Visual-speech Synthesis of Exaggerated Corrective Feedback

Yaohua Bu*[1], Weijun Li*[2], Tianyi Ma[3,4], Shengqi Chen[3], Jia Jia†[3,4,5], Kun Li[6], Xiaobo Lu[1]
[1] Academy of Arts & Design, Tsinghua University
[2] School of Information Science and Technology, Northeast Normal University
[3] Department of Computer Science and Technology, Tsinghua University
[4] Key Laboratory of Pervasive Computing, Ministry of Education
[5] Beijing National Research Center for Information Science and Technology [6] SpeechX Ltd.

ABSTRACT

To provide more discriminative feedback for the second language (L2) learners to better identify their mispronunciation, we propose a method for exaggerated visual-speech feedback in computer-assisted pronunciation training (CAPT). The speech exaggeration is realized by an emphatic speech generation neural network based on Tacotron, while the visual exaggeration is accomplished by **ADC Viseme Blending**, namely increasing **A**mplitude of movement, extending the phone's **D**uration and enhancing the color **C**ontrast. User studies show that exaggerated feedback outperforms non-exaggerated version on helping learners with pronunciation identification and pronunciation improvement.

KEYWORDS

Corrective feedback, emphatic speech synthesis, visual-speech exaggeration, pronunciation learning

1 INTRODUCTION

Due to the influence of language transfer [5, 7, 16, 18, 23], learners tend to replace an unfamiliar phone in the second language by a phone from their first language(L1). It will cause many inconspicuous mispronunciations [15] that L2 learners hardly notice or correct them. There are many kinds of corrective feedbacks [1, 2, 6, 14, 26, 27] to increase the awareness between L1 and L2 in CAPT, such as speech, articulatory animations, etc. However, currently few feedback methods focus on the learners' intuitive to realize their mistakes. Thus, offering an identifiable and perceptible feedback is necessary for the development of proper pronunciation.

When the learners are having difficulties to realize their mispronunciation, teachers widely use the exaggerated method in teaching, such as speaking loudly and slowly to show the movements of mouth clearly [21]. Inspired by this, we propose a method for visual-speech synthesis of exaggerated feedback to point out user's mispronunciation by using emphatic speech and exaggerated articulatory animation. The speech and visual exaggeration are respectively realized by end-to-end emphatic speech synthesis and ADC Viseme Blending. Then they are combined to form the visual-speech synthesis[26]. By this method, we can provide more distinguishable feedback to learners and correct their mispronunciations.

2 IMPLEMENTATION

2.1 Speech Exaggeration

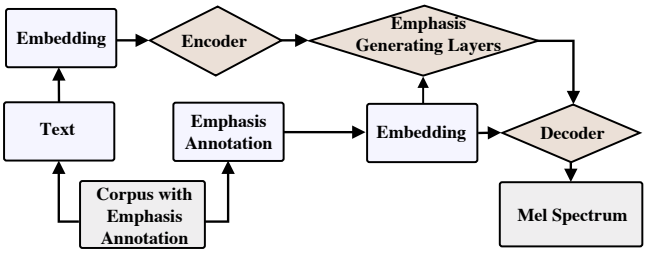

Figure 1: Architecture of speech exaggeration network

To synthesize emphatic speech, 8000 text prompts with phone-level emphasis annotations are carefully designed, each containing one or more emphatic phones that the Chinese ESL speakers often mispronounce [9, 17]. Two comparative speech utterances are recorded for each prompt: one with neutral intonation throughout the utterance and the other with exaggerated intonation with emphasis placed on the emphatic phones in the sentence.

We design a novel exaggerated speech generation neural network based on the previous Tacotron architecture [19, 20, 22, 25]. The network synthesizes exaggerated speech from input pairs, each including a sentence text and an emphasis vector containing phone level exaggeration annotation of the corresponding sentence. Formally the network could be described as $W = G(T_{l\times h}, A_{l\times 1})$, where W is the sound wave with exaggeration, T is the text input, A is the annotation, l is the length of a text,h is the length of a text vector that represents a word and G represents the whole network.

