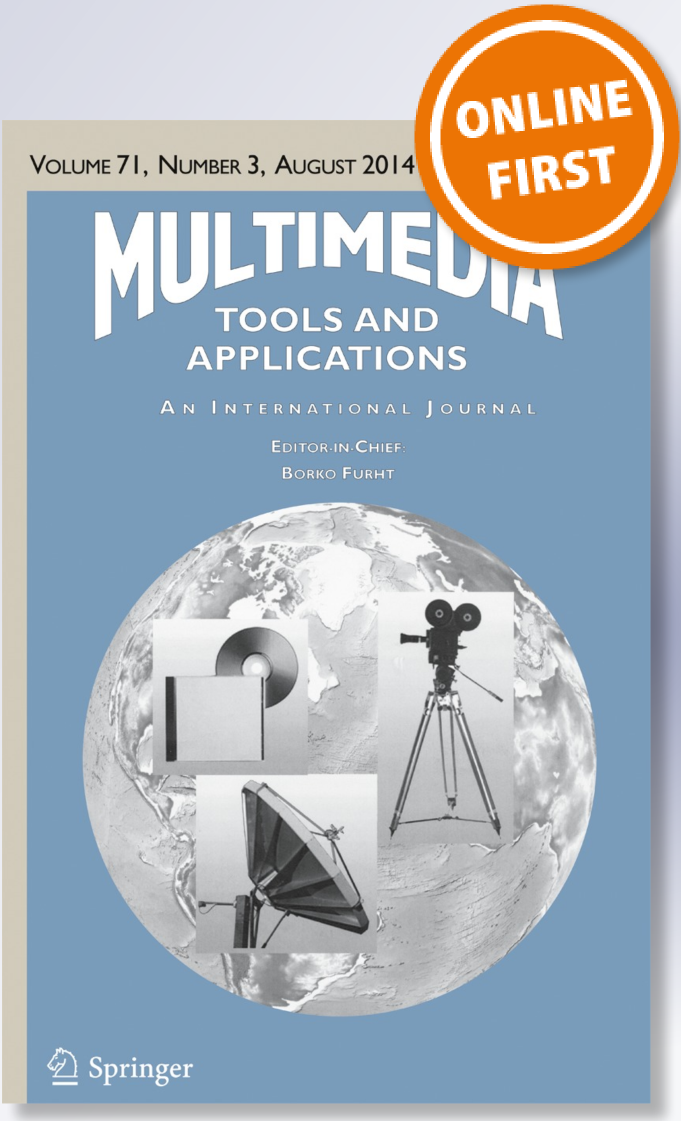
Generating emphatic speech with hidden Markov model for expressive speech synthesis

Zhiyong Wu, Yishuang Ning, Xiao Zang, Jia Jia, Fanbo Meng, Helen Meng & Lianhong Cai

Multimedia Tools and Applications
An International Journal

ISSN 1380-7501

Multimed Tools Appl
DOI 10.1007/s11042-014-2164-2



Multimed Tools Appl
DOI 10.1007/s11042-014-2164-2

Generating emphatic speech with hidden Markov model for expressive speech synthesis

Zhiyong Wu · Yishuang Ning · Xiao Zang · Jia Jia · Fanbo Meng · Helen Meng · Lianhong Cai

Received: 5 March 2014 / Revised: 1 June 2014 / Accepted: 23 June 2014

Abstract Emphasis plays an important role in expressive speech synthesis in highlighting the focus of an utterance to draw the attention of the listener. As there are only a few emphasized words in a sentence, the problem of the data limitation is one of the most important problems for emphatic speech synthesis. In this paper, we analyze contrastive (neutral versus emphatic) speech recordings considering kinds of contexts, i.e. the relative locations between the syllables and the emphasized words. Based on the analysis, we propose a hidden Markov model (HMM) based method for emphatic speech synthesis with limited amount of data. In this method, decision trees (DTs) are constructed with non-emphasis-related questions using both neutral and emphasis corpora. The data in each leaf node of the DTs are classified into 6

Z. Wu · Y. Ning · X. Zang (✉) · J. Jia · F. Meng · H. Meng · L. Cai
Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, and Shenzhen Key Laboratory of Information Science and Technology, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
e-mail: zangxiaocs@163.com

Z. Wu
e-mail: zywu@se.cuhk.edu.hk

Y. Ning
e-mail: ningys13@mails.tsinghua.edu.cn

J. Jia
e-mail: jjia@tsinghua.edu.cn

F. Meng
e-mail: skywing32@gmail.com

H. Meng
e-mail: hmmeng@se.cuhk.edu.hk

L. Cai
e-mail: clh-dcs@tsinghua.edu.cn

Z. Wu · H. Meng
Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong, SAR, China

Z. Wu · Y. Ning · X. Zang · J. Jia · F. Meng · L. Cai
Tsinghua National Laboratory for Information Science and Technology (TNList), and Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Published online: 12 July 2014

emphasis categories according to the emphasis-related questions. The data in the same emphasis category are grouped into one sub-node and are used to train one HMM. As there might be no data of some specific emphasis categories in the leaf nodes of the DTs, a method based on cost calculation is proposed to select a suitable HMM in the same leaf node for predicting parameters. Further a compensation model is proposed to adjust the predicted parameters. We conduct a series of experiments to evaluate the performances of the approach. Experiments indicate that the proposed emphatic speech synthesis models improve the emphasis quality of synthesized speech while keeping a high degree of the naturalness.

