# A REAL-TIME CANTONESE TEXT-TO-AUDIOVISUAL SPEECH SYNTHESIZER

*Jian-Qing Wang[1], Ka-Ho Wong[2], Pheng-Ann Heng[1], Helen M. Meng[2] and Tien-Tsin Wong[1],*
[1]Virtual Reality, Visualization and Imaging Research Centre,
Department of Computer Science and Engineering
[2]Human-Computer Communications Laboratory
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong SAR, China
[1]{jqwang, pheng, ttwong}@cse.cuhk.edu.hk, [2]{khwong, hmmeng}@se.cuhk.edu.hk

**ABSTRACT**

This paper describes the design and development of a Cantonese TTVS synthesizer, which can generate highly natural synthetic speech that is precisely time-synchronized with a real-time 3D face rendering. Our Cantonese TTVS synthesizer utilizes a homegrown Cantonese syllable-based concatenative text-to-speech system named CU VOCAL. This paper describes the extension of CU VOCAL to output syllable labels and durations that correspond to the output acoustic wave file. The syllables are decomposed and their initials/finals are mapped to the nearest IPA symbols that correspond to static viseme models. We have authored sixteen static viseme models together with two emotion-based face models. In order to achieve 3D face rendering, we have designed and implemented a blending technique that computes the linear combinations of the static face models to effect smooth transitions in between models. We demonstrate that this design and implementation of a TTVS synthesizer can achieve real-time performance in generation.

## 1. INTRODUCTION

This paper describes our first attempt to develop a text-to-speech audiovisual speech synthesis system that can accept Chinese textual input and generate in real time synthetic speech that is lip-synchronized with a talking face in real-time. Such a system offers a multimedia/multimodal presentation of dynamic information, e.g. for news and weather information reporting,[1] for applications in edutainment, for personified dialog systems,[2] or as an aid for the hearing-impaired [1] where the simulated lip movements can help the user decipher the spoken message. The talking face can also convey non-verbal communicative signals, such as emotions [2].

Development of TTVS synthesizers has been ongoing for a long time. The synthesizer in [3] is one of the earliest prototypes. Previous approaches include an image-based synthesizer as in [4] that concatenates *viseme* images. A *viseme* is a facial image treated as a unit in video that corresponds to a phoneme unit in speech. An alternative approach involves parameterized lip shapes, such as the facial animation parameters (FAPs) in the MPEG-4 standard.[3] This has been applied to three-dimensional facial animation. Since the parameterized model has many fewer parameters than the three-dimensional face model, pre-defined functions (like radial basis) are used to compute the coordinates of the entire three-dimensional model based on the parameters in order to achieve three-dimensional rendering. Additionally, there has been previous work in TTVS for languages aside from English, such as the Italian talking head described in [5].

In this work, we have developed a TTVS synthesizer based on a Chinese syllable-based concatenative synthesizer, CU VOCAL, together with the use of a *blending technique* for three-dimensional face animations. Our objective is to (i) use the most natural-sounding Chinese text-to-speech (TTS) synthesizer that is readily available; (ii) map the phoneme units derived from the synthesized speech into time-synchronized visemes for video generation; (iii) compose target viseme models that can be used to produce a high-quality and realistic talking head; (iv) smooth the image transitions in between visemes to produce natural movements; and (v) design and implement a computationally-inexpensive smoothing technique (known as blending) in order to achieve real-time three-dimensional facial animation for TTVS.