The architecture of the network is shown in Figure 1. The pre-net embeds text with word vector containing phonetic level information, which are in the same shape and then processed separately in the encoder layer. In the exaggeration generation layers, text and annotation vectors are combined, whose information are retained in output vector. To enhance the effect of exaggeration annotation, it is mixed with the output of generation layer again in the decoder layer, whose output is transformed into Mel spectrum. The

*Equal contribution
†Corresponding author, jjia@tsinghua.edu.cn

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10.
https://doi.org/10.1145/3394171.3414444

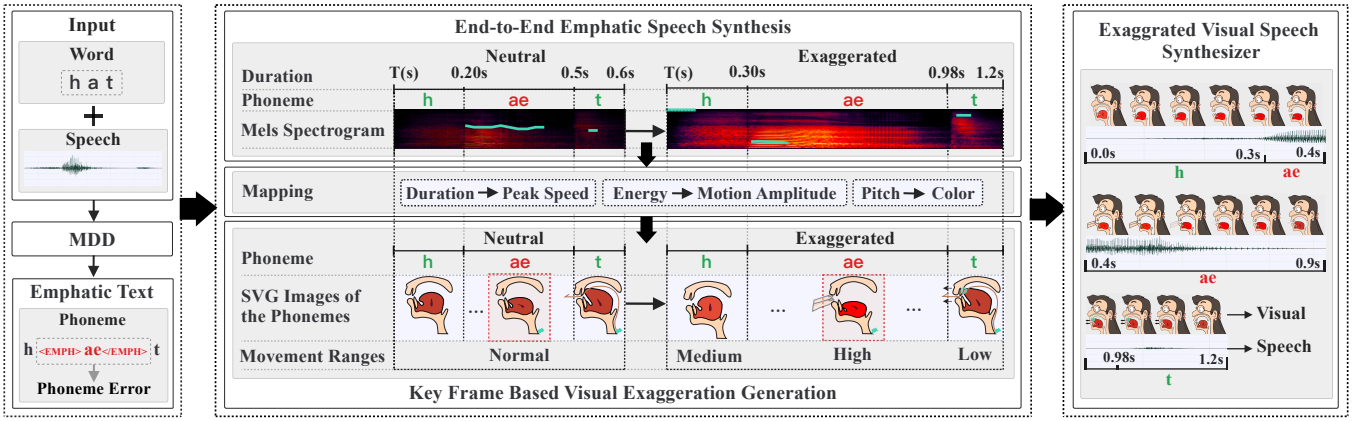

Figure 2: Flow chat of visual-speech synthesis of exaggerated corrective feedback

decoder also generates a sequence of the duration of each phoneme, which is later used by visual exaggeration. Finally, WaveNet [19] is appended to convert the Mel spectrum to speech with exaggeration.

2.2 Visual Exaggeration

We generate the exaggerated version of speech animation mainly by **ADC** Viseme Blending, which means increasing the **A**mplitude of articulatory movement, the **D**uration of movement around the key actions[29] and color **C**ontract of movement.

For each phone, we draw its corresponding visemes in SVG image according to the research on phonetics [4, 10] and articulatory phonology [3, 8, 28] in four exaggeration levels, namely *Normal*, *Low*, *Medium*, and *High*. Each image includes tongue, chin, teeth, soft jaw and lips (see as Figure 2). At each level, consonant and unit sounds each corresponds to one viseme, while each vowel has two. We then build a database for all these images.

To generate the animation of a specific phone sequence, we first sort out its corresponding viseme sequence, and obtain their duration from the network in Figure 1. Then the acoustic features of phones are mapped to visual features, which is shown in Figure 2 and described below.

