A Cantonese Speech-Driven Talking Face Using Translingual Audio-to-Visual Conversion

Lei Xie[1], Helen Meng[1], and Zhi-Qiang Liu[2]

[1] Human-Computer Communications Laboratory
Dept. of Systems Engineering & Engineering Management
The Chinese University of Hong Kong, Hong Kong
{lxie, hmmeng@}@se.cuhk.edu.hk
[2] School of Creative Media
City University of Hong Kong, Hong Kong
zq.liu@cityu.edu.hk

Abstract. This paper proposes a novel approach towards a video-realistic, speech-driven talking face for Cantonese. We present a technique that realizes a talking face for a target language (Cantonese) using only audio-visual facial recordings for a base language (English). Given a Cantonese speech input, we first use a Cantonese speech recognizer to generate a Cantonese syllable transcription. Then we map it to an English phoneme transcription via a translingual mapping scheme that involves symbol mapping and time alignment from Cantonese syllables to English phonemes. With the phoneme transcription, the input speech, and the audio-visual models for English, an EM-based conversion algorithm is adopted to generate mouth animation parameters associated with the input Cantonese audio. We have carried out audio-visual syllable recognition experiments to objectively evaluate the proposed talking face. Results show that the visual speech synthesized by the Cantonese talking face can effectively increase the accuracy of Cantonese syllable recognition under noisy acoustic conditions.

1 Introduction

With the recent advances in multimedia technologies, animated characters, such as talking faces/heads, are playing an increasingly important role in human-computer communication. Talking faces can be driven by input text or input speech[1]. While text-driven talking faces employ both synthesized voices and faces, constituting text-to-audiovisual speech (TTAVS); speech-driven talking faces involve synthesizing visual speech information from real speech. A speech-driven talking face may serve as an aid to the hearing-impaired as the visual speech signal can effectively augment the audio speech signal (eg. by lip-reading) in order to enhance clarity in speech perception. The timing information needed for visual speech synthesis must be synchronized to the input audio speech signal. Such timing information may be obtained by means of a speech recognizer. Hence, speech-driven talking face synthesis is an interesting and feasible research problem [1].

Q. Huo et al.(Eds.): ISCSLP 2006, LNAI 4274, pp. 627–639, 2006.

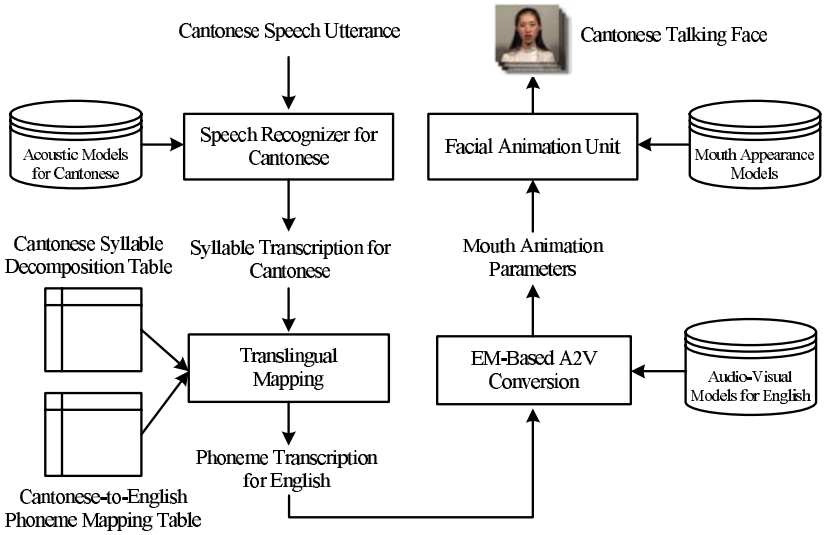

Fig. 1. Block Diagram of the Cantonese Talking Face System

During the last decade, various talking faces have been proposed, pursuing either a natural 3D facial mesh [2] or video-realistic effects [3]. These talking faces are mostly driven by English phonetics (or visemes). Recently we also see talking faces driven by Finnish [4], Italian [5], Chinese Mandarin (Putonghua) and Cantonese [6]. A related problem is how to animate a talking face designed based on phonetics in one language, with input audio speech in another (target) language. For example, Verma *et al.* [7] have proposed a Hindi talking face based on a translingual mapping between Hindi and English phonemes. In this paper, we extend our previous work on an English talking face [8], such that it may be driven by input Cantonese speech. Such translingual audio-visual associations enhance the inter-operability between audio speech analysis and visual speech synthesis.

