HMM-BASED EMPHATIC SPEECH SYNTHESIS FOR CORRECTIVE FEEDBACK IN COMPUTER-AIDED PRONUNCIATION TRAINING

Yishuang Ning[1,3], Zhiyong Wu[1,2,3], Jia Jia[1,3,], Fanbo Meng[1,3], Helen Meng[1,2], Lianhong Cai[1,3]*

[1]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Shenzhen Key Laboratory of Information Science and Technology,
Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
[2]Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China
[3]Tsinghua National Laboratory for Information Science and Technology (TNList),
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

ningys13@mails.tsinghua.edu.cn, zywu@se.cuhk.edu.hk, jjia@tsinghua.edu.cn,
skywing32@gmail.com, hmmeng@se.cuhk.edu.hk, clh-dcs@tsinghua.edu.cn

ABSTRACT

This paper investigates the incorporation of hidden Markov model (HMM) based emphatic speech synthesis for audio exaggeration into an audio-visual speech synthesis framework for the corrective feedback in computer-aided pronunciation training (CAPT). To improve the voice quality of the synthetic emphatic speech, this paper proposes a new method for HMM training. In this method, the contextual questions for decision tree building are extended by considering the emphasis-related information. HMMs are then trained using a small scale emphatic corpus together with a large scale neutral corpus. The emphatic corpus is used to ensure the quality of the emphatic speech segments whereas the neutral corpus is to further improve the quality of both the non-emphatic speech segments and the emphatic ones. Finally, emphatic speech synthesis is achieved by extending the Flite+hts_engine. Experimental results show that our method can synthesize emphatic speech with high quality and make the feedback more discriminatively perceptible.

Index Terms—computer-aided pronunciation training (CAPT), emphatic speech synthesis, hidden Markov model (HMM)

1. INTRODUCTION

Multimodal information processing plays an important role in computer-aided pronunciation training (CAPT) [1][2] which uses exaggerated audio and visual speech to make the feedback discriminatively perceptible. The exaggerated audio and visual speeches are realized by audio and video exaggeration respectively. Our previous work has used perturbation model [4][5] to synthesize exaggerated audio speech and indicated audio-visual synthesis of exaggerated speech has much to offer for the multimodal corrective feedback [4]. The process of audio exaggeration is equivalent to emphatic speech synthesis. The framework of exaggerated audiovisual synthesis is shown in Fig. 1.

However, the problem of this method is that the speech quality may degrade a lot when the perturbation ratios are larger than a threshold. To address this problem, this work tries to use the parametric speech synthesis based on hidden Markov model (HMM) [6] to synthesize exaggerated speech considering its flexibility and high naturalness.

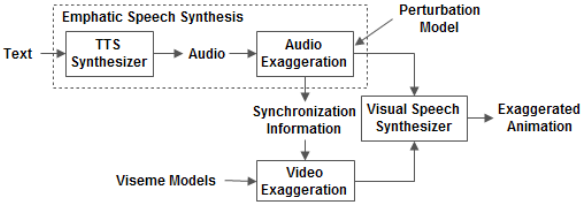

Fig. 1. The framework of exaggerated audiovisual synthesis.

This paper aims to implement audio exaggeration with the HMM framework. Different from the previous work, we achieve the purpose of synthesizing emphatic speech by extending the Flite+hts_engine [7][8] with the emphatic speech synthesis model. Previous work shows that context model [9] can achieve the highest quality and naturalness of the synthesized speeches. Therefore, we choose this model to synthesize emphatic speech. We also propose a new method for HMM training to further improve the voice quality of the synthesized speech. Subjective evaluation experiments show that our method can synthesize emphatic speech with high quality, thus obtaining better effect of language learning.

2. CORPORA

2.1. Emphatic corpus

To synthesize emphatic speech, a set of text prompts are carefully designed and contrastive speech utterances are recorded for the analysis and modeling of emphasis.

