AUDIOVISUAL SYNTHESIS OF EXAGGERATED SPEECH FOR CORRECTIVE FEEDBACK IN COMPUTER-ASSISTED PRONUNCIATION TRAINING

Junhong Zhao[1,2], Hua Yuan[3], Wai-Kim Leung[4], Helen Meng[4], Jia Liu[3] and Shanhong Xia[1]

[1] State Key Laboratory on Transducing Technology, IECAS, Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100190, China
[3] TNList, Department of Electronic Engineering, Tsinghua University, China
[4] Department of Systems Engineering and Engineering Management, CUHK, Hong Kong SAR, China
junhong.iecas@gmail.com

ABSTRACT

In second language learning, unawareness of the differences between correct and incorrect pronunciations is one of the largest obstacles for mispronunciation correction. In order to make the feedback more discriminatively perceptible, this paper presents a novel method for corrective feedback generation, namely, exaggerated feedback, for language learning. To produce exaggeration effect, the neutral audio and visual speech are both exaggerated and then re-synthesized synchronously based on the audiovisual synthesis technology. The audio speech exaggeration is realized by adjusting the acoustic features related to duration, pitch and energy of the speech according to different phonemes conditions. The visual speech exaggeration is realized by increasing the articulatory movement range and slowing down the movement around the key action. The results show that our methods can effectively generate bimodal exaggeration effect for feedback provision and make them more discriminative to be perceived.

Index Terms— Computer-assisted pronunciation training, exaggerated feedback, visual-speech synthesis

1. INTRODUCTION

In second language acquisition, a Computer-Assisted Pronunciation Training (CAPT) system is often used to diagnose mispronunciations and offer corrective feedback for pronunciation training. As an important aspect of pronunciation training, perceptual training aims to to develop the ability of discriminating different sounds in a language. Some studies [1][2] suggest that discriminative perceptual training can improve the production, and the availability of good corrective-feedback is beneficial for reducing pronunciation errors.

The traditional feedback way of CAPT is to show the pairwise differences between the target speech and the mispronunciation in various aspects, like the speech wave, speech formant, articulation [3][4], articulatory animation [5][6], etc. But sometimes the learners still cannot recognize their mistakes from such straightforward feedback methods. In this case, making the feedback more discriminative to be perceived will be helpful to solve the problem. In the CASTLE system [7][8], to assist second language learners to perceive stress patterns, a perceptual assistance module is used to enlarge the differences between stressed and unstressed syllables from a teacher's speech.

The work presented here focus on the phonetic rather than prosodic feedback. We propose an exaggerated-feedback mode to improve the perceptual effect of the reference phonemes in both audio and visual modalities. We exaggerate the audio speech to make it discriminatively perceptible. Simultaneously, we exaggerate the visual speech by increasing the range of articulatory movements and slowing down the articulatory action. Using this bimodal exaggeration method, we can provide more perceptible and discriminative feedback for learners to correct their mispronunciations.

2. THE EXAGGERATION METHOD

Our exaggerated feedback is implemented on the previous visual-speech synthesizer reported in [9][10], which visualized pronunciation movement from midsagittal and front views of the vocal tract. It focuses on providing corrective feedback and can offer a reliable visualization for coarticulation. As the flowchart in Fig. 1 illustrates, to generate exaggerated visual-speech animation, we use the audio and visual exaggeration modules to adjust audio and video synthesis respectively. The audio produced by text-to-speech (TTS) synthesizer is first exaggerated. The basic elements of the exaggeration are phonemes in the synthesized sentences.

This work was partly conducted when the first author was a summer intern in the Human-Computer Communications Laboratory in The Chinese University of Hong Kong. It was supported by a grant from the NSFC-RGC Joint Scheme (project number N_CUHK414/09) and also a grant from the Innovation and Technology Fund (project number ITF/163/12) from the Hong Kong SAR Government. The project was also supported by the NSFC under Grant No. 61273268, No. 61005019 and No. 90920302.

Then the corresponding visual speech is exaggerated based on viseme models and the modified synchronization information. The adjusted audio and video streams are finally synchronized by visual-speech synthesizer.