The rest of the paper is organized as follows. The following section describes the block diagram of our talking face system. In Section 3, the translingual audio-to-visual conversion scheme is presented in detail. Section 4 describes our facial animation unit. In Section 5, experiments are carried out to evaluate our Cantonese talking face. Finally conclusions are drawn in Section 6.

2 System Overview

Fig. 1 shows the block diagram of the proposed Cantonese talking face system. The system is composed of four main phases—a Cantonese speech recognizer, a translingual mapping unit, an expectation maximization (EM)-based audio-to-visual (A2V) converter and a facial animation unit.

The initial audio-visual model is developed based on English phonetics. Input English audio speech is fed into the A2V converter which generates mouth animation parameters. This A2V converter adopts an EM-based conversion algorithm which generates mouth parameters frame by frame under the maximum likelihood (ML) sense, which is frame-synchronized to the audio input. These generated mouth images are "stitched" onto a background facial image sequence using a facial animation unit. Since this work presents a Cantonese speech-driven talking face, we need to extend the existing framework to cover the target language of Cantonese, as described below.

Different from our previous English talking face, in this work we use a Cantonese speech recognizer to generate a Cantonese syllable transcription for the input audio. Subsequently, the translingual mapping unit is in charge of mapping the Cantonese syllable transcription into a reasonable English phoneme transcription where each phonetic unit is associated with estimated timing information. As the initializations, the corresponding visual model means associated with the English phonetic string, together with the input Cantonese audio, is fed into the A2V converter.

3 Translingual Audio-to-Visual Conversion

We have developed a translingual audio-to-visual conversion scheme that is capable of converting speech input in the target language (namely Cantonese) into mouth animation parameters corresponding to the base language (i.e., English) of the existing audio-visual model. This facilitates inter-operability between the audio speech analysis component and the visual speech synthesis component. In this way, we do not need to record a new visual database for visual speech synthesis.

3.1 Audio-Visual Modelling in the Base Language of English

English is the base language of our audio-visual model since we have already proposed a video-realistic talking face [8] that learned audio-visual associations for spoken English from audio-visual facial recordings. These facial recordings involve head-and-shoulder front-view videos of a female speaker uttering 524 TIMIT sentences.[1] Each acoustic feature vector includes 12 MFCCs with log energy and their first and second order derivatives (hence 39 dimensions in total). The mouth region-of-interest (ROI) was first tracked, and encoded using the principal component analysis (PCA). To achieve video-realistic animation, we used PCA to get the visual features that capture mouth appearance in a low dimension (30 PCA coefficients here).

We used multi-stream hidden Markov models (MSHMMs) [9] to model the audio-visual articulation process in terms of context-dependent (CD) phoneme

[1] For details on the AV recordings, please refer to http://www.cityu.edu.hk/rcmt/mouth-synching/jewel.htm

Table 1. Cantonese phonetic decomposition table (partial)

		Initial	Nucleus	Coda	E.g. Character
uk	phn. string	×	*u*	*k*	屋
	duration	0	0.5	0.5	
bing	phn. string	*b*	*i*	*ng*	並
	duration	0.2	0.4	0.4	
baang	phn. string	*b*	*aa*	*ng*	崩
	duration	0	0.5	0.5	
loeng	phn. string	*l*	*eo*	*ng*	涼
	duration	0.3	0.35	0.35	
jyun	phn. string	*j*	*yu*	*n*	員
	duration	0.25	0.375	0.375	

Note: '×' denotes a NULL phoneme.

models (triphones and biphones). We used two-stream, state-synchronous MSHMMs in audio-visual modelling, where two observation streams are incorporated to describe audio and visual modalities respectively. In its general form, the class conditional observation likelihood of the MSHMM is the product of the observation likelihoods of its single-stream components, where stream exponents are used to capture the reliability of each modality.