## 2. PHONEMES AND VISEMES RELATED TO CANTONESE

CU VOCAL is a syllable-based concatenative text-to-speech (TTS) synthesizer for Cantonese, a major dialect of Chinese predominant in Hong Kong, South China and many overseas Chinese communities [6]. Cantonese is monosyllabic in nature (like Chinese) and the dialect has a rich tonal structure with between six to nine tones. Coarticulatory effects in CU VOCAL are captured in terms of distinctive features. The TTS engine also uses right tonal context for unit selection. Figure 1 illustrates typical input and output for CU VOCAL. Chinese does not have explicit word delimiters and a word may contain one or more characters. Hence the input Chinese character string is tokenized into Chinese words by a greedy algorithm with reference to a lexicon and the word pronunciations are looked up from a dictionary.[4] For example, in Figure 1, the first character in the input text string, "你" (meaning: you), is pronounced as /nei5/ (i.e. the syllable is /nei/ with tone 5. The syllable inventory adopted in CU VOCAL follows the LSHK[5] convention. CU VOCAL generates the synthetic speech output (in .wav format). The TTS engine has also been extended to explicitly generate the syllable sequence with timing information, e.g. the first syllable unit /nei5/ has a duration of 0.39 second, the fourth unit LP indicates a pause (silence) for 0.504 second and the last two syllables are /lam4/ of duration 0.32 second each. The syllable unit can be further subdivided into an optional onset (i.e. the consonant that starts the syllable), a nucleus (i.e. the core vowel/diphthong) and an optional coda (i.e. the consonant that ends the syllable). The Chinese syllable unit is often subdivided into an initial (i.e. the onset) and the final (i.e. the nucleus and coda). For example, the syllable /nei/ has initial /n/ and final /ei/ (or onset /n/ and nucleus /ei/). The syllable /lam/ has initial /l/ and final /am/ (or onset /l/, nucleus /a/ and final /m/).

[1] See Ananova, http://www.ananova.com/video/
[2] See KurzweilAI.net, http://kurzweilAI.net
[3] FAP Specifications, http://www-dsp.com.dist.unige.it/~pok/RESEARCH/MPEG/fapspec.htm
[4] The Chinese University of Hong Kong (CUHK) AOPA Text Processing Resource is made publicly available and contains a word tokenizer together with a Cantonese pronunciation dictionary with over 200K lexical entries.
[5] Linguistic Society of Hong Kong, http://144.214.20.92/lshk/

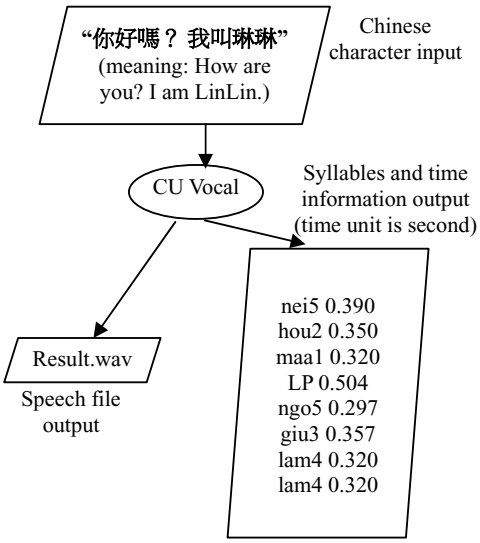


Figure 1. The sample input and output of CU Vocal.

Since much previous work defined visemes in relation to phones (e.g. those from the IPA inventory), our approach involves decomposing a syllable into its onset, nucleus and coda and mapping these to their closest IPA phonetic symbol. We use a total of 28 IPA symbols. Examples of our mapping are illustrated in Table 1.

| LSHK representation | Corresponding IPA symbol |
|---|---|
| /aa/ | /a/ |
| /b/ | /p/ |
| /d/ | /t/ |
| /e/ | /e/ |
| /g/ | /k/ |
| /k/ | /k’/ |
| /o/ | /p’/ |

Table 1. Sample mappings from LSHK to IPA syllable.

Since different phonetic symbols may correspond to the same lip shape, the 28 symbols are mapped to only 16 visemes in total. Examples of mappings from symbols to visemes are provided in Table 2. Figures 2 also illustrate the viseme models of /b/ and /a/ respectively.

Figure 2. Static viseme models for the IPA symbols /b/ (left) and /a/ (right).

| LSHK representation | IPA symbol | Viseme label |
|---|---|---|
| /b/, /p/ | /p/, /p’/ | /b/ |
| /d/, /t/ | /t/, /t’/ | /t/ |
| /aa/ | /a/ | /a/ |
| /g/, /k/ | /k/, /k’/ | /k/ |
| /eo/, /oe/ | /œ/, /œ/ | /œ/ |

Table 2. Examples of viseme definitions.

