking face animation has become a popular research topic in human-computer interaction. Various methods have been proposed to obtain MPEG-4 compliant 3D or photo realistic 2D face animations. Moreover, several virtual avatars have been tentatively used in real life applications, such as SynFace[1], Baldi[2], CrazyTalk[3] and VideoRewrite[4]. The talking head animation systems have an essential problem, lip synchronization. The synthesized mouth movement of the talking head has to match the corresponding context and intensity of the input speech. The simple key frame based approaches [5, 6] cause jerky mouth animations since they cannot capture the real speech dynamics or coarticulation. To overcome the jerky phenomenon, Massaro [2] adopted the dominance functions leading to a blending over time of the articulatory commands related to adjacent segments. Instead of concatenating key frames of visemes, some talking-face systems adopted other units containing context information, such as diphones or triphones [4], as well divisemes (viseme pairs) [7, 8]. The disadvantage of these methods is that a large amount of captured data is required to produce natural results.

Some researchers adopt the machine learning strategy by considering mouth synching as an audio-to-visual conversion problem. E.g. Choi et al. [9] exploited the Hidden Markov Model inversion (HMMI) technique for an MPEG-4 facial animation system. Given an audio input and the trained multi-stream HMM (MSHMM) parameters, visual parameters are learned based on the Maximum Likelihood Estimation (MLE) of an auxiliary function. Later, Terissi et al. [10] expanded the HMMI technique to a general case of full covariance matrices, and proposed a speech driven MPEG-4 compliant facial animation system. In [11], a similar audio to visual conversion approach was performed using an audio visual dynamic Bayesian network model with articulatory features (AF_DBN), in which all the articulatory features (AFs) are asynchronous without any constraint. Natural and realistic mouth animations have been obtained. However, the assumption that all the AFs move with unlimited asynchrony along the sentence does not fit the physical mechanism of the articulator organs. This has been considered by Livescu et al. in [12] who proposed an audio visual DBN model with articulatory features (AF_AVDBN) for speech recognition, in which the maximum asynchrony between the AFs can be controlled. The authors demonstrated the validity of the model by producing high recognition rates. However, the authors only discussed the structure of the AF_AVDBN model, and did not define clearly the conditional probability distributions (CPDs) of the nodes.

In our previous work [13], referring to the model structure of AF_AVDBN, we proposed using articulatory DBN models with constrained asynchrony for isolated words, along with an acoustic speech to realistic mouth animation conversion method for isolated words. In the training of the articulatory DBN models, downsampled YUV spatial frequency features of the interpolated mouth image sequences are extracted as visual features. For reproducing the mouth animation sequence, from the learned visual features, a spatial upsampling and a temporal downsampling have been applied. Both qualitative and quantitative results show that through the articulatory DBN models of isolated words, the accuracy of the constructed mouth shapes is improved compared to the
state of the art HMMI and AF_DBN based methods. In this work, we extend the approach of [13] by 1) considering a continuous speech AF_AVDBN model and synthesizing mouth animations for continuous speech, and 2) using the active appearance model (AAM) based visual features to improve the cleanness of the synthesized mouth images. Moreover, to compare the performance of the proposed approach, we did build an audio-visual state synchronous DBN model [14] (SS_DBN, actually the DBN implementation of MSHMM), as well an audio-visual state asynchronous DBN model with constrained asynchrony (SA_DBN) [14], and synthesize mouth animations from the learned AAM features based on these models. All the above DBN models have been trained using the GMTK [15] toolkit. In our experiments, a database of connected digits of high quality video and audio has been used, and mouth animations have been synthesized for 30 connected digit speech sentences. Objective evaluations showed that the learned visual features using AF_AVDBN track the real parameters much more closely than those from the SA_DBN and SS_DBN models. Subjective evaluation results showed that, by considering the asynchronies between the articulatory features in the AF_AVDBN (as well between the audio and visual states in the SA_DBN), the synchronization between the speech and mouth movements are well obtained. Moreover, since the AF_AVDBN captures the dynamic movements of articulatory features and models the pronunciation process more precisely, the obtained accuracy of the mouth animation is much better than those from the SA_DBN and the SS_DBN models.