Given the bimodal observation $\mathbf{o}_t^{av} = [\mathbf{o}_t^a, \mathbf{o}_t^v]$ at frame t, the state emission likelihood of a MSHMM is

$$P(\mathbf{o}_t^{av}|c) = \prod_{s\in\{a,v\}} \left[\sum_{k=1}^{K_{sc}} \omega_{sck}\mathcal{N}_s(\mathbf{o}_t^s; \mu_{sck}, u_{sck})\right]^{\lambda_{sct}}, \quad \sum_s \lambda_{sct} = 1 \qquad (1)$$

where λ_{sct} denotes the stream exponents, which are non-negative, and a function of modality s, the HMM state c, and frame t. The state dependence is to model the local, temporal reliability of each stream. We set $\lambda_{sct} = 0.5$ for all s, c and t supposing audio speech and visual speech have the same contribution. $\mathcal{N}_s(\mathbf{o}_t^s; \mu_{sck}, u_{sck})$ is the Gaussian component for state c, stream s, and mixture component k with mean μ_{sck} and covariance u_{sck}. In total we trained 423 MSHMMs for triphones, biphones and monophones. Each MSHMM has 3 emitting states with 6 continuous density Gaussian mixtures.

3.2 Translingual Mapping

The current work aims to integrate Cantonese audio speech analysis and English visual speech synthesis. This involves a translingual mapping of two levels:

– **Symbols:** Different languages have different phonological units, e.g. syllables are commonly used for Cantonese and phonemes are commonly used for English. This also entails different contextual representations, e.g. initial-finals

for Cantonese and triphones/biphones for English. The different symbolic representations need to be bridged.

– **Timing:** Phonetic units of different languages may have different time durations. The audio frame rates used in the recognizer may be different from the video frame rate of the audio-visual models. Therefore, time alignments must be considered along with the mapping across symbolic representations.

3.2.1 Mapping Across Different Symbolic Representation Systems

Previous work in Translingual speech-driven talking face has involved Hindi and English[7]. These two languages are both Indo-European, and can be accomplished by simple phoneme-to-phoneme mapping. Our approach involves mapping between Chinese and English that their phonological architectures are quite different [10].

The Chinese spoken languages (e.g. Cantonese) do not have explicit word delimiters and a word may contain one or more characters. Each character is pronounced as a *syllable*, and an utterance is heard as a string of momosyllabic sounds with tones. If we ignore the tonal variations, the syllable unit is commonly referred to as a *base syllable*. In general, there are about 600 base syllables in Cantonese. Each base syllable is decomposed into an *initial* and a *final*, and a final can be further subdivided into a *nucleus* and a *coda*. For example, the syllable /nei/( ) is composed of a initial /n/, a nucleus /ei/ and a null coda. In Cantonese, there are about 20 initials and 53 finals. If we categorize these units (initials, nucleus, and codas) with the same (or similar) pronunciations into a phonetic class, there are altogether about 28 "phonetic" classes. Recall that this work needs to map symbolic representation of Cantonese phonetics to that of English phonetics. Based on the above phonetic classifications, our approach involves the following two steps.

– Decompose a base syllable into a sub-syllable string with an initial, a nucleus and a coda, which constitutes a Cantonese "phonetic" string;

– Map the Cantonese "phonetic" string to an English phonetic string via a translingual mapping table.

Table 1 and Table 2 show fragments of the Cantonese phonetic decomposition table and the translingual mapping table respectively. Note that the phoneme durations in Table 1 are obtained from Cantonese syllable samples, and also some Cantonese phonemes are mapped to English phoneme pairs in Table 2. For example, /yu/ is mapped to {/ih/, /uw/}.