350 text prompts are designed, with each text prompt containing one or more emphatic words. These emphatic words are located at different positions in the sentences. For the emphatic words, they might be monosyllabic or polysyllabic, with the primary stressed syllables at different places in the words. Besides, the design of these text prompts considers all kinds of

pronunciation mechanisms of phones. The context characteristics of the phones are also covered by the text prompts as many as possible.

For each of the text prompts, two contrastive speech utterances are recorded - one with neutral intonation throughout the utterance and the other with expressive intonation with emphasis placed on the emphatic words in the sentence. A female speaker with a high level of English proficiency was invited to record the contrastive speech utterances in a sound proof studio. Hence we have 700 recorded utterances, saved in the wav format as sound files (16 bit mono, sampled at 16 kHz).

From the 350 text prompts, 20 prompts are randomly selected as the test set for experimentation, all the others are used to train an emphatic HMM with emphasis-related questions (details on HMM training will be elaborated in Section 4).

2.2. Neutral corpus

In order to ensure the naturalness and quality of the synthesized speech, the CMU US ARCTIC clb corpus [10] with neutral speech recordings is used as the neutral corpus, which is actually an adaptation corpus.

The corpus contains 1,132 phonetically balanced utterances recorded by an US female speaker, and stored in the 16bit mono format as wav files with 16 kHz sampling rate. The corpus is automatically annotated by FestVox [11]. The phone, syllable and word boundaries are then generated from the annotation result. The context features related to phone, syllable, word, position, lexical stress, etc. are also derived.

2.3. Phone classification of emphatic corpus

From our previous research, we find that emphatic words will often affect the changes of the acoustic features of their neighboring words. For example, speaker tends to decrease the f0s of the post-emphasized words. To consider such effects, we classify the phones into 6 emphasis categories based on the locations and positions of the phones in relation with the nearest emphasized word and its primary stressed syllables at word and syllable layers:

- **Class 1 (I-P-E)**: phones *In* the *Primary* stressed syllable of an *Emphasized* word;
- **Class 2 (B-P-E)**: phones *Before* the *Primary* stressed syllable of an *Emphasized* word;
- **Class 3 (A-P-E)**: phones *After* the *Primary* stressed syllable of an *Emphasized* word;
- **Class 4 (N-B)**: phones in the *Neutral* word *Before* the emphasized word;
- **Class 5 (N-A)**: phones in the *Neutral* word *After* the emphasized word;
- **Class 6 (O-R)**: all *Other Remaining* phones.

The phones of a syllable are assigned the class with the lowest class number if they fall into more than one class. Fig. 2 illustrates this method of phone classification, where "PETERSON" and "OCCASION" are the emphasized words in the sentence.

I	have	met	PETERSON	on	one	OCCASION.
6	4	1 3	5	4	2 1	3

Fig. 2. An example of phone classification based on the location of stressed syllables in emphasized words, where "PETERSON" and "OCCASION" are the emphasized words.

3. METHOD

As has been introduced above, our previous work has used perturbation model to synthesize exaggerated audio speech. The speech quality may degrade greatly when the perturbation ratios exceed a threshold. This paper tries to implement audio exaggeration with HMM-based emphatic speech synthesis framework by extending the Flite+hts_engine [7][8] due to its flexibility and high naturalness. The Flite+hts_engine is an English TTS (Text-to-Speech) system, which integrates Flite and HTS engines to realize the two parts respectively.

3.1 Implementation of emphatic speech synthesis based on Flite+hts_engine

The emphatic speech synthesis based on Flite+hts_engine is divided into two parts: the text analysis part and the speech synthesis part. As shown in Fig. 3 and Fig. 4, in the text analysis part, the emphatic text is first processed to generate raw text and record the position of each emphatic word. The raw text is then passed through the text analysis module of the original Flite engine to generate the full labels. Full emphatic labels are then obtained from the full labels and the position of emphatic words. In the speech synthesis part, acoustic feature parameters of the emphatic speech are predicted by the emphatic speech synthesis model (i.e. improved context model in Fig. 3) and full emphatic labels with the HTS engine. In order to improve the quality of the synthesized speech, this paper proposes a new method for HMM training to get the improved context model using a small scale emphatic speech corpus together with a large scale neutral speech corpus.

