ALLOPHONIC VARIATIONS IN VISUAL SPEECH SYNTHESIS FOR CORRECTIVE FEEDBACK IN CAPT

Ka-Ho WONG, Wai-Kit LO and Helen MENG

Human-Computer Communications Laboratory
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
Hong Kong SAR, China
{khwong, wklo, hmmeng}@se.cuhk.edu.hk

ABSTRACT

Abstract—***This paper presents a visual speech synthesizer providing midsagittal and front views of the vocal tract to help language learners to correct their mispronunciations. We adopt a set of allophonic rules to determine the visualization of allophonic variations. We also implement coarticulation by decomposing a viseme (visualization of all articulators) into viseme components (visualization of tongue, lips, jaw, and velum separately). Viseme components are morphed independently while the temporally adjacent articulations are considered. Subjective evaluation involving 6 subjects with linguistic background shows that 54% of their responses prefer having allophonic variations incorporated.***

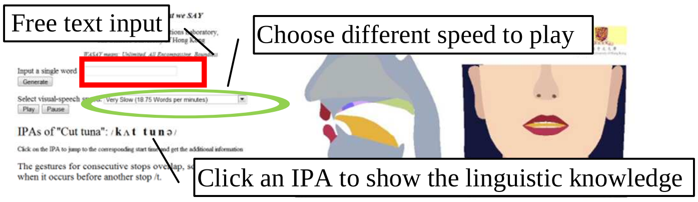
Index Terms— Audiovisual, coarticulation, allophone, language learning, synthesizer

1. INTRODUCTION

Computer-Assisted Pronunciation Training (CAPT) can be used for pronunciation practice and mispronunciation identification in a self-directed learning environment. Conventional CAPT offers feedback as pronunciation scores. There is also a system [1] that presents illustrations of correct pronunciation when a learner chooses an incorrect one. Ville [2] uses iconic indicators to present correctness of pronunciation in several aspects (e.g., duration, stress, reduction) at the same time. Our goal is to provide mispronunciation diagnosis with corrective feedback. Visual corrective feedback is a reliable and efficient way to precisely and concretely show the learners which articulators are involved, and what the correct places and manners of articulation are. Animated visualization is preferred over a static image because the change of articulation over time can also be shown. Engoll [3] showed that providing articulatory instructions relating to mispronunciations can help learners improve their pronunciations. Wik [4] also indicated that the midsagittal view can improve the perception of some Swedish where the tongue movement is invisible in a front view.

WASAY (WAtch what we SAY) (see Figure 1) is a visual speech synthesizer. It uses FreeTTS [5] to convert free text input into a phoneme string, together with the corresponding phone durations and syllable boundaries. WASAY generates synchronized animations of the speech articulators in the midsagittal and the front views. The initial implementation of WASAY [6] uses context-independent visemes, and blending such visemes does not offer a reliable visualization for coarticulation.

We aim to devise a simple and efficient articulatory model of visual speech such that a language teacher can easily explain the articulation for a pronunciation and a learner can easily understand how to pronounce. The model focuses on the illustration of articulations that are relevant to the mispronunciations. We also improve the visual speech output over previous work by realizing salient allophonic variations, such as coarticulation and incorporating linguistic knowledge to provide instructional animations.


Free text input
Choose different speed to play
Click an IPA to show the linguistic knowledge
Generate
Play
Pause

Input a single word

Select visual-speech ...: Very Slow (18.75 Words per minutes)

IPAs of "Cut tuna": /kʌt tunə/

Click on the IPA to jump to the corresponding start time and get the additional information

The gestures for consecutive stops overlap, s... when it occurs before another stop /t.

Figure 1: The screen shot of the WASAY system. The articulatory movements shown on the right side include the lips, tongue, jaw, velum, and voicing at the larynx.