In our approach, we used a homegrown Cantonese base syllable recognizer [11] to transcribe input Cantonese speech. In this recognizer, the acoustic models includes three-state HMMs for syllable initials and five-state HMMs for syllable finals. These acoustic models are context-dependent HMMs, namely *initial-final* models, with 16 Gaussian mixtures. They were initially trained with clean, read speech from CUSENT, [2] and then adapted with studio anchor speech recorded

[2] http://dsp.ee.cuhk.edu.hk/speech/cucorpora/

Table 2. Cantonese-to-English phoneme mapping table (partial)

Cantonese Phoneme	E.g.	English Phoneme	E.g.
b	**b**ou(報)	*b*	**b**oy
m	**m**ei(美)	*m*	li**m**it
h	**h**aa(下)	*hh*	**h**air
gw	**gw**ok(國)	*g-w*	**g**row-al**w**ays
i	s**i**(士)	*ih*	h**i**lls
oe	j**oe**ng (洋)	*er*	m**ur**der
yu	j**yu**n (員)	*ih-uw*	h**i**lls-t**w**o

from the news broadcasts of the Hong Kong TVB Jade Channel. (about 40 minutes). The syllable recognition accuracy is 59.3%. Further details on the recognizer can be found in [11].

We first collected the base syllable transcription for a Cantonese utterance, and subsequently aligned the transcription to initial-final symbols via the Viterbi algorithm [9]. The core sub-syllables, e.g., /F_ei/ in /I_g−F_ei+I_gw/,[3] were mapped to English phonemes via Table 1 and Table 2, as illustrated in Fig. 2 (a) and (b). Since we used context-dependent AV models (triphone and biphone MSHMMs) to catch the coarticulation phenomena, we further expanded the English phonemes to triphones or biphones by considering the nearest neighbors, as shown in Fig. 2 (d). The triphones and biphones were selected from the 423 AV models. If a triphone (or biphone) match cannot be found in the model list, a simple phoneme model is chosen.

3.2.2 Time Alignment

Previous research have shown that humans are quite sensitive to the timing relations between audio and visual speech [3]. Therefore, we use the following steps to capture reasonable time relations:

Step 1: If the sub-syllable is an initial (I_*), its duration is directly obtained from the alignment of the speech recognition against the recognized sub-syllable units. If the sub-syllable unit is a final (F_*), the durations of its nucleus and coda are assigned via the durations defined in Table 1. Note that state durations are merged to the model level. For example in Fig. 2 (a), the durations of initials /g/ and /gw/ are directly obtained from the alignment result. The durations of Cantonese "phonemes" /e/ and /i/ are obtained from Table 1. The durations of the three states of /F_ei−I_gw+F_ok/ are merged.

Step 2: Cantonese "phoneme" durations are directly assigned to English phoneme durations. If the Cantonese "phoneme" is mapped to an English phoneme pair, the duration of each English phoneme is a half duration of the Cantonese "phoneme". For example in Fig. 2 (b), the durations of /g/, /eh/, /ih/, /ao/ and /k/ are directly obtained from /g/, /e/, /i/, /o/ and /k/, while the durations of /g/ and /w/ are half durations of /gw/.

[3] Where 'I' denotes Cantonese syllable initial, and 'F' denotes syllable final.