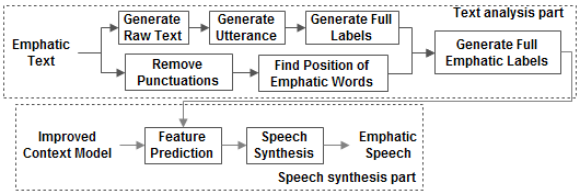

Fig. 3. The flowchart of emphatic speech synthesis based on Flite+hts_engine.

Emphatic Text: I have met <EMPH>Peterson</EMPH> on one <EMPH>occasion</EMPH>.
Raw Text: I have met Peterson on one occasion.
Position of Emphatic Words: 4 Peterson 7 occasion
Full Labels:
19574913 20545864 n^ax-k+ey=zh@1_2/A:0_0_1/B:1-1-2@2-2&10-2#6-1$5-3!2-0;2-1|ey/C:0+1+3/D:content_1/E:content+3@7+1&5+1#1+0/F:0_0/G:0_0/H:11=7^1=1|L-L%/I:0=0/J:11+7-1 occasion one 0 1 0 0 occasion occasion occasion
Full Emphatic Labels:
19574913 20545864 n^ax-k+ey=zh@1_2/A:0_0_1/B:1-1-2@2-2&10-2#6-1$5-3!2-0;2-1|ey/C:0+1+3/D:content_1/E:content+3@7+1&5+1#1+0/F:0_0/G:0_0/H:11=7^1=1|L-L%/I:0=0/J:11+7-1/EM:1

Fig. 4. Illustration of the terms appeared in Fig. 3, where "/EM:*i*" indicates that the emphasis category of the current phoneme is *i* which is between 1 and 6 (as stated in Section 2.3).

4. IMPROVED CONTEXT MODEL FOR EMPHATIC SPEECH SYNTHESIS

The framework of HMM-based speech synthesis generally can be divided into two stages: the training stage and the synthesis stage. The synthesis stage includes the text analysis part and the speech synthesis part. In this paper, to synthesize emphatic speech, the training stage is realized by training the improved context model, and the synthesis stage is realized based on the Flite+hts_engine.

4.1. Framework of emphatic speech synthesis

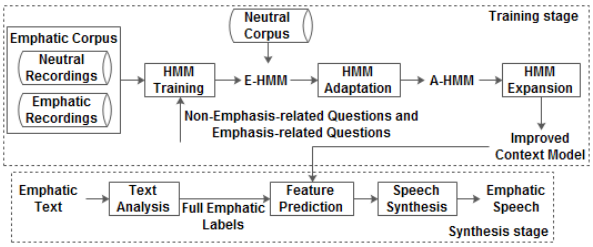

Fig. 5. The framework of emphatic speech synthesis.

Fig. 5 illustrates the framework of emphatic speech synthesis, including both training and synthesis stages.

During the training stage, two kinds of HMMs are involved: the emphatic HMM (E-HMM) and the adapted HMM (A-HMM). Firstly, the general decision tree (DT) is constructed using the emphatic corpus with minimum description length (MDL) [12] criterion and the E-HMMs are trained with the standard context questions (non-emphasis-related questions) from the official HTS toolkit [6] and emphasis-related questions.

In deriving the emphasis-related questions, we follow our phone categorization scheme (introduced in Section 2.3). These categories are used to compose 6 emphasis-related questions in the form of "Does the current phone belong to category *i*?" where *i* is one of the 6 emphasis categories listed above.

The purpose of the E-HMMs is to predict the acoustic feature parameters. These predicted parameters are generated from the E-HMMs trained from the emphatic corpus, thus ensuring the quality of emphatic speech segments. This is done by adding extended labels which contain the emphasis category information and emphasis-related questions to grow the decision tree of E-HMMs.