The remainder of the paper is organized as follows. In section 2, we summarize the considered audio and visual features. Section 3 describes the state based SS_DBN and SA_DBN models. Section 4 defines the conditional probability distributions of the nodes in the AF_AVDBN model. Section 5 deduces the visual feature learning algorithm and describes the process of synthesizing mouth images from the learned AAM features. The experimental results are analyzed in section 6. and section 7 discusses the conclusions and future work.

2. Audio Visual Speech Features

2.1. Audio Features

From the audio speech, 42 audio features, corresponding to 13 perceptual linear prediction (PLP) features and energy, plus their first and second order differential coefficients, are extracted with a frame length of 50 milliseconds and frame shift of 40 milliseconds.

2.2. Visual Features

2.2.1. Face Tracking and Lip Contour Extraction

The Constrained Bayesian Tangent Shape Model (CSM) of [16] has been used for the detection and tracking of a shape model defined by 83 facial feature points, over a facial image sequence. For each image, a 64x48 mouth region of interest (ROI) is extracted, based on the 12 tracked feature points of the outer lip contour defined by the CSM shape [16]. After training of the AAM, each mouth image is represented by the vector $c$ of dimension 80, which is then used as the input visual feature to the SS_DBN, SA_DBN and the AF_AVDBN models, respectively.

3. State Based Audio Visual DBN Models

Using the audio visual multi-stream DBN models of [14], Figure 1 and Figure 2 illustrate the structures of the SS_DBN model and the SA_DBN model respectively. Each DBN is consisting of a Prologue part (initialization), a Chank part that is repeated every audio-visual frame, and a closure of a sentence with an Epilogue part. Every horizontal row of nodes depicts a separate temporal layer of random variables. At each frame of SS_DBN, the audio and visual observation features (AudioObs and VisualObs) share the same state variable, i.e. they are forced to be synchronous at the state level. While in the SA_DBN model, AudioObs and VisualObs are associated with their states StateA and StateV separately. StateA and StateV can transit asynchronously, but the extent of asynchrony is constrained by the node “CheckSync”.

Figure 1: State synchronous DBN model (SS_DBN)

Figure 2: State asynchronous DBN model (SA_DBN)
4. AF_AVDBN and Definition of Conditional Probability Distributions

Figure 3: The structure of AF_AVDBN

Figure 3 shows the structure of the AF_AVDBN model. It consists of three levels: (i) the upper level with Word and Word Transition, (ii) the articulatory feature (AF) level L, T and G, denoting the states of lips (lip location and lip opening), tongue tip and tongue body, as well as glottis and velum, and (iii) the observation level with audio and visual speech features.

The detailed definitions of the nodes are:

- **Word (W):** word instance.
- **Word Transition (WT):** decides if the word transits in the next frame.
- **LPosition (LP)/TPosition (TP)/GPosition (GP):** position of the AF in the current word.
- **LTransition (LT)/TTransition (TT)/GTransition (GT):** decides if the AF transits. If LT, TT, or GT is 1, then the corresponding AF transits in the next frame.
- **L/T/G:** the AF instance.
- **LTransition (LT)/TTransition (TT)/GTransition (GT):** decides if the AF transits. If LT, TT, or GT is 1, then the corresponding AF transits in the next frame.
- **Audio/Visual Obs (av):** audio or visual observation features, i.e. PLP features and AAM features.

In the training process, the AF transcriptions are obtained by mapping the phonemes to AF sequences. In the later mouth synthesi s process, a single stream AF_ADBN model, which has the similar structure as Figure 3, is used for the AF_AVDBN model, which has the similar structure as Figure 3 but all the arcs connecting with the visual observation vector are removed, is also trained with the audio features.