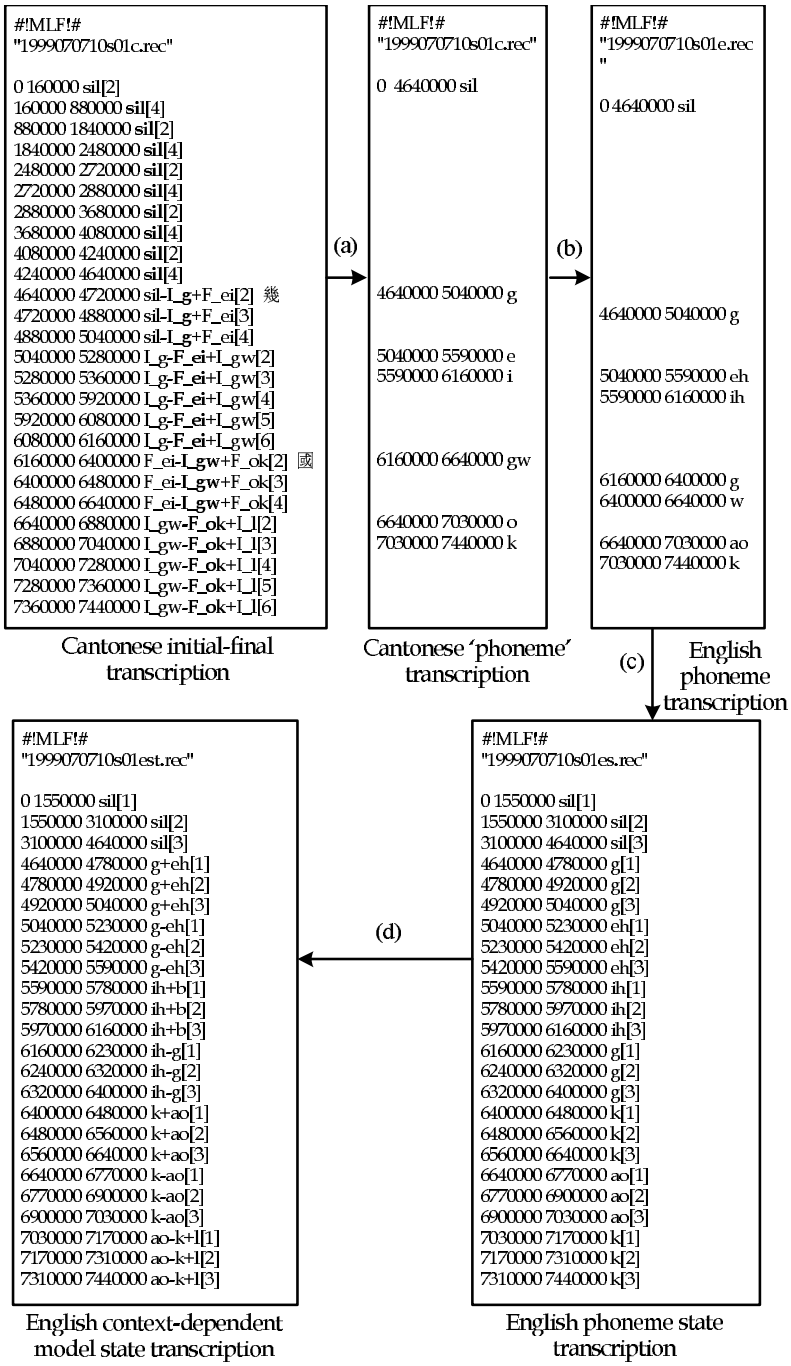

Fig. 2. An example of the translingual mapping process. Label format: start_time end_time phonetic_label[state].

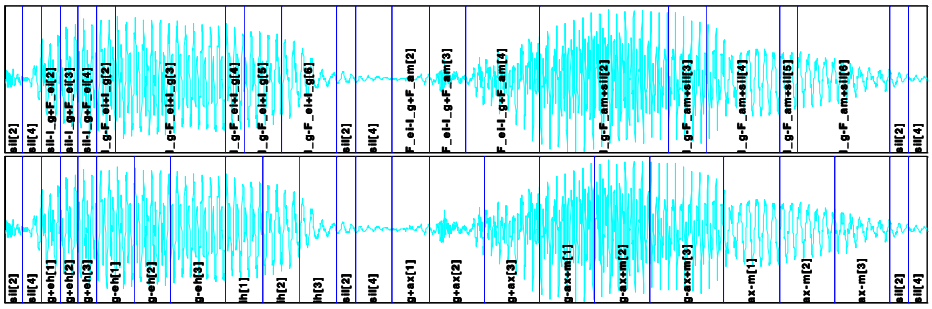

Fig. 3. Time alignment result for a speech fragment. Up: Cantonese initial-final transcription, Bottom: English context-dependent model state transcription.

Step 3: The duration of a state is 1/3 of that of a phoneme (or triphone, biphone). For example in Fig. 2 (c), the duration of state /g[1]/ is 1/3 of that of /g/.

Note that we directly use the the durations for initials generated from the recognizer in Step 1 since they are more accurate for the specific utterance as compared to the statistics from syllable samples. Fig. 3 shows a time alignment result for a Cantonese speech fragment using the above steps.