Since the emphatic corpus is very limited, only using these utterances for emphatic speech synthesis will degrade the quality of synthesized speeches greatly. Therefore, the large scale neutral corpus (proposed in Section 2.2) is used in the training process to improve the voice quality of the non-emphatic speech segments. Besides, seeing that the emphatic corpus and neutral corpus are recorded by different speakers, maximum likelihood linear regression (MLLR) [13] is used to adapt the parameters of the E-HMM using the data from the neutral corpus for each leaf node of the DT. After the above process, the A-HMMs will be generated.

However, due to the sparseness of emphasis data in the corpus, there may be no data in some sub-nodes, which results in no HMMs being trained for these sub-nodes. To address this problem, such sub-trees are expanded from their sibling nodes and parent nodes of the same emphasis-related sub-trees. The explanation of this method is to be elaborated in Section 4.2.

During the synthesis stage, full emphatic labels (proposed in Section 3.1) are provided to the improved context model to predict the acoustic feature parameters of the emphatic speech. Finally, the emphatic speech is synthesized by the Vocoder in the official hts_engine.

4.2. Improved context model

4.2.1. Motivation

When the emphasized words are located in different contexts, their acoustic characteristics are different. In the DT clustering process, the DT will cluster the data which are close to each other according to the context characteristics. The acoustic features of the nodes whose context characteristics are similar in the DT are close to each other. For example, the acoustic features of the data in the leaf nodes are similar to their sibling nodes and parent nodes. Therefore, it is reasonable to predict feature parameters from their sibling nodes and parent nodes when the current leaf node doesn't contain emphasis data of some emphasis categories. Based on this point, [9] proposes the context model. The training process of this model is as follows:

1) The neutral HMMs (N-HMMs) are trained from the large scale neutral corpus with non-emphasis-related questions.

2) Adapt these N-HMMs to the neutral speeches of the emphatic corpus with MLLR. After this step, the A-HMMs are generated.

3) Synthesize all the 1,132 speeches of the neutral corpus with the A-HMMs model. The purpose of these steps is to enlarge the neutral corpus, thus ensuring the naturalness of the synthesized speech and improving the quality of the non-emphatic speech segments.

4) The speech utterances synthesized from step 3), together with all the speech recordings of the emphatic corpus are used to train the E-HMMs.

In the above training process of the context model, synthetic speeches have been used to train the E-HMMs. However, the use of these synthetic speeches will influence the prediction of acoustic feature parameters, thus degrading the quality of the emphatic speech generated by the E-HMMs. Therefore, we propose the improved context model which uses the emphatic corpus to ensure the quality of emphatic speech segments and the neutral corpus to ensure the quality of non-emphatic speech segments.

4.2.2. Training the improved context model

To train the improved context model, five steps are involved.

1) The general DT is constructed using the training data from the emphatic corpus with MDL criterion. This clusters the training data (i.e. phones) into different leaf nodes according to non-emphasis-related questions. These leaf nodes are further classified into 6 categories with emphasis-related questions.

2) The data in each leaf node are used to train the E-HMMs.

3) Because of the small size of the emphatic corpus, using the speech parameters generated from these E-HMMs will affect the quality of the synthesized speeches. Therefore, this paper used a large scale neutral corpus to adapt the speech parameters of the E-HMMs. However, since the emphatic corpus and neutral corpus are recorded by different speakers, MLLR is then used to adapt the parameters of the E-HMMs using the data from the neutral corpus for each leaf node of the general DT. The phones of each sub-node belonged to the same emphasis category are used to adapt the HMMs from the leaf node of the general DT with MLLR to generate the A-HMMs.

4) Because the general DT is constructed with both emphasis-related questions and non-emphasis-related questions, some sub-nodes may not have data in some emphasis categories. To address this problem, the large scale neutral corpus which only has data in category 6 is used to expand such sub-nodes.