The probability of emitting the audio visual observation $p(O_{av}^t | W_{av}, j, G_{av})$ is defined similarly as in Equation (2), which means that when $G$ changes slower (or faster) than the mean position of $(L, T)$ for more than $S$ indices, the state of $G$ will change according to the same rule of $T$.

The probability of emitting the audio visual observation features by the combined articulatory feature state $j$, e.g. $j = \{L = protruded & narrow; T = alveolar & narrow & uvular & medium; G = closed & voiceless \}$, is modeled as a multiply of Gaussian mixture models (GMMs).

$$p(G_{av}^t | L_t = q, T_t = r, G_t = s) = p(O_{av}^t | G_{av}) = \prod_{d \in \{PLP, AAM\}} \sum_{k} c_{dk}^j \cdot N(\mu_{dk}^j, \Sigma_{dk}^j)$$

For each feature stream $d$ (PLP features for the audio stream, and AAM features for the visual stream), the parameters $c_{dk}^j$, $\mu_{dk}^j$, and $\Sigma_{dk}^j$ are the weight, mean and covariance matrix, of the Gaussian mixture $k$ of the state $d$, respectively. $M$ is the number of Gaussian mixtures, fixed to 2 in our current experiments. $\alpha_d$ is the weight adjusting the influence of the stream $d$, with the constraint $\alpha_a + \alpha_v = 2$. In the training process of the AF_AVDBN model, $\alpha_a$ and $\alpha_v$ are set to 1, respectively.

In the training process, for each stream $d$, the GMM parameter set $\lambda^d$, of all possible combined articulatory feature states in the AF_AVDBN, or combined audio visual states in the SA_DBN and SS_DBN, are estimated using the Expectation Maximization (EM) algorithm.

For the later mouth synthesis process, a single stream AF_ADBN model, which has the similar structure as Figure 3 but all the arcs connecting with the visual observation vector are removed, is also trained with the audio features.

5. Visual Speech Synthesis Based on AF_AVDBN

5.1. Visual Feature Learning Algorithm

Let $\psi_{t}$ be the set of all hidden variables $(LP_t, TP_t, GP_t, LT_t, TT_t, GT_t, L_t, T_t, G_t, CLT_t, CLTG_t)$ at frame $t$. The probability of an audio visual speech ($\omega^a, \omega^v$) evolving
along a hidden variable path \( \psi = (\psi_1, \psi_2, ..., \psi_T) \) can be concisely defined as

\[
P(O^a, O^v, \psi | k) = \prod_{t=1}^{T} p(O^a_t | L_t, T_t, G_t)p(O^v_t | L_t, T_t, G_t)p(\psi_t | \psi_{t-1})
\]  

(4)

Given an input audio sequence \( O^a \) and the trained model set \( \lambda = (\lambda^a, \lambda^v) \), the Maximum Likelihood (ML) criterion is used to find the optimal visual feature sequence by iteratively maximizing an auxiliary function \( \Omega(\lambda; O^a, O^v, O^v') \) defined as:

\[
\Omega(\lambda; O^a, O^v, O^v') = \sum_{\psi \in \Phi} P(O^a, O^v, \psi | k) \log P(O^a, O^v, \psi | k)
\]

(5)

where \( O^v' \) is the newly estimated visual feature sequence, and \( O^v \) is the estimated visual feature sequence in the last iteration, respectively. The optimal visual feature \( o^v_t \) can be obtained by setting the derivative of \( \Omega(\lambda; O^a, O^v, O^v') \) with respect to \( o^v_t \) equal to zero, i.e.

\[
0 = \sum_{\psi_{k}} P(O^a, O^v, \psi | k) c_{\psi_{k}} (O^v_{\psi_{k}})'^{-1} (o^v_t - p^v_{\psi_{k}}) = 0
\]

(6)

where \( k \) denotes the \( k \)th Gaussian mixture.