## 3. THREE-DIMENSIONAL FACIAL ANIMATION

### 3.1. Three-dimensional (3D) Face Model

The basic face model is provided by a computer graphics artist[6] and shows no emotion, no lip movement and no blinking of the eyes. Hence we will also refer to this model as the “neutral” face model (*NeutralFace*).

This model defines such information as position coordinates that determine feature positions, normal coordinates for computing light reflection effects, texture and its coordinates for texture mapping, and other information. Light reflection and texture help create a more realistic appearance of the face model. The position coordinates are linked together to form a network of polygons in order to determine the shapes of facial features (Figure 3). Therefore we can define a 3D model in terms of a sequence of position coordinates (also known as *vertices* in a 3D object), i.e.,

$$F_j = (x_{0j}, y_{0j}, z_{0j}, x_{1j}, y_{1j}, \ldots \ldots, x_{mj}, y_{mj}, z_{mj}) \quad (1)$$

where $F_j$ is a face model $j$, $m$ is the number of position coordinates in a face model, $x_{kj}, y_{kj}, z_{kj}$ are the $k^{th}$ coordinate for x-axis, y-axis and z-axis in face model $j$,

We are able to modify the neutral face model by a 3D character design software tool to form the target viseme models (16 in total) and target emotion models (2 in all for *SmileFace* and *WorryFace*). We have modified the models manually with reference to [7]. Examples are shown in Figures 2 and 4. Target models are static models that correspond to the visemes and emotions. Hence our TTVS system stores 19 models in all (one neutral and 18 target models).

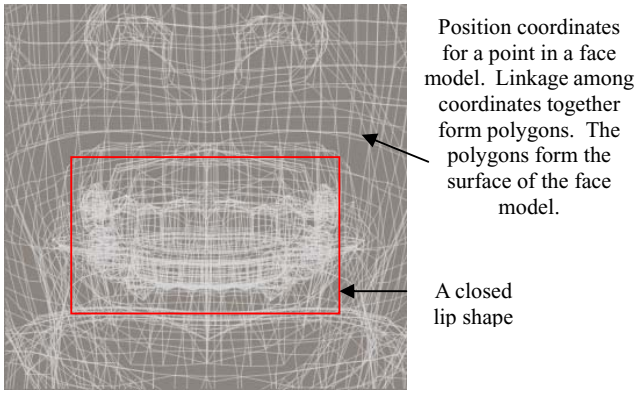


Figure 3. The position coordinates and their linkages.

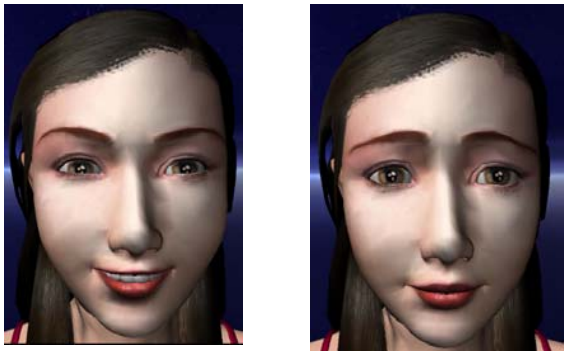

Figure 4. Static emotion models of “smile” (left) and “worry” (right).

[6] See acknowledgements.