4.1) For a sub-node to be expanded, if the data in this node are all with the same emphasis category *x*, following steps are involved to generate a leaf node with category *i* (*i*≠*x*). First, a new leaf-node is generated and assigned as the child of the node to be expanded. The emphasis category of the newly added leaf node is set to be *i*. Then, the mean value of the acoustic feature parameters (such as f0, duration or energy) of this leaf node is calculated from the data of its parent nodes with this category. This is equivalent to predict feature parameters from its parent nodes, and is reasonable because the context characteristics of the leaf node are similar to its parent nodes.

4.2) If the sub-node to be expanded doesn't contain data in some category, the mean value of the feature parameters of newly added leaf node is calculated as follows. Suppose *m* is the category of the node to be expanded, $em\_mean_{K,j}$ and $em\_num_{K,j}$ are the mean value of the feature parameters and total number of the data in emphasis category *j* of the *K*-th sub-node. Suppose *P* is the parent node of the leaf node and contains data in category *m* (if parent doesn't contain such data, then check its grandparent node until find such node). Let *c* be the category of the data whose mean value is the closest to the data in category *m* of *P*. Then *c* can be calculated as:

$$c = \arg\min \left| em\_mean_{P,j} - em\_mean_{P,m} \right|, (1 \le j, m \le 6) \quad (1)$$

The mean value of the feature parameters of the leaf node in category *m* of the *K*-th sub-node is then calculated as:

$$em\_mean_{K,m} = em\_mean_{K,c} \times em\_mean_{P,m} / em\_mean_{P,c} \quad (2)$$

5) Check if each sub-tree has data in 6 emphasis categories. If a sub-tree doesn't have data in some category, then repeat step 4).

After the expansion, all the sub-trees of the DT have data in all categories, thus improving the voice quality of the emphatic speech segments.

5. EXPERIMENTS AND RESULTS

Two subjective listening tests were conducted to test our proposed approach.

The first listening test is to validate if the proposed method can generate emphasis that can be easily perceived. In this test, 20 prompts in the test set (as mentioned in Section 2.1) were used, with each prompt containing one or more emphatic words (embedded by "<EMPH>" and "</EMPH>"). For each prompt, we synthesized its neutral and emphatic speeches with the proposed methods. Both the raw text and the two speeches were provided to the subjects. The subjects were asked to identify the emphatic words in each pair, and also to give the perception degree of the identified emphatic words based on a five-point scale: '1' (too weak to be perceived); '2' (slight emphasis); '3' (moderate emphasis); '4' (strong emphasis); '5' (exaggerated emphasis). 10 subjects were invited in the test. The results are as follows. The precision and recall of perceived emphatic words are 98% and 94%. The mean perception degrees of the perceived emphatic words for the emphatic speech and neutral speech are 3.85 and 2 respectively. This means the emphatic words in the emphatic speeches are much easier to be perceived; although some of the neutral speeches are perceived to carry emphasis, it is really very hard to discriminate. The results prove that emphatic speech can efficiently improve the discrimination of the feedback.

The second listening test is to evaluate the quality of the synthesized speeches of the improved context model. In this test, 3 HMM-based models for emphatic speech synthesis are compared. The first model is the context model (denoted by "context"), the second one is trained without the neutral corpus (denoted by "non-neu") and the last one is the improved context model (denoted by "improved"). The same 20 prompts in the test set were used. For each prompt, 3 emphatic speeches were synthesized with the 3 models respectively. 10 subjects were invited and asked to listen to the synthetic emphatic speeches and to indicate the voice quality based on a five-point scale: '1' (bad); '2' (poor); '3' (fair); '4' (good); '5' (excellent). The average mean opinion scores (MOS) of different models are shown in Table 1, where the confidence intervals (CI) at confidence level 0.95 (α=0.95) are also given. As can be seen, the synthesized speeches of the "improved" model have the highest MOS score, which means the voice quality of our proposed method is the best among the 3 models.

Table 1. Experimental results on the quality of the synthesized speeches, where the confidence intervals (CI) of the mean opinion scores (MOS) are given at the confidence level α=0.95.