2. ALLOPHONIC VARIATIONS

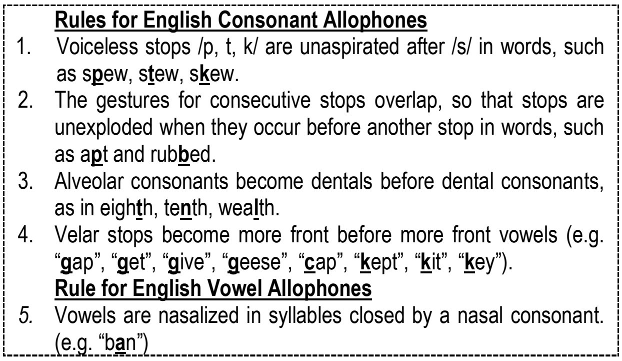
Allophonic variations may be classified as either *free variations* or *complementary variations*. Our current focus is on the allophonic variations that are specific to the language (English in this work) and the involuntary variations due to the coarticulation.

2.1. Allophone rules of English

For a spoken language, there are cases where the pronunciation of a phoneme is changed under different contexts according to allophonic rules. A total of 25 typical rules in English are summarized by Ladefoged [7]. We incorporated five of these rules (as listed in Figure 2), which satisfy the following criteria:

- Variations other than phoneme durations,
- Variations independent of speaking style or rate, and
- Visually salient, either in phonation or in articulation.

Let us take “rubbed” /r ʌ b d/ (relating to Rule 2 in Figure 2) as an example. The phoneme /b/, which is released normally, is pronounced as an unreleased allophone in continuous speech when it is immediately followed by another stop /d/.



Rules for English Consonant Allophones

1. Voiceless stops /p, t, k/ are unaspirated after /s/ in words, such as spew, stew, skew.
2. The gestures for consecutive stops overlap, so that stops are unexploded when they occur before another stop in words, such as apt and rubbed.
3. Alveolar consonants become dentals before dental consonants, as in eighth, tenth, wealth.
4. Velar stops become more front before more front vowels (e.g. “gap”, “get”, “give”, “geese”, “cap”, “kept”, “kit”, “key”).

Rule for English Vowel Allophones

5. Vowels are nasalized in syllables closed by a nasal consonant. (e.g. “ban”)

Figure 2: The allophone rules selected from [7]. Rule 1 is reflected by changes in phonation. Rules 2, 3, 4, and 5 are reflected by changes in both phonation and articulation of the visemes.

WASAY also handles assimilation in the visual aspect. Assimilation refers to “*the change of one sound into another, making it more similar to a neighboring sound*” [7]. For example, the /s/ in “Gas shortage” /g æ s ʃ ɔ r t ə dʒ/ is assimilated to a /ʃ/. To cater for this phenomenon, we incorporated an additional rule:

- /s/ is produced like similar to /ʃ/ when followed by /ʃ/.

2.2 Allophonic variations due to coarticulation

Coarticulation means “*the overlapping of adjacent articulations*” [7]. For example, it is known that the lip configuration of /u/ is rounded, and the lip configuration of /d/ is normally unrounded (e.g. “did”). Under the context of a /u/, lip configuration of the /d/ in “do” will be rounded due to the anticipatory effect (see Figure 3). In another example, “alcohol,” the vocal fricative /h/ does not have stringent requirement on the configuration of the articulators. As a result, the articulator positions of the /h/ first follow that of the preceding /ə/ (schwa). Towards the end of the phoneme /h/, it follows that of the low-back vowel /ɔ/. These variations are supported by the use of viseme components with different configurations relating to the articulators – details will be described in the following section.

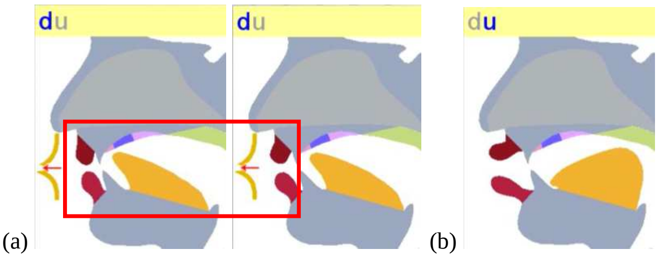


Figure 3: The left most picture in (a) is the articulation of /d/ in “do” without coarticulation. The lips in /d/ are unrounded until the articulation of /u/ starts. With coarticulation (the middle picture), the lips in /d/ start rounding and transit to fully rounded in /u/ as shown in (b).