For different motion **A**mplitudes determined by the phonemes' energy, we choose SVG file of different exaggeration levels for each viseme, which are used as the key frames of the animation and then interpolated with non-linear functions for smoother and more natural transition between non-exaggerated and exaggerated phones. Also they are made more prominent by extending its **D**uration. Meanwhile, colors with higher **C**ontrast, purity and brightness are used to make the exaggerated key frames easier to identify. We also add auxiliary graphics, such as arrows and airflow, to help users to better understand the pronunciation through visualization.

2.3 Visual-Speech Exaggeration Synthesis

The flow chart describing the whole process of visual-speech synthesis of exaggerated corrective feedback could be seen in Figure 2. We first decide the appropriate emphatic text by utilizing the Mispronunciation Detection & Diagnosis (MDD) model[11–13]. The output text from MDD with emphatic marks is fed to the network introduced in Section 2.1 to generate exaggerated speech feedback. Then we create the visual exaggeration in the format of SVG animation from the exaggerated speech by our proposed mapping rules as described in Section 2.2. Finally, speech and animation are combined to form the exaggerated visual-speech feedback.

3 USER STUDY

Three experiments are conducted to compare the exaggerated (E) visual-speech feedback with the non-exaggerated (N) version.

First, we prepare 28 questions containing visual-speech animation of 14 pairs of easily mispronounced words (such as bed and bad), and randomly decide whether each animation should use E or N-feedback. 32 participants are asked to finish the questionnaire by distinguishing the two words in the pair. The average identification accuracy of E and N-feedback animation is 93.30% and 75.45% respectively, which proves the effectiveness of exaggerated feedback on pronunciation identification.

The second experiment is to verify whether exaggerated feedback could improve the learner's pronunciation accuracy. 14 participants with average English level are divided into N and E groups and take courses with N and E-feedback respectively. The result shows that the average increase of accuracy of E and N group is 19.92% and 11.22% separately, which indicates exaggerated feedback could better improve the accuracy of pronunciation.

In addition, in order to make exaggerated feedback more in line with learners' cognition and play a better teaching effect, we invite 29 learners and 20 professional teachers to rate the degree of exaggeration with average opinion score (MOS) method[24]. For learners, the medium level of exaggeration gets the most three-point votes. For teachers, the medium level gets 86 three-point votes, which also prevails. Hence, the medium level is the most appropriate degree for exaggerated mispronounced phonemes.

ACKNOWLEDGEMENT

This work is supported by the state key program of the National Natural Science Foundation of China (No. 61831022) and the innovative research group project of the National Natural Science Foundation of China (No. 61521002).

REFERENCES

[1] Chesta Agarwal and Pinaki Chakraborty. 2019. A review of tools and techniques for computer aided pronunciation training (CAPT) in English. *Education and Information Technologies* 24, 6 (2019), 3731–3743.

[2] Pierre Badin, Atef Ben Youssef, Gérard Bailly, Frédéric Elisei, and Thomas Hueber. 2010. Visual articulatory feedback for phonetic correction in second language learning. In *Second Language Studies: Acquisition, Learning, Education and Technology*.

[3] Catherine P Browman and Louis Goldstein. 1992. Articulatory phonology: An overview. *Phonetica* 49, 3-4 (1992), 155–180.

[4] Philip Carr. 2019. *English phonetics and phonology: An introduction.* John Wiley & Sons.

[5] Nuria Calvo Cortés. 2005. Negative language transfer when learning Spanish as a foreign language. *Interlingüística* 16 (2005), 237–248.

[6] Tracey M Derwing and Murray J Munro. 2005. Second language accent and pronunciation teaching: A research-based approach. *TESOL quarterly* 39, 3 (2005), 379–397.

[7] James E Flege. 1995. Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research* 92 (1995), 233–277.

[8] Cécile Fougeron and Patricia A Keating. 1997. Articulatory strengthening at edges of prosodic domains. *The journal of the acoustical society of America* 101, 6 (1997), 3728–3740.