Finally, the average values of visual Gaussian means associated with each model states

$$\mathbf{o}_t^v = \sum_k w_{\theta_t^v k} \mu_{\theta_t^v k} \tag{2}$$

were used as the initializations of mouth animation parameters, where θ_t^v is the mapped phonetic model state at t. These initialized values were fed into the EM-based A2V converter.

3.3 EM-Based AV Conversion

We used an EM-based audio-to-visual conversion method [8] which directly resulted in mouth parameters (i.e., estimated PCA coefficients) framewise under the ML criterion. The EM-based conversion method has been shown robust to speech degradations, resulting in decent mouth parameters [8].

Given the input audio data $\mathbf{O}^a$ and the trained MSHMMs λ, we seek the missing visual observations (i.e. parameters) $\hat{\mathbf{O}}^v$ by maximizing the likelihood of the visual observations. According to the EM solution of ML, we maximize an auxiliary function:

$$\hat{\mathbf{O}}^v = \arg \max_{\mathbf{O}^{v'} \in \mathcal{O}^v} \mathcal{Q}(\lambda, \lambda; \mathbf{O}^a, \mathbf{O}^v, \mathbf{O}^{v'}), \tag{3}$$

where $\mathbf{O}^v$ and $\mathbf{O}^{v'}$ denote the old and new visual observation sequences in the visual observation space $\mathcal{O}^v$ respectively.

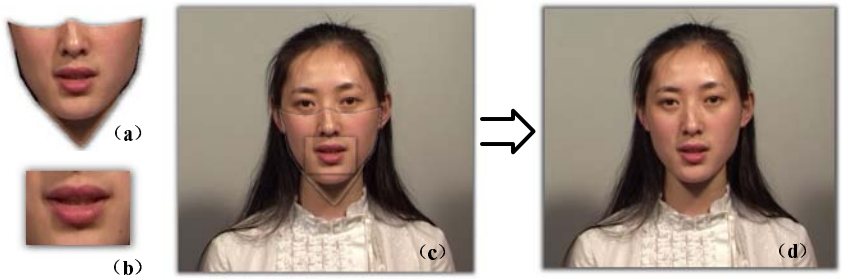

Fig. 4. The three-layer overlay process. (a) a jaw candidate, (b) a synthesized mouth, (c) stitching to face and (d) a resultant frame.

By taking derivative of $\mathcal{Q}(\lambda, \lambda; \mathbf{O}^a, \mathbf{O}^v, \mathbf{O}^{v'})$ respect to $\mathbf{o}_t^{v'}$ to zero, we get [8]

$$\hat{\mathbf{o}}_t^v = \frac{\sum_{q_t} \sum_k \gamma_t(q_t, k) \omega_{q_t v k} u_{q_t v k}^{-1} \mu_{q_t v k}}{\sum_{q_t} \sum_k \gamma_t(q_t, k) \omega_{q_t v k} u_{q_t v k}^{-1}}, \tag{4}$$

where q_t is the possible state of t, and the *occupation* probabilities $\gamma_t(q_t, k)$ can be computed using the forward-backward algorithm described in the E-Step of the EM algorithm. Since the EM algorithm converges to a local minimum, a good parameter initialization is essential for accurate mouth parameters. Therefore, we adopted the visual Gaussian means associated with the mapped English phonetic transcriptions (see Eq. (2)) as the initializations.

4 Video-Realistic Facial Animation

The facial animation unit first smoothes the estimated mouth parameters (i.e., PCA coefficients) by a moving average filter (width=3) to remove possible jitters, and then augments the fine appearance details through a performance refinement process indicated in [12]. Mouth images are generated from the estimated PCA coefficients by the PCA expansion process. Finally, the synthesized mouth frames are overlaid onto a base facial video clip.

We used a three-layer overlaying process (see Fig. 4), where the synthesized mouth, the corresponding jaw, and the face background snippet are sewed up by the Poisson image editing technique [14]. We associated an appropriate jaw from a jaw candidate set to each synthesized mouth according to the mouth opening scale and the waveform energy [12]. To avoid jerky animation induced by stitching coordinates errors, we used a facial feature tracking method [3] with sub-pixel accuracy. Fig. 5 illustrates some snapshots from a synthesized talking face video clip.