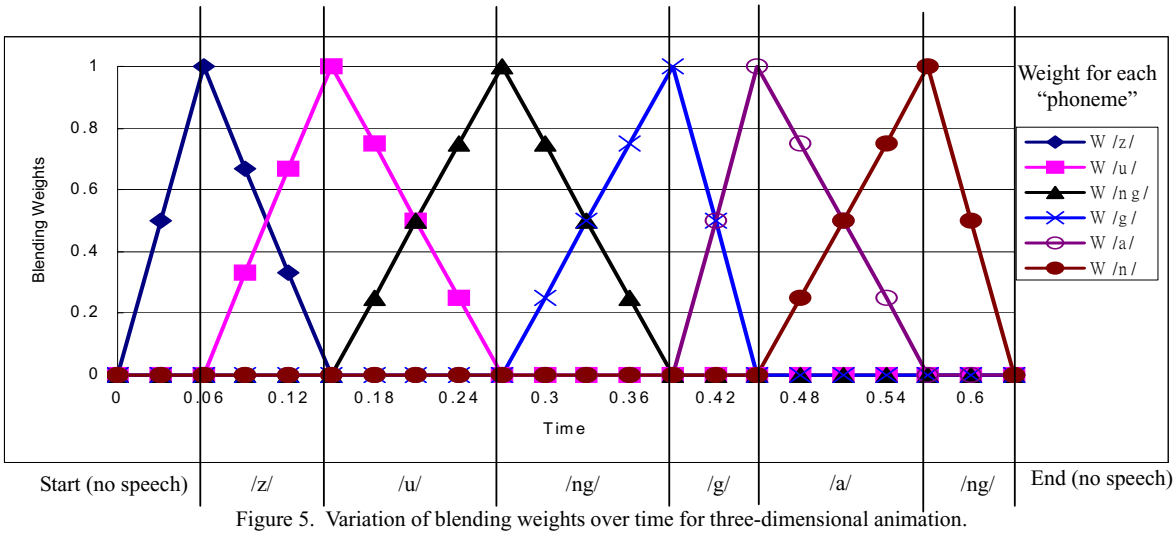


Figure 5. Variation of blending weights over time for three-dimensional animation.

### 3.2. The Blending Process for Animation

We have designed and implemented a simple blending process based on weighted morphing [7] to achieve real-time animation. Animation focuses on the lip shapes and emotions. Each face model is represented by a vector in a $3\times m$-dimensional space ($m$ is defined in Equation 1 known as the *FaceSpace*. All the 3D face models form the *FaceSet*, as indicated in Equation 2.

$$FS = \{F_1, F_2, \ldots \ldots, F_n\} \quad (2)$$

where $FS$ is a *FaceSet*, $F_j$ is a face model $j \in FaceSpace$ and $n$ is the total number of face models (n = 19)

Facial animation can be views as migrating from one point in the *FaceSpace* to another point. A common approach generates a new face model (*NewFace*) can generated by a linear combination of the face models in the *FaceSet* by the use of *blending weights* (see Equation 3). Each blending weight controls the dominance of its corresponding face model (in the *FaceSet*) in the *NewFace* model.

$$NewFace = \Sigma_{i=1}^{n} a_i F_i \quad (3)$$

where $a_i$ is the blending weight for each face model in the face set.

In this work, we incorporate some modifications of the above method. Animation is achieved by the use of *deformation vectors* (DV) derived from the target viseme/emotion models. For example the DV for *smile* is defined in Equation 4.

$$DV_{Smile} = SmileFace - NeutralFace \quad (4)$$

Different DVs can be linearly combined with the neutral face model (*NeutralFace*) to form a new face model (*NewFace*), as illustrated in Equation 5.

$$NewFace = NeutralFace + \Sigma_{i=1}^{n} a_i DV_i \quad (5)$$

where $DV_i$ is the DV for the $i^{th}$ face model in the *FaceSet* and $a_i$ is the blending weight for face model $i$.