Models	MOS	CI (α=0.95)
context	2.67	[2.03, 3.31]
non-neu	3.19	[2.58, 3.80]
improved	3.58	[3.12, 4.03]

6. CONCLUSIONS AND FUTURE WORK

This paper investigates incorporating HMM-based emphatic speech synthesis for audio exaggeration into an audio-visual speech synthesis framework in the CAPT system to offer corrective feedback. Addressing the voice quality of the synthesized speech, we propose an improved context model – a new method for HMM training, which uses the emphatic corpus to ensure the quality of the emphatic speech segments and the neutral corpus to ensure the quality of both the emphatic speech segments and the non-emphatic ones. The emphatic speech system is implemented by extending the Flite+hts_engine. Experiments show this approach can not only synthesize emphatic speeches with high quality, but also effectively improve the discrimination of the feedback.

Future work will incorporate this system into the CAPT system to offer corrective feedback for pronunciation training.

7. ACKNOWLEDGEMENTS

This work was conducted when the first author was a summer intern in the Human-Computer Communications Laboratory, The Chinese University of Hong Kong (CUHK). We wish to acknowledge the CUHK OALC summer internship program. The work is jointly supported by the research funds from the Hong Kong SAR Government's Research Grants Council (N-CUHK414/09), the National Natural Science Foundation of China (61375027, 61370023), and the National Society Science Foundation of China (13&ZD189).

8. REFERENCES

[1] F.B. Meng, Z.Y. Wu, J. Jia, H. Meng, L.H. Cai, "Synthesizing English emphatic speech for multimodal corrective feedback in computer-aided pronunciation training," *Multimedia Tools and Applications*, 2014, 73(1): 463-489, DOI: 10.1007/s11042-013-1601-y.

[2] H. Meng, W.K. Lo, A.M Harrison, P. Lee, K.H. Won, W.K. Leung, F.B. Meng, "Development of automatic speech recongition and synthesis technologies to support Chinese learners of English: The CUHK experience," in *Proc. of Asia Pacific Signal and Information Processing Association (APSIPA)*, 2011.

[3] Z.Y. Wu, Y.S Ning, X. Zang, J. Jia, F.B. Meng, H. Meng, L.H. Cai, "Generating emphatic speech with hidden Markov model for expressive speech synthesis," *Multimedia Tools and Applications*, 2014, DOI: 10.1007/s11042-014-2164-2.

[4] J.H. Zhao, H. Yuan, W.K. Leung, J. Liu, S.H. Xia and H. Meng, "Audiovisual Synthesis Of Exaggerated Speech For Corrective Feedback In Computer-Assisted Pronunciation Training," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 8218-8222, 2013.

[5] F.B. Meng, H. Meng, Z.Y. Wu, L.H. Cai. "Synthesizing expressive speech to convey focus using a perturbation model for computer aided pronunciation training," in *Proc. of the Second Language Studies: Acquisition, Learning, Education and Technology*, 22-27, 2010.

[6] K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A. Black, T. Nose, "The HMM-based speech synthesis system (HTS) version 2.1," *http://hts.sp.nitech.ac.jp/*, 2008.

[7] http://www.speech.cs.cmu.edu/flite/.

[8] http://hts-engine.sourceforge.net/.

[9] F.B. Meng "Analysis and generation of focus in continuous speech," *PhD. Thesis*, Tsinghua University, Beijing, 2013.

[10] J. Kominek, A.W. Black, "CMU ARCTIC databases for speech synthesis," *Tech. Rep. CMU-LTI-03-177*, Carnegie Mellon University, 2003.

[11] http://www.cstr.ed.ac.uk/projects/festival/.

[12] Shinoda, K., Watanabe, T., "MDL-based context-dependent subword modeling for speech recognition," *Acoust. Soc. Japan (E)*, 21:79-86, 2000.

[13] Leggetter, C.J., Woodland, P.C., "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Computer Speech and Language*, 9: 171-186, 1995.