From Equation (6), \( o^v_t \) is estimated as

\[
o^v_t = \sum_{\psi_{k}} P(O^a, O^v, \psi | k) c_{\psi_{k}} (O^v_{\psi_{k}})'^{-1} p^v_{\psi_{k}}
\]

(7)

where \( p(O^a, O^v, \psi | k) \) is the probability of the audio visual sequence \( (o^a, o^v) \) passing through \( \psi_t \).

For a given input audio speech \( O^a \), we firstly do speech recognition on the audio only AF_ADBN model to obtain the best AF path \( \{\hat{\psi}_1, \hat{\psi}_2, ..., \hat{\psi}_T\} \), then \( o^v_t \) is initialized as

\[
o^v_t = \frac{M_t}{\sum_{k=1}^{K} c_{\hat{\psi}_{t,k}}^p \mu_{\psi_{t,k}}}
\]

(8)

In each iteration of estimating the visual features, we firstly perform audio visual speech recognition using GMTK on the AF_ADBN model, with the audio feature sequence \( O^a \) and the estimated visual feature sequence of the last iteration \( o^v \) as input. N-Best paths are deduced from the output file with the item –verobs 80. Then in Equation (7), we estimate \( o^v \) by replacing the sum over all possible states of the hidden variables \( \psi_t \), by the sum over their states in the N-Best paths.

5.2. Mouth Image Sequence Reconstruction

The learned visual features \( \{o^v_1, o^v_2, ..., o^v_T\} \), with \( T \) being the length of the input audio sequence at frame rate of 25 frames/s, provide the optimal AAM features \( \{a_1 = o^v_1, a_2 = o^v_2, ..., a_T = o^v_T\} \) of the mouth image sequence to be synthesized. In our experiments, we adopted the AAM API from http://bagpuss.smb.man.ac.uk/~him/software/ for the estimation/extraction of the AAM models of mouth images, as well as for the mouth synthesis, using the learned AAM features.

6. Experiments and Analysis

In our experiments, an audio video database of connected digits has been used. The database consists of 100 sentences, each containing 2 to 5 digits (oh and zero to nine) following the scripts of the audio speech database Aurora 5.0. Among the 100 audio visual speech sentences, 70 sentences were selected as the training data and the other 30 sentences as the testing set.

For each audio frame with frame shift of 40 milliseconds, 42 PLP features have been extracted as described in section 2.1. For the visual features, following the approach described in Section 2.2, 80 AAM features are estimated for each of the 5371 frames of the 100 sentences.

6.1. Audio Visual Speech Recognition

To verify the performance of the AF_AVDBN model, we firstly made audio visual speech recognition experiments, and compare the results with those from the SS_DBN model, as well as the SA_DBN model using different asynchrony constraints.

Table 1. AF_AVDBN Recognition rates

<table>
<thead>
<tr>
<th>Model</th>
<th>AF (1)</th>
<th>AF (2)</th>
<th>AF (3)</th>
<th>AF (4)</th>
<th>AF (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>94.74</td>
<td>94.74</td>
<td>94.74</td>
<td>93.68</td>
<td>93.68</td>
</tr>
</tbody>
</table>

Table 2. SA_DBN and SS_DBN Recognition rates

<table>
<thead>
<tr>
<th>Model</th>
<th>SA (1)</th>
<th>SA (2)</th>
<th>SA (3)</th>
<th>SA (4)</th>
<th>SA (none)</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>98.83</td>
<td>97.95</td>
<td></td>
</tr>
</tbody>
</table>

The obtained recognition rates are given in Table 1 and Table 2 for the AF_AVDBN model, and the SS_DBN as well SA_DBN models, respectively. The numbers between brackets are the maximum allowed asynchrony between the audio and visual states in the SA_DBN, or the asynchrony constraint between the articulatory features in the AF_AVDBN, (none) denotes that there is no constraint on the asynchrony between the audio and visual streams.