5 Evaluations

To evaluate the proposed Cantonese talking face, we carried out objective evaluations using audio-visual speech recognition (AVSR) experiments. This kind

Fig. 5. Some snapshots from a synthesized video

Table 3. Evaluation systems

System	Features & Models	Training & Testing
AO	MFCCs+Δ+Δ^2 (39); CD-HMMs with 16 mixtures	*Training*: Original audio (40 mins); *Testing*: Original, 20db, 10dB (207 secs)
AV-*nontrans*	Audio: MFCCs+Δ+Δ^2 (39); Video: PCA Coefs. (30); CD-MSHMMs with 16 mixtures for audio and 6 mixtures for video; Without translingual mapping	*Training (Audio)*: Original audio (40 mins); *Testing (Audio)*: Original, 20db, 10dB (207 secs)
AV-*trans*	Audio: MFCCs+Δ+Δ^2 (39); Video: PCA Coefs. (30); CD-MSHMMs with 16 mixtures for audio and 6 mixtures for video; With translingual mapping	*Training (Video)*: Estimated PCA Coefs. from original audio (40 mins); *Testing (Video)*: Estimated PCA Coefs. from originial audio (207 secs)

of lipreading test by machine was used to evaluate the quality of the mouth animation (i.e. visual speech) in terms of the improvement in speech recognition accuracy of an AVSR system versus an audio-only ASR system. It provides a way to evaluate the quality of visual speech synthesis by means of machine perception.

5.1 Experiment Setup

We used the hand-transcribed anchor speech (about 40 minutes) from the Cantonese news broadcasts of the Hong Kong TVB Jade channel (described in Section 3.2) as the training data, and another 207 seconds anchor speech were used as the testing set. Speech babble noise (simultaneous speech from multiple speakers collected from cafeteria environment) was added to the testing speech at two signal-noise-ratio (SNR) conditions (20dB and 10dB). As a sanity check, we also developed a talking face without the translingual mapping, where Cantonese

Table 4. Experimental results

	AO	AV-*nontrans*	AV-*trans*
Ori.	59.3	59.4	59.6
20dB	40.6	46.2	50.0
10dB	19.0	28.8	34.3

input speech was directly converted to an English phonetic transcription by an English recognizer. The English recognizer was trained using the audio data from the English audio-visual facial recordings described in Section 3.1. We carried out syllable recognition experiments, and collected syllable accuracy rates for an audio-only ASR system and two AVSR systems. In the AVSR systems, we also adopted the state-synchronous context-dependent MSHMMs described in Section 3.1 as the audio-visual modelling scheme, and the estimated animation parameters (i.e., PCA coefficients) from the original speech were used as the visual features. The stream exponents were selected by minimizing the syllable error rate.

In the experiments, we also used the Cantonese syllable recognizer described in Section 3.2 as the audio-only (AO) baseline system to benchmark the test. The AO system was trained using the same training data. Experiments were performed under mismatched training-testing conditions, i.e., the recognizer was trained using original clean speech, while tested using contaminated speech (10dB and 20dB SNR). Table 3 summarizes the system configurations.

5.2 Experimental Results

From results in Table 4, we can clearly observe that the AO system is heavily affected by additive noise. When the SNR is decreased to 10dB, the syllable accuracy is only 19.0%. The insertion errors contribute a lot to the accuracy decrease. This also shows that training-testing mismatch can drastically affect the performance of a recognizer. Not surprisingly, with the help of the visual speech information provided by the talking faces, both the AV-*nontrans* and the AV-*trans* systems significantly improve the accuracy rates at noisy conditions, with the latter (with the translingual mapping) being superior, yielding a 3.8% and a 5.5% absolute accuracy increase at 20dB and 10dB SNR respectively as compared with the former (without the translingual mapping). These promising results show that the visual speech synthesized by the proposed talking face contains useful lipreading information that can effectively increase the accuracy of machine speech perception under noisy conditions.