3. VISEME COMPONENTS

In WASAY, every phoneme is associated with two visemes. A viseme is used to define a *key frame* of audiovisual speech. A viseme is defined by the configurations of tongue, lips, jaw, and velum, which we will refer as *viseme components*. Each viseme component has its own *configuration*. The values of the configurations specify their shape (e.g. rounded for lips), position (e.g. low back for tongue), manner (e.g. retroflexion for tongue), status (e.g. closed for velum), or left unspecified as in the case of /h/. Tables 1, 2, and 3 list all possible values of the viseme components for both the midsagittal view and front view.

We design viseme components for the midsagittal view with reference to [8], and for the front view with reference to a set of audiovisual speech recordings from a Canadian speaker who is a university English instructor. The recording was performed in a studio with two synchronized cameras placed at an angle of 90°. The speaker is presented with a computer screen in front and reads the displayed recording script, which covers all 43 English phonemes.

A point worth noting is that the jaw and lips are combined into one component in the front view. This is because rotation of the jaw and rounding of the lips can be animated independently in the midsagittal view, but lowering/lifting of the jaw in the front view also changes the shape of the lips (Figure 4).

4. PHONATION AND TIMING

We also defined two phonation attributes, namely vocality and airflow (see Table 4). Vocality indicates whether a phoneme is voiced and it is depicted as a vibrating cord in the visual speech animation for voiced phonemes. For unvoiced phonemes, a stationary vocal cord is shown. For airflow, we make use of pictorial indicators to convey the message to the learners. For example, a nasal consonant /n/ is voiced and hence the vocal cord will be vibrating. More importantly, we depict the intended path of airflow for the phoneme by showing a clear nasal cavity with indicator arrows (see Figure 5).

Viseme Component: Velum		
Possible Values	**Midsagittal view**	**Front view**
CLOSED	●	
OPEN	●	

Table 1: This table shows the possible values for the velum attributes, which are applicable to the midsagittal view only.

Viseme Component: Tongue		
Possible Values	**Midsagittal view**	**Front view**
NEUTRAL	●	●
VELAR	●	●
VELAR RELEASE	●	
HIGH-BACK	●	
MID-HIGH BACK	●	
LOW-BACK	●	
MIDDLE-CENTRAL	●	
LOW-FRONT	●	
MIDDLE-FRONT	●	
ALVEOLAR RELEASE	●	
POST-ALVEOLAR RELEASE	●	
HIGH-FRONT	●	
MID-HIGH FRONT	●	●
ALVEOLAR	●	●
RETROFLEXION	●	
INTER-DENTAL	●	●
UNSPECIFIED	●	●

Table 2: This table shows the possible values for the tongue attributes. A "●" indicates a value for the corresponding articulator attribute. For the front view, a single "●" spanning across multiple rows for the midsagittal view means that the tongue has the same value (i.e. same visual position) for those values in the midsagittal view.

Viseme Component: Jaw and lips			
	Midsagittal view		**Front view**
Possible Values	**Jaw**	**Lips**	**Jaw and Lips**
NEUTRAL	●	●	●
SLIGHTLY OPEN	●		●
OPEN	●		●
BILABIAL	●	●	●
DENTAL	●	●	●
LABIO-DENTAL	●	●	●
SLIGHTLY ROUND		●	●
ROUND		●	●
MORE ROUND		●	●
DENTAL AND SLIGHTLY ROUND			●
UNSPECIFIED	●	●	●

Table 3: This table shows the possible values for the jaw and lip attributes. A "●" indicates a value for the corresponding articulator attribute. For the front view, the attributes for the jaw and the lips are combined.

Phonation Attributes		**Normalized Timing**
Vocality	**Airflow**	
VOICED, UNVOICED	NOTHING, NASAL, EXPLODED, ASPIRATED, EXPLODED AND THEN ASPIRATED	0 to 1

Table 4: This table lists all possible values for the attributes related to phonation and timing. For different values for the phonation attributes, we show different pictorial indicators to the users. The normalized timing of "0" and "1" refer to the start and end times of a phoneme respectively [6].