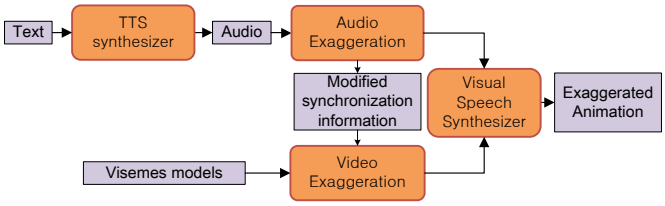

Fig. 1. The flowchart of exaggerated audiovisual synthesis.

2.1. Audio exaggeration with STRAIGHT

Many studies of emphatic speech analysis and synthesis have reported that emphatic speech often has higher pitch with dramatic variation, amplified energy and longer duration. So to realize phoneme exaggeration of the target speech, we enhance the following five acoustic features: (i) pitch level, (ii) pitch variation, (iv) energy, (v) pause length, and duration. Related analysis and re-synthesis are both based on the STRAIGHT algorithm [11].

2.1.1. Pitch level & pitch movement enhancements

Both pitch level and pitch movements contribute to speech emphasis. Let $F_i(n)$ be the temporal sequence of F0 for the i_{th} phoneme of the speech, s_i and e_i be the corresponding start and end indexes, α_1 be the enhancement ratio of pitch level. The enhanced F0 that is denoted by $F_i'(n)$ is calculated by (2). We use the Hamming window denoted by (1) to smooth the boundary transition. Let α_2 and α_3 be the modulation ratio for pitch minimum and pitch maximum. To enlarge the pitch movement, we should reduce the pitch minimum and increase the maximum. So here we set $\alpha_2 < 1$ while $\alpha_3 > 1$. Based on the enhanced $F_i'(n)$ and its weaken minimum and enlarged maximum, the final pitch contours are generated according to (4). The adjustment of pitch-related acoustic features are only for vowels and voiced consonants since voiceless consonants has no vocal cord vibration.

$$H(n) = 0.538 - 0.462\cos(\frac{2\pi(n-s_i)}{e_i - s_i}), n \in [s_i, e_i] \quad (1)$$

$$F_i'(n) = \alpha_1 F_i(n) H(n) \quad (2)$$

$$F''_{i,min} = \alpha_2 F'_{i,min}, \quad F''_{i,max} = \alpha_3 F'_{i,max} \quad (3)$$

$$F_i''(n) = F''_{i,min} + \frac{F''_{i,max} - F''_{i,min}}{F'_{i,max} - F'_{i,min}}(F_i'(n) - F'_{i,min}) \quad (4)$$

2.1.2. Energy enhancement

To amplify the energy, we multiply a constant enhancement ratio for all frequency components of spectrum. The Hamming window is also adopted to ensure gradual boundary transition of phonemes. Let k be the total number of frequency components, $M_i(mn)$ be the temporal sequence of the m_{th} frequency component. The enhanced temporal sequence $M_i'(mn)$ is obtained as:

$$M_i'(mn) = \beta M_i(mn) H(n), \{n \in [s_i, e_i], m \in [1, k]\} \quad (5)$$

2.1.3. Pause addition & duration elongation

To indicate to learners that the next phoneme is the one which they have mispronounced and is about to be exaggerated in reference speech, a pause is added before this phoneme. The duration of the incremental pause is chosen as 0.2 second. Then the duration of the phoneme is lengthened by multiplying a factor γ. Let p be the length of the pause sequence, the start index s_i', duration D_i' and end index e_i' of the i_{th} phoneme are adjusted as follows:

$$s_i' = s_i + p, \quad D_i' = \gamma D_i, \quad e_i' = s_i' + D_i'. \quad (6)$$

2.1.4. Synthesize exaggerated phoneme with STRAIGHT and overall utterance generation

After obtaining the modified pitch contour $F_i''(n)$, the enhanced spectrum $M_i'(mn)$ and the lengthened duration D_i', exaggerated speech synthesis of the i_{th} phoneme is implemented with the STRAIGHT algorithm, which is denoted in term of $f(\cdot)$ in (7). In this process, the corresponding values of extended time-axis are obtained by interpolation.