### 3.3. Connectivity between Visemes

The transition between two phonemes in synthesized speech corresponds to the transition between two visemes in facial animation. Smooth transition is achieved by controlling the weights in the blending technique. We will elaborate on this point by means of an example. Consider TTVS for the Chinese word 中間 (meaning: center) pronounced as /zung/ /gan/ in LSHK syllables. For a given syllable, we reference the CU VOCAL syllable corpus[7] to get the average duration among the occurring instances. For example, the syllable /zung/ averages 0.33 second in duration. We also reference the corpus to get the average fraction of the syllable's duration that is occupied by its initial and final respectively. For example, the syllable /zung/ has the initial /z/ and final /ung/. The initial /z/ takes up about a quarter of the syllable's duration on average, while the remaining three quarters is taken up by the final /ung/. The final can be further subdivided into the nucleus /u/ and coda /ng/. For the sake of simplicity, we assume the nucleus and coda for the final /ung/ have equal average durations. Hence about 0.5 of the average duration of /zung/ is occupied by the syllable onset /z/, about 0.375 by the syllable nucleus /u/ and the remaining fraction of 0.375 by the syllable coda /ng/. In order to use this information for facial animation, we locate the visemes that correspond to the IPA symbols /z/, /u/ and /ng/ respectively. Since these are static viseme models, we need to determine the blending weights that correspond to these visemes for 3D animation A linear interpolation is used as shown in Figure 5. Each viseme starts with a unity weight at its start instant, and linearly decreases to zero weight at its end point. This defines the variation of the blending weights over time and our system demonstrates that this achieves a realistic and smooth facial animation effect.

The variation of blending weights for emotion face models (see Figure 4) are defined manually by the user by means of a slider rule in our system's interface (see Figure 6). These weights are used in a similar way for 3D face rendering, as compared with the viseme weights. The overall control flow for our TTVS system is depicted in Figure 7.

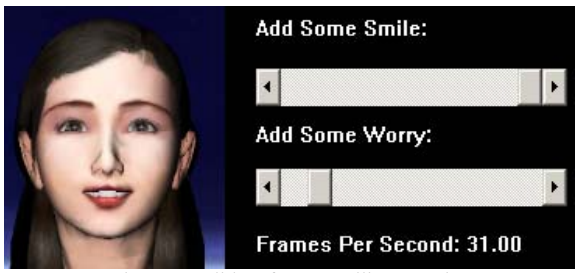


Figure 6. Sliders for controlling emotions.

[7] The CU VOCAL syllable corpus stores syllables extracted from prompt-based recordings. These syllables are used for concatenation in synthesis.

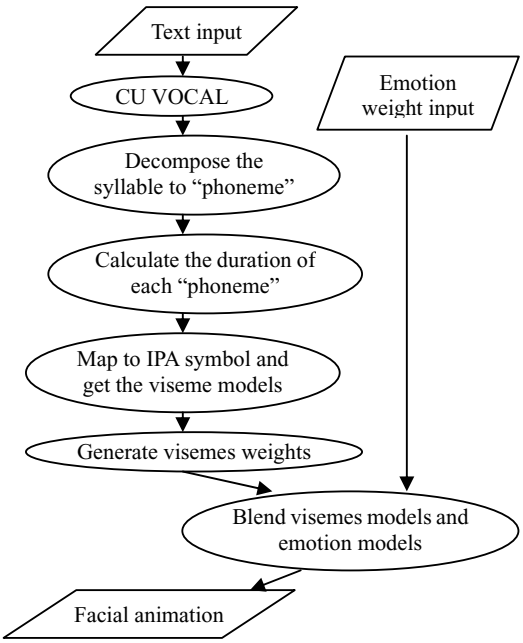


Figure 7. The overall flow about text to audiovisual speech.

## 4. EVALUATION

We ran user perception experiments with 12 randomly generated seven-digit strings. For each digit string we generate either (i) an audio recording of the synthesized speech in a noisy (cafeteria) environment; or (ii) a video file that augments the noisy synthesized audio with a talking face. Our tests involve 16 Cantonese-speaking subjects. Each subject is presented with the 12 audio/video files and asked to write down the digit string that was spoken. Subjects have no prior knowledge of the lengths of the digit strings. Table 3 shows the experimental results in terms of substitution (S), deletion (D) and insertion (I) errors. *Accuracy(%)=1 – total error rates(%).*

| | S | D | I | Accuracy |
|---|---|---|---|---|
| Voice only | 3.1% | 14.7% | 1.6% | 80.6% |
| Voice with face | 4.8% | 3.4% | 2.8% | 89.0% |

Table 3. Results of the user perception experiments.