6 Conclusions

This paper presents a video-realistic, speech-driven talking face for Cantonese using only audio-visual facial recordings for English. We have developed a translingual audio-to-visual conversion scheme, which is composed of a Cantonese speech

recognizer, a translingual mapping scheme and an EM-based audio-to-visual converter. The translingual mapping involves symbol mapping and also time alignment from Cantonese syllables to English phonemes. With the help of the translingual audio-to-visual conversion scheme, Cantonese speech is converted to mouth animation parameters using audio-visual English phonetic models. The mouth parameters are resembled to mouth images, and stitched onto a background facial image sequence. We have demonstrated that the visual speech synthesized by the proposed Cantonese talking face can effectively improve the syllable recognition accuracy of machine speech perception under noisy acoustic conditions, for example improving the syllable accuracy rate from 19.0% to 34.3% at 10dB SNR.

The promising results in this work have shown that given recorded facial video clips for one language, it is possible to synthesize reasonable facial animation with speech from another language. Since perceptual evaluations by human viewers are more appropriate for visual speech synthesis, we are currently performing subjective evaluations.

Acknowledgements

We would like to thank the Television Broadcasts Limited for providing the Cantonese news audio. We gratefully acknowledge Ms. Ye Zhu who has done a lot in recording the facial videos. We also would like to thank Ms. Pui Yu Hui for providing the Cantonese syllable recognizer and her help on the translingual mapping. This work is partially funded by the CUHK Shun Hing Institute of Advanced Engineering affiliated in the Microsoft-CUHK Joint Lab. for Human-Centric Computing & Interface Technologies.

References

1. Ostermann, J., Weissenfeld, A.: Talking Faces–Technologies and Applications. Proc. 17th ICPR (2004)
2. Pighin, F., Hecker, D., Lischinski, R., Szeliski, D. H.: Synthesizing Realistic Facial Expressions from Photographs. Siggraph (1998) 75–84
3. Cosatto, E., Ostermann, J.: Lifelike Talking Faces for Interactive Services. Proceedings of IEEE. **91**(9) (2003) 1406–1429
4. Olives, J.-L., Sams, M., Kulju, J., Seppaia, O., Karjalainen, M., Altosaar, T., Lemmetty, S., Toyra, K., Vainio M.: Towards a High Quality Finnish Talking Head. IEEE 3rd Workshop on Multimedia Signal Processing (1999) 433–437
5. Pelachaud, C. E., Magno-Caldognetto, Zmarich, C., Cosi, P.: Modelling an Italian Talking Head. Proc. Audio-Visual Speech Processing (2001) 72–77
6. Wang, J.-Q., Wong, K.-H., Heng, P.-A., Meng, H., Wong, T.-T.: A Real-Time Cantonese Text-To-Audiovisual Speech Synthesizer. Proc. ICASSP (2004) 653–656
7. Verma, A., Subramaniam, V., Rajput, N., Neti, C.: Animating Expressive Faces Across Languages. IEEE Trans. on Multimedia. **6**(6) (2003) 791–800
8. Xie, L., Liu, Z.-Q.: An Articulatory Appraoch to Video-Realistic Mouth Animation. Proc. of ICASSP (2006) 593–596

9. Young, S., Evermann, G., Kershaw, D., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book (for HTK Version 3.2). Cambridge University Engineering Department (2002), [online] http://htk.eng.cam.ac.uk/
10. Linguistic Society of Hong Kong. Cantonese Transcription Scheme (1997)
11. Hui, P. Y., Lo, W. K., Meng, H.: Tow Robust Methods for Cantonese Spoken Document Retrieval. Proc. of 2003 ISCA Workshop on Multilingual Spoken Document Retrieval (2003) 7–12
12. Xie, L., Liu, Z.-Q.: A Coupled HMM Approach to Video-Realisic Speech Animation. Pattern Recognition, submitted (2006)
13. Cosatto, E.: Sample-Based Talking-Head Synthesis. Ph.D Thesis of Swiss Federal Institue of Technology (2002)
14. Pérez, P., Gangnet, M., Blake, A.: Poisson Image Editing. Siggraph (2003) 313–318