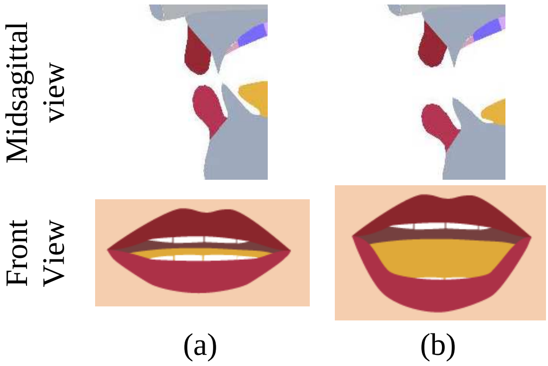


Figure 4: When the jaw opens from (a) to (b), the lower lip shape is unchanged and only a rotation occurs in the midsagittal view. In the front view, the lower lip shape is deformed from a flat u-shape to a deeper u-shape.

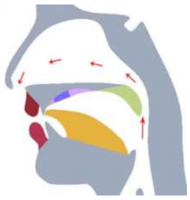


Figure 5: A clear nasal cavity together with some red arrows is used to indicate a nasal or nasalized phoneme.

5. THE VISUAL SPEECH SYNTHESIS PROCESS

The viseme components, phonation and timing define the temporal position and visual appearance of the articulators in the key frames of the visual speech. The basic idea of the visual speech generation process is based on linear morphing. As mentioned before, every phoneme is associated with two visemes, and every viseme defines a *key frame*. Between key frames, intermediate frames are morphed from the preceding and succeeding key frames. The details of synthesis process are listed as follows:

1. For each phoneme, the corresponding visemes are identified. For example, the $V_{t,1}$ and $V_{t,2}$ for /t/ in Figure 6.
2. Any allophone rule triggered will be applied. For example, the first stop plosive /t/ is immediately followed by another /t/. According to Rule 2 in Figure 2, the first /t/ will become unexploded and unaspirated.
3. All triggered allophone rules during the synthesis process will be displayed as a message for reference by the learners and teachers.
4. We apply linear morphing [6] to all viseme components. In case the viseme component has an unspecified value, e.g., the articulation of the lips for both /t/s in Figure 6 are generated by linearly morphing from "NEUTRAL" in /ʌ/ (the preceding phoneme) to "MORE ROUND" in /u/ (the succeeding phoneme).
5. For each intermediate frame, viseme components are combined. We first calculate the rotation angle for the jaw. The lower lip and tongue are also rotated with the same angle. The other viseme components are then positioned with reference to the reference lines as shown in Figure 7. The three reference lines are as follows:

a. Tongue: the line between the tongue and jaw
b. Upper lip: the line between the upper lip and upper jaw
c. Lower lip: the line between the lower lip and jaw

The velum is always drawn at a fixed position with respect to the head.

WASAY also takes physiological and physical constraints into consideration. Every frame is checked after morphing to ensure that they are physiologically reasonable. For example, when the jaw is "OPEN," the lip cannot be "MORE ROUND" (e.g., /aʊ/). As another example, when the lip value is "LABIO-DENTAL," the jaw value must be "LABIO-DENTAL" (e.g., /f/) too. We also checked every frame to ensure that they are physically reasonable. For instance, the tongue cannot penetrate through the hard palate.