The digits '5' and '2' are pronounced in Cantonese as /ng3/ and /yi6/ respectively. These are often misrecognized due to their low energies. Furthermore, their visemes look similar – both have a slightly open lip shape. When the synthetic face is included, we observe a slight increase in substitution errors. This is caused by substitutions between '2' and '5'. The significant decrease in deletion errors is predominantly due to better perception of '5' when the viseme is included. The slight increase in substitution errors is due to the insertion of '2' at the end of the digit string – a slight smile in the talking face at the end of the utterance misled the subjects to believe that the viseme for '2' was realized.

## 5. EXTENSIBILITY AND APPLICATIONS

As mentioned previously, we have implemented a TTVS model that involves mapping the LSHK annotation convention for Cantonese syllable initials and finals to the IPA symbols. Many of the visemes thus derived from Cantonese can be used for other Chinese dialects. An example is illustrated for Mandarin Chinese. Table 4 shows examples of mapping Mandarin syllable initials/finals to IPA symbols and their corresponding viseme number. Analogous mapping processes can be applied to other languages to extend the facial animation in our TTVS system to other languages. As a demonstration, we have generated the IPA phonetic sequences from the lyrics of four songs (in Cantonese, Mandarin, Japanese and English respectively) and then derive the viseme sequence (with timing information) for real-time facial animation. Hence in addition to TTVS in Cantonese, the virtual character LinLin can sing in the four different languages. A demonstration is available at www.se.cuhk.edu.hk/TTVS.

| Mandarin pronunciation symbol | IPA symbol | Viseme label |
|---|---|---|
| /b/, /p/, /m/ | /p/, /p'/ | /b/ |
| /d/, /t/ | /t/, /t'/ | /t/ |
| /g/, /k/ | /k/, /k'/ | /k/ |

Table 4. Example mappings between IPA and Mandarin sub-syllable structures.

## 6. SUMMARY

We have developed a Cantonese TTVS synthesizer, which can generate highly natural synthetic speech that is precisely time-synchronized with a real-time 3D face rendering. Our Cantonese TTVS synthesizer utilizes a home-grown Cantonese syllable-based concatenative text-to-speech system named CU VOCAL. This paper describes the extension of CU VOCAL to output syllable labels and durations that correspond to the output acoustic wave file. The syllables are decomposed and their initials/finals mapped to their nearest IPA symbols that correspond to static viseme models. We have also defined two static face models that correspond to emotions. In order to achieve 3D face rendering, we have designed and implemented a blending technique that computes the linear combinations of the static face models to effect smooth transitions in between models. We have demonstrated that this design and implementation of a TTVS synthesizer can achieve real-time performance in generation.

## 7. ACKNOWLEDGEMENTS

We thank Mr. K. Yamato, a famous computer graphics artist from Japan for granting us permission to use his 3D virtual character, LinLin, in our project. This project is partially supported by the Direct Grant from The Chinese University of Hong Kong.

## REFERENCES

1. Karlsson, I.., A. Faulkner and G. Salvi, "SYNFACE – a talking face telephone." In the *Proc. of Eurospeech,* Geneva, Sweden, pages 1297—1300, September 2003.
2. Picard, R., *Affective Computing,* MIT Press, 1997.
3. Parke, F., "Computer Generated Animation of Faces." In the *Proc. of ACM National Conference,* volume 1, pages 451—457, 1972
4. Ezzat, T. and T. Poggio, "Visual Speech Synthesis by Morphing Visemes." In the *International Journal of Computer Vision,* volume 38, no.1, pages 45—57, 2000.
5. Pelachaud, C., E. Magno-Caldognetto, C. Zmarich and P. Cosi, "Modelling an Italian Talking Head." In the *Proc. of Audio-Visual Speech Processing,* Aalborg, Denmark, pages 72—77, September 2001.
6. Lee, T., H. M. Heng, W. Lau, W. K. Lo and P. C. Ching, "Microprosodic Control in Cantonese Text-to-Speech Synthesis", In *Proc. of Euorspeech,* volume 4, pages 1855—1858, September 1999.
7. Parke, F. I.. and K. Waters, *Computer Facial Animation*, A. K. Peters Ltd., 1996.