One can notice that, the recognition rates using SS_DBN and SA_DBN are higher than the ones using the AF_AVDBN model. This is reasonable because the state and state transition based models are good for small vocabulary word recognition, and they need less training data compared to the complicated AF_AVDBN model. However, the state based models cannot describe the detailed articulatory movements of the speech production process. Nevertheless, the speech recognition results using the AF_AVDBN model are also satisfactory. Moreover, from the above results, the SA_DBN with asynchrony constraints produces higher performance than the state synchronous SS_DBN model, as well the SA_DBN model without asynchrony constraint between the audio and visual states. For the AF_AVDBN model, the asynchrony constraint should be set appropriately to get a high performance. When the allowed asynchrony between the articulatory features is too high (4 or 5), the recognition rate decreases.
6.2. Objective Evaluation of the Learned Visual Features

For each of the 30 testing audio speech sentences, 3 mouth animations are synthesized using the AAM features learned from the SS_DBN model, the SA_DBN(1) model and the AF_AVDBN(1) model, respectively. Figure 4 plots the time trajectories of one of the AAM features, estimated from the original image (real estimated parameter) as well as the learned feature using the 3 considered DBN models. One can notice that the learned visual features using AF_AVDBN(1) and SA_DBN(1) track much more closely the real parameter than those from the SS_DBN model. To quantitatively assess the learned visual features, we estimate the mean relative distance (MRD), over the 30 testing sequences, between the original, \( v_{ktjo} \), and the learned visual features, \( \hat{v}_{ktjo} \):

\[
\text{MRD} = \frac{\sum_{k=1}^{30} \sum_{i=1}^{N_k} (v_{ktjo} - \hat{v}_{ktjo})}{80 \times \sum_{k=1}^{30} N_k}
\]

where \( N_k \) is the frame number of the \( k \)th mouth sequence. The obtained MRD scores are 10.1947 for SS_DBN, 9.9042 for the SA_DBN(1) and 9.5091 for the AF_AVDBN(1). These results show that much more accurate visual features can be learned from the AF_AVDBN(1) model due to its capability of modeling the temporal dynamic movements of the articulatory features.

6.3. Subjective Evaluation of the Mouth Animations

Figure 5 depicts an example of a synthesized mouth image sequence. First of all, one can notice that, the AAM features allow synthesizing mouth images very close to the original ones. Moreover, compared to the results of the SS_DBN model, which is actually a DBN implementation of the HMMI method, the synthesized mouth images using the SA_DBN and the AF_AVDBN models, with constrained asynchrony, produce photo-realistic animations close to the real ones. Finally, the synthesis results using the AF_AVDBN are more accurate, i.e. more like the real mouth images (see the last two frames of “Three”, the last frame of “Five”, and all the frames of “Two”) than the ones of the SA_DBN.

In the previous section, the mean relative distance gives an objective distance between the synthesized mouth parameters and the original mouth parameters. However, it is not certain whether these values reflect the error perceived by a viewer. Therefore we performed subjective tests to evaluate the quality of the synthesized mouth motion. 15 students have been asked to perform an acceptability test on the 30 testing image sequences, to investigate the differences among the DBN models by focusing on the naturalness of the synthesized mouth movement. The indicators are: the naturalness, the accuracy of the synthesized mouth parameters, the synchronization between the mouth shapes and the audio speech, and the clearness. The Mean Opinion Score (MOS) was used as a measure to evaluate the naturalness of the synthesized mouth movement. The subjects assigned scores on a five point scale: 1 (bad), 2 (poor), 3 (fair), 4 (good) and 5 (excellent).