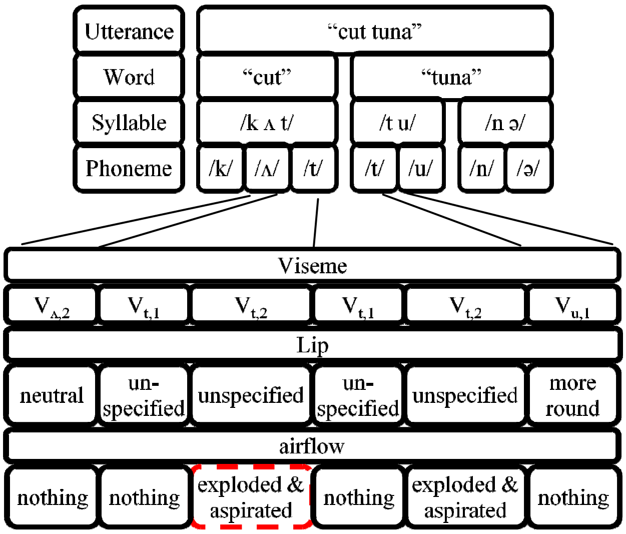


Figure 6: Breakdown of the sample utterance "cut tuna" and the structure of the medial phonemes "c**ut tu**na." The /t/ in "cut" is an allophonic variation which is unreleased. It is because the /t/ (as indicated in the box with a dashed line) is followed by another stop /t/ (and Rule 2 applies).

6. EVALUATION

We invited six subjects with linguistic background (Group A) and eight English learners (Group B) to participate in the subjective evaluation. Each of them is presented with eight pairs of videos. Each pair includes a video incorporating allophonic variations and a control video without. The order of playback of the pair of videos is randomized. Subjects are requested to indicate which of the videos offer a better visualization, or if they are equally good.

Based on the responses from Group A, 54% (26/48) preferred having the allophonic variations incorporated, 27% (13/48) showed no preference and the remaining 19% (9/48) preferred no incorporation of allophonic variations. Most of the subjects in Group A (83%, 5/6) prefer the incorporation of the anticipatory effect – e.g. lip rounding of /d/ in "do". The following three allophonic variations have the second highest preference (67%, 4/6) – the unaspirated plosive (e.g. /k/ after /s/ as in "school"), fronting of velar stops (67%, 4/6, e.g. /g/ in "gill" vesus "gall") and assimilation (67%, 4/6, e.g. /s/ in "gas shortage"). The three allophonic variations with the lowest preferences (all at 33%, 2/6) are nasalized vowels (e.g. /æ/ in "ban"), dentalized consonants (e.g. /n/ in "tenth") and coarticulation of /h/ (e.g. "alcohol").

Based on the responses from Group B, only 20% (13/64) preferred having the allophonic variations incorporated, 52% (33/64) showed no preference and the remaining 28% (18/64) preferred no incorporation of allophonic variations. This result is expected, as the subjects may not be aware of effects of allophonic variations.

7. CONCLUSIONS

We have developed a visual-speech synthesizer offering articulatory visualization with synchronized midsagittal and front views for corrective feedback in CAPT. We have implemented allophonic variations in the visualization by incorporating allophonic rules for English through the use of viseme components with different configurations relating to the articulators. Evaluation shows that 54% of the time a subject with linguistic background indicates a preference to having incorporated allophonic variations in the visualization. This shows that our approach for incorporation of allophonic variations offers a better visualization of articulation for corrective feedback in CAPT. Evaluation also indicates that regular learners need more training to raise awareness of allophonic variations.

8. ACKNOWLEDGMENT

This work has been partially supported by the National Natural Science Foundation of China (60928005).

9. REFERENCES

[1] Y. Tsubota, T. Kawahara and M. Dantsuji, "Recognition and verification of English by Japanese students for computer-assisted language learning system", in *the Proceeding of ICSLP*, pages 749-752, 2002.

[2] P. Wik and A. Hjalmarsson, "Embodied conversational agents in computer assisted language learning", *Speech Communication*, Vol - 51, October 2009.

[3] O. Engwall, "Can audio-visual instructions help learners improve their articulation? - An ultrasound study of short term changes", in *the Proceedings of Interspeech*, 2008.

[4] P. Wik and O. Engwall, "Looking at tongues - Can it help in speech perception?", in *the Proceedings of Fonetik*, 2008.

[5] FreeTTS 1.2, http://freetts.sourceforge.net/docs/index.html

[6] K. H. Wong, W. K. Leung, W. K. Lo and H. Meng, "Development of an articulatory visual-speech synthesizer to support language learning", in *the Proceedings of ICSLP*, Taiwan, 2010.

[7] P. Ladefoged, *A course in phonetics*, 2006.

[8] D. L. F. Nilsen and A. P. Nilsen, *Pronunciation contrasts in English*, 1973.