[9] Jia Jia, Wai-Kim Leung, Yu-Hao Wu, Xiu-Long Zhang, Hao Wang, Lian-Hong Cai, and Helen M Meng. 2014. Grading the Severity of Mispronunciations in CAPT Based on Statistical Analysis and Computational Speech Perception. *Journal of Computer Science and Technology* 29, 5 (2014), 751–761.

[10] Daniel Jones. 1922. *An outline of English phonetics.* BG Teubner.

[11] Kun Li, Jing Li, Yufang Song, and Hewei Fu. 2015. Rating Algorithm for Pronunciation of English Based on Audio Feature Pattern Matching. In *MATEC Web of Conferences*, Vol. 22. EDP Sciences, 01032.

[12] Kun Li, Xiaojun Qian, Shiyin Kang, Pengfei Liu, and Helen Meng. 2015. Integrating acoustic and state-transition models for free phone recognition in L2 English speech using multi-distribution deep neural networks.. In *SLaTE*. 119–124.

[13] Kun Li, Xiaojun Qian, and Helen Meng. 2016. Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 1 (2016), 193–207.

[14] Pengfei Liu, Ka-Wa Yuen, Wai-Kim Leung, and Helen Meng. 2012. menunciate: Development of a computer-aided pronunciation training system on a cross-platform framework for mobile, speech-enabled application development. In *2012 8th International Symposium on Chinese Spoken Language Processing*. IEEE, 170–173.

[15] Roy Lyster. 1998. Negotiation of form, recasts, and explicit correction in relation to error types and learner repair in immersion classrooms. *Language learning* 48, 2 (1998), 183–218.

[16] Helen Meng, Eric Zee, and Wai Sum Lee. 2007. A contrastive phonetic study between Cantonese and English to predict salient mispronunciations by Cantonese learners of English. *Unpublished article. The Chinese University of Hong Kong* (2007).

[17] Yishuang Ning, Zhiyong Wu, Jia Jia, Fanbo Meng, Helen Meng, and Lianhong Cai. 2015. HMM-based emphatic speech synthesis for corrective feedback in computer-aided pronunciation training. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4934–4938.

[18] Terence Odlin. 1989. *Language transfer*. Vol. 27. Cambridge University Press Cambridge.

[19] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).

[20] Aaron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George van den Driessche, Edward Lockhart, Luis C Cobo, Florian Stimberg, et al. 2017. Parallel wavenet: Fast high-fidelity speech synthesis. *arXiv preprint arXiv:1711.10433* (2017).

[21] Ellen Ricard. 1986. Beyond Fossilization: A Course in Strategies and Techniques in Pronunciation for Advanced Adult Learners. *TESL Canada Journal* (1986), 243–253.

[22] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4779–4783.

[23] Winifred Strange. 1995. Speech perception and linguistic experience: Theoretical and methodological issues.

[24] Robert C Streijl, Stefan Winkler, and David S Hands. 2016. Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. *Multimedia Systems* 22, 2 (2016), 213–227.

[25] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135* (2017).

[26] Ka-Ho Wong, Wai-Kit Lo, and Helen Meng. 2011. Allophonic variations in visual speech synthesis for corrective feedback in capt. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5708–5711.

[27] Ka-Wa Yuen, Wai-Kim Leung, Peng-fei Liu, Ka-Ho Wong, Xiao-jun Qian, Wai-Kit Lo, and Helen Meng. 2011. Enunciate: An internet-accessible computer-aided pronunciation training system and related user evaluations. In *2011 International Conference on Speech Database and Assessments (Oriental COCOSDA)*. IEEE, 85–90.

[28] Linghan Zhang, Sheng Tan, and Jie Yang. 2017. Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 57–71.

[29] Junhong Zhao, Hua Yuan, Wai-Kim Leung, Helen Meng, Jia Liu, and Shanhong Xia. 2013. Audiovisual synthesis of exaggerated speech for corrective feedback in computer-assisted pronunciation training. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 8218–8222.