$$S_i'(n) = f(F_i''(n), M_i'(mn), D_i'), \{n \in [s_i', e_i'], m \in [1, k]\} \quad (7)$$

Then the final speech utterance $S'(n)$ is generated by concatenating the re-synthesized exaggerated phoneme speech with all the preceding and following speech segments.

$$S'(n) = \{S_1(n), \cdots, S_i'(n), \cdots, S_N(n)\} \quad (8)$$

2.1.5. Rules for controlling the degree of exaggeration

To produce the expressive effect without destroying the correct articulatory characteristic by over-exaggeration, the degree of exaggeration should be adjusted according to different phonemes in various contexts.

For vowels, we adjust the exaggeration degree according to the following cases:

- *Stressed & unstressed*: For stressed vowels, the original expressions are already strong. Further exaggeration will drown out surrounding speeches, which is undesirable. So they are enhanced slightly when compared with unstressed vowels.

- *Long & short*: Here "long" category including long monophthongs and all diphthongs; "short" category including all short vowels. The short vowels should be pronounced in a short time with concentrated strength. Hence, to exaggerate them, the pitch and energy are enhanced more largely while the lengthening extent is smaller than long vowels.

Generally speaking, to yield good exaggeration effects, the degree of exaggeration for consonants should be higher than vowels because of their weaker expression. For consonants, we adjust exaggeration degree according to the following cases.

- *Adjacency to stressed vowels (SA) & ustressed vowels(USA)*: The consonants will sound weaker when they are adjacent to the stressed vowels because of the auditory masking effect. So in order to make the generated effect as expressive as in the USA case, the exaggeration degree of consonants in SA case is higher.

- *Voiced & voiceless*: Compared to the voiced consonants, the voiceless ones have lower energy. So they are enhanced to a higher degree.

- *Plosive & non-plosive*: The articulation of the plosives require the air in the vocal tract to be released in a short time. To keep this burst characteristic, the lengthening extent of plosives are more smaller than the non-plosives.

2.2. Realization of visual exaggeration

In our visual-speech synthesis system, each phoneme is assigned to two visemes as the key frames for animation generation [9][10]. Each viseme can be assumed as the representation of a key articulatory action. One viseme constitutes of many viseme components, including velum, tongue, jaw and lips. The configuration of each viseme component revealed its specific status of shape, position, manners etc. All these viseme components will be processed independently and then they are combined together to become one intermediate frame. The animations are synthesized by applying linear-weight morphing technique for the blending of these successive visemes.

We generate the exaggerated visual speech of articulatory animation using the following methods: (i) Increase the movement range of articulatory organs. We select some of the viseme components and exaggerate their corresponding configurations. (ii) Slow down the articulatory movement around the key action to be emphasized. In visual speech, a segment with relatively slower articulatory movement can be perceived more easily for learners. It is implemented by slowing down the change of blending weight around this viseme when morphing.

Table 1. Phonemes and viseme components needed to be adjusted based on the manners of articulation when exaggeration.

Manner of articulation	Viseme components	Phonemes
Open	Jaw from midsagittal view Jaw and lips from front view	aa ae ah ay eh er ey ax axr
Round	Lip from midsagittal view Jaw and lips from front view	ow oy w uw uh ao aw jh sh zh
Bilabial	Lip from midsagittal view Jaw and lips from front view	b p m

2.2.1. Increase articulators

Compared with the articulators like the tongue and velum, the jaw and lips often have wider range of movements. Here we choose to increase the movement range of jaw and lips. In practice, we should use different adjustment methods according to different articulatory manners (open, round, etc.) for exaggeration. The final exaggeration effect depends on the manner requirement of original articulation. For example, if the original articulation requires the lips be rounded (e.g. /ow/), in the exaggerated animation, the lips should be rounder accordingly. The articulatory manners we choose to exaggerate are listed in Table 1, with their corresponding viseme components that should be adjusted and the specific set of phonemes that will trigger these kinds of exaggerations.