Keywords Emphasis · Emphatic speech synthesis · Hidden Markov model (HMM) · Decision tree clustering · Parameter compensation

1 Introduction

State-of-the-art speech synthesis technologies can generate speech with high naturalness [2, 11, 14]. However, effective human-computer interaction needs the generation of expressive speech to properly convey the message [9]. Emphasis is an important feature of expressivity and plays an important role in the expression of spoken language. Emphasis synthesis can be useful in many human-computer interaction scenarios, e.g. to highlight important words to attract user's attention in spoken dialog system [11, 14, 18], to provide corrective feedback for computer-aided pronunciation training [7, 9], etc.

To synthesize emphasis, two kinds of methods have been proposed, including concatinative based method to concatenate units carrying emphasis [11], and parametric speech synthesis with hidden Markov model (HMM) [13]. With the framework of concatenative based synthesis, [5] analyzed the duration pattern of emphasis and proposed a rule-based emphasis synthesis model. [19] proposed an acoustic feature prediction model with decision tree (DT) and Gaussian mixture model (GMM) to guide unit selection. [17] tried to model the most perceivable sections of the pitch curves of emphasis. For HMM based emphatic speech synthesis, [6, 10] added emphasis-related questions in the traditional HMM framework [13] for DT construction. However, directly putting emphasis-related questions into the question set may have little effect. It is because that the emphasis data is much less than non-emphasis data leading to the problem that the emphasis-related questions are rarely used while growing the DT. To address the problem, several methods have been proposed to supervise the growth of DT. [18] proposed two methods for growing DT. The first one is the two-pass DT. Emphasis-related questions are first used to grow the main DT and then non-emphasis-related questions are used to expand all the leaf nodes of the main DT. The second one is the factorized DT. First, two DTs are grown with non-emphasis-related questions and emphasis-related questions separately. The structure of the latter DT is used to expand all the leaf nodes of the former DT. Besides, [15] proposed the method of re-sampling by repeating the emphasis data which has low proportion in all data. With this method, the discriminative powers of emphasis-related questions can be improved as the proportion of emphasis data is improved.

In general, the challenges for synthesizing emphatic speech based on HMM are the following:

- How to cluster the training data based on DTs as both accurately and flexibly as possible while avoiding data sparseness?
- How to generate the acoustic parameters of emphasis of different contexts, especially the contexts out of training set?

To address the above challenges, we try to conduct a systematic investigation of the problem. A framework for emphatic speech synthesis based on HMM is proposed. This framework contains three improvements for emphatic speech synthesis compared to the traditional HMM-based speech synthesis model: 1) At the stage of growing DT, a main DT is first grown using non-emphasis-related questions to maintain the naturalness of synthesized speech, and then its leaf nodes are expanded using emphasis-related questions to cluster the emphasis data and the non-emphasis into different sub-nodes. With the DT of this specific structure, the problem of data sparseness only occurs in some sub-nodes, while there are enough data in their parent nodes and their neighborhood sub-nodes which can help deal with the problem of data sparseness. 2) Due to the limitation of the training data, there are no emphasis data in some leaf nodes of the main DT. The framework contains a method based on cost calculation to select suitable HMMs for synthesizing emphasis out of training data. 3) Though the suitable HMMs are selected for synthesizing emphasis out of training data, there may be still some differences between the parameters generated by the HMMs and those of emphasis target. Hence, the framework contains a compensation model to adjust the generated parameters to improve the emphasis intensity of synthesized speech.

The rest of the paper is organized as follows. Section 2 presents the corpora used for data analysis and model training, together with the acoustic analysis of the English emphasis corpus. Section 3 gives details of the models for emphatic speech synthesis. Section 4 presents and discusses the experiments conducted to evaluate different models and the overall approach. Finally, Section 5 lays out conclusions.

2 Corpora

2.1 Emphasis 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.

2.1.1 Design of text prompts

We carefully designed a set of text prompts (351 in all) for recording the emphasis corpus by considering the factors affecting the expression of emphasis at word, syllable and phone layer. For word layer, one or more emphasized words are contained in each text prompt, with each emphasized word located at different positions in the sentences. For syllable layer, these emphasized words might be monosyllabic or polysyllabic, with the primary stressed syllables at different places in the words. For phone layer, the phones with all kinds of pronunciation mechanisms are covered by the text prompts. The contexts of the phones are also covered by the text prompts as many as possible.

Two example text prompts are shown as follows, with emphasized words in bold face and capitalized:

"*Fighting **THIRST** is the **FIRST** thing to be done in this country.*" and
"*I have met **PETERSON** on one **OCCASION***."

2.1.2 Contrastive speech recordings

Two contrastive speech utterances are recorded for each of the text prompt - one with neutral intonation throughout the utterance and the other with expressive intonation with emphasis placed on the emphasized 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). Phone boundaries are located automatically by means of forced alignment with an automatic speech recognizer that is trained on the TIMIT database [8]. Pitch tracking is done by Praat [1]. Smoothing is performed in the f0 trajectory and phone segments with obvious errors (amounting to about 5 %) are excluded from subsequent analysis.

From the 350 text prompts, 20 prompts are randomly selected as the test set for experimentation, all the other prompts are used as the training set.

2.2 Neutral corpus

In addition to the emphasis corpus, the CMU US ARCTIC clb corpus [3] with neutral speech recordings is used as the neutral corpus. This neutral