<table>
<thead>
<tr>
<th>Table 3. Subjective Evaluation on Original Animation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scores 1 2 3 4 5 MOS</td>
</tr>
<tr>
<td>synchronization 0 0 0 48 402 4.89</td>
</tr>
<tr>
<td>accuracy 0 0 0 56 394 4.88</td>
</tr>
<tr>
<td>clearness 0 0 0 12 438 4.96</td>
</tr>
<tr>
<td>naturalness 0 0 0 35 415 4.92</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4. Subjective Evaluation on AF_AVDBN(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scores 1 2 3 4 5 MOS</td>
</tr>
<tr>
<td>synchronization 0 0 50 328 72 4.05</td>
</tr>
<tr>
<td>accuracy 0 0 75 282 93 4.04</td>
</tr>
<tr>
<td>clearness 0 0 136 296 18 3.73</td>
</tr>
<tr>
<td>naturalness 0 0 63 317 70 4.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 5. Subjective Evaluation on SA_DBN(1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scores 1 2 3 4 5 MOS</td>
</tr>
<tr>
<td>synchronization 1 9 48 305 87 4.03</td>
</tr>
<tr>
<td>accuracy 1 6 139 255 49 3.77</td>
</tr>
<tr>
<td>clearness 0 0 190 250 10 3.60</td>
</tr>
<tr>
<td>naturalness 0 0 95 303 52 3.90</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 6. Subjective Evaluation on SS_DBN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scores 1 2 3 4 5 MOS</td>
</tr>
<tr>
<td>synchronization 2 90 314 44 0 2.89</td>
</tr>
<tr>
<td>accuracy 2 135 289 24 0 2.74</td>
</tr>
<tr>
<td>clearness 0 60 349 41 0 2.96</td>
</tr>
<tr>
<td>naturalness 0 123 298 29 0 2.79</td>
</tr>
</tbody>
</table>

Table 3, Table 4, Table 5, and Table 6 gives the MOS for the mouth animations from original audiovisual speech, AF_AVDBN, SA_DBN and SS_DBN, respectively. One can notice that, by considering the asynchronies between the articulatory features, or between audio and visual states, the overall MOS values of AF_AVDBN and SA_DBN are much higher than the SS_DBN model, especially the synchronization between the audio speech and the mouth movements. Moreover, the obtained results using the AF_AVDBN give higher MOS performances on each evaluation item compared to the ones using SA_DBN. This is due to the fact that the AF_AVDBN captures the dynamic movements of articulatory features and thus model the pronunciation process more precisely.
7. Conclusions and Future Work

In this paper, we proposed an audio to visual conversion method based on an articulatory DBN model (AF_AVDBN) which allows constrained asynchronies between the articulatory features, such as lips, tongue and glottis/velum. PLP features from the audio speech and active appearance model (AAM) features from mouth images have been adopted to train the AF_AVDBN model for continuous speech. An EM-based optimal visual feature learning algorithm is deduced given the input auditory speech and the trained AF_AVDBN parameters. Finally, photo-realistic mouth images are synthesized from the learned AAM features. Objective evaluations show that the learned visual features using the AF_AVDBN allow synthesizing mouth images which are more close to the original ones than those from the state asynchronous DBN model SA_DBN, and the state synchronous DBN model SS_DBN. Subjective evaluation results show that by considering the asynchronies between the articulatory features in the AF_AVDBN (or between the audio and visual states in SA_DBN), the synchronization between the audio speech and the synthesized mouth animations are well obtained. Moreover, since the AF_AVDBN captures the dynamic movements of articulatory features and model the pronunciation process more precisely, very high quality mouth animations can be obtained, with an accuracy that is much better than those from the SA_DBN and the SS_DBN models.

In our future work, we would expand the experiments for large vocabulary continuous speech, and try to compare the results with data provided by international challenges such as the LIPSYNCH challenge.

8. Acknowledgements

This work is supported within the framework of the LIAMA-CAVSA project, the EU FP7 project ALIZ-E (grant 248116), Shaanxi Provincial Key International Cooperation Project (2011KW-04), and the NPU Foundation for Fundamental Research (NPU-FFR-JC200043).

9. References