In our system, the movement range for each manner of articulation is divided into several levels. To keep the original relative difference of movement range across phonemes, we uniformly raise the movement range by same levels for the exaggeration of all phonemes (two levels in our experiment) based on their original states. For example, as Fig. 2 illustrates, the opening range of the jaw is divided into three levels before exaggeration as neutral, slightly open, open. For exaggeration, we create two extra higher levels: widely open and more widely open, by empirically adjusting the related viseme components. For the phonemes with "slightly open" manner in neutral speech production, we use the "widely open" manner instead when generate the exaggerated visual speech.

2.2.2. Non-linear morphing for key action emphasis

After obtaining the phoneme string and the corresponding duration information, the intermediate frames are generated by morphing. The morphing results between the k_{th} viseme V_k and the $(k+1)_{th}$ viseme V_{k+1} are generated according to (9), where $w(t)$ is the blending weight at time t. To emphasize one key action, we slow down the rate of change in blending weights around the corresponding viseme in focus. To be specific, let the two visemes indexed by n and $(n+1)$ belong to one phoneme and we will emphasize its first viseme V_n by non-linear morphing. After obtaining its lengthened duration produced by the speech exaggeration as mentioned above, we implement the non-linear morphing using the cosine blending

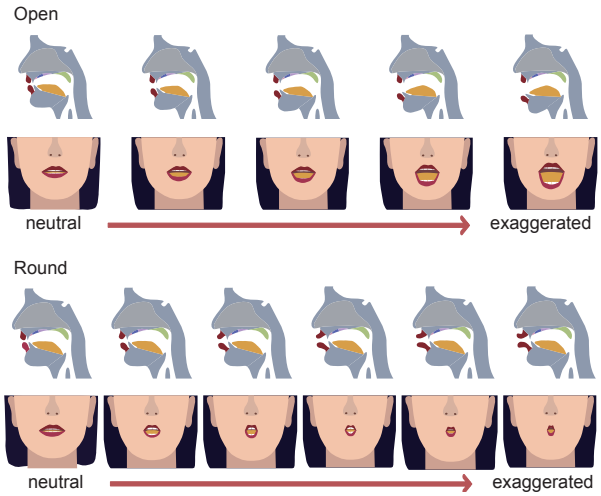

Fig. 2. The movement range levels for "open" and "round" articulatory manners, as illustrated by the viseme components of jaw and lips from midsagittal view, and mouth shape from the front view.

weight function depicted in (10). Referring the power of the cosine function denoted by ω, the smaller the value, the flatter is the curve of the blending weight around the V_n viseme. The comparison of the changes in blending weight before and after emphasis is illustrated in Fig. 3.

The burst characteristic of the plosive requires the corresponding articulatory motion to be completed in a very short time. Thus, it is not reasonable to slow down the movement between its two visemes. But we can slow down the movements at preparation and ending phrases. So when exaggerating one plosive, only the sections before the first viseme and after the second viseme morph in a non-linear way, and the morphing of the section between the two visemes remains linear.

$$V_{k,k+1}(t) = V_k w(t) + V_{k+1}(1 - w(t)) \tag{9}$$

$$w(t) = \begin{cases} 1 - \frac{t}{d}, (k < n-1 \bigcup k > n) \\ 1 - \cos^{\omega}(\frac{\pi}{2}\frac{t}{d}), (k = n-1, n) \end{cases} \tag{10}$$

3. EVALUATIONS

We have invited five English learners to participate in the subjective evaluation. The first listening test is to evaluate the exaggeration degree the proposed methods can achieve. We show 40 pairs of animations to the subjects. These pairs cover all cases that mentioned in Section 2.1.5. Each pair contains a neutral animation and its exaggerated version. The subjects are requested to indicate the exaggeration degree based on a five-point scale: '1' (too weak to be perceived); '2' (slight exaggeration); '3' (moderate exaggeration); '4' (strong exaggeration); '5' (excessive exaggeration). From this objective test, we get a mean opinion score of 3.67. It proves that our method can achieve a good exaggeration effect.

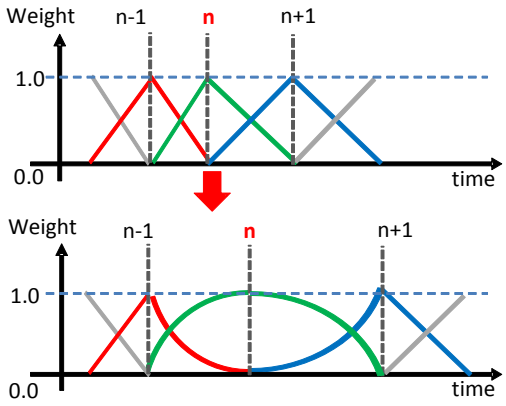

Fig. 3. The comparison of changes in blending weights before and after the emphasis of the n_{th} viseme.

The other listening test is to evaluate improvements in perception due to the exaggerated mode of feedback generation. We choose 13 minimal word pairs in this test. Each pair differs in one phoneme. For each minimal word pair, we generated one pair of neutral animations and another pair of exaggerated animations by exaggerating the differing phoneme. Animations in each pair were randomly placed. We played the overall 26 animation pairs to subjects and let them to discriminate between the given minimal word pair. Through this test, our subjects achieve 84.6% correctness on average for exaggerated animation pairs and 61.8% for non-exaggerated animation pairs. The results prove that the exaggeration methods can effectively improve the discrimination of the feedback. In the test, we found that our system achieved relatively worse results for the plosive pairs because the articulation characteristic of plosive is difficult to be captured and exaggerated. We will address this problem in our future work.

4. CONCLUSIONS

In this paper, we introduce. Audio speech is exaggerated by enhancing the pitch level, pitch variation and energy, pause insertion and phoneme duration lengthening. The exaggeration degree is controlled according to different phoneme conditions. For exaggerating the corresponding visual speech, we increase the articulatory movement range of the jaw and lips for "open" and "round" articulatory manners, and slow down the motions around the key action. The subjective evaluations prove that these methods can achieve fine exaggeration effects and make the feedback more discriminatively perceptible.

5. REFERENCES

[1] A. Neri, C. Cucchiarini, and H. Strik, "Feedback in computer assisted pronunciation training: technology push or demand pull?," *in Proc. ICSLP*, pp. 1209–1212, 2002b.

[2] R. Akahane-Yamada, Y. Tohkura, A. R. Bradlow, and D. B. Pisoni, "Does training in speech perception modify speech production," *in Proc. ICSLP*, pp. 606–609, 1996.

[3] Y. Tsubota, T. Kawahara, and M. Dantsuji, "Recognition and verification of english by japanese students for computer-assisted language learning system," *in Proc. Interspeech*, pp. 1205–1208, 2002.

[4] Y. Tsubotay, M. Dantsujiz, and T. Kawaharay, "Computer-assisted english vowel learning system for japanese speakers using cross language fomant structuers," *in Proc. ICSLP*, pp. 566–569, 3 2000.

[5] Y. Iribe, S. Manosavanh, K. Katsurada, and T. Nitta, "Generating animated pronunciation from speech through articulatory feature extraction," *in Proc. Interspeech*, pp. 1617–1621, 2011.

[6] Y. Iribe, S. Manosavanh, K. Katsurada, R. Hayashi, C. Zhu, and T. Nitta, "Improvement of animated articulatory gesture extracted from speech for pronunciation training," *in Proc. ICASSP*, pp. 5133–5136, 2012.

[7] J. Lu, R. Wang, L. C. D. Silva, Y. Gao, and J. Liu, "Castle: a computer-assisted stress teaching and learning environment for learners of english as a second language.," *in Proc. Interspeech*, pp. 606–609, 2010.

[8] J. Lu, R. Wang, and L. C. D. Silva, "Automatic stress exaggeration by prosody modification to assist language learners perceive sentence stress," *International journal of speech technology*, vol. 15, no. 2, pp. 87–98, 2012.

[9] K. H. Wong, W. K. Leung, W.K. Lo, and H. Meng, "Development of an articulatory visual-speech synthesizer to support language learning," *in Proc. ISCSLP*, pp. 139–143, 2010.

[10] K. H. Wong, W. K. Lo, and H. Meng, "Allophonic variations in visual speech synthesis for corrective feedback in capt," *in Proc. ICASSP*, pp. 5708–5711, 2011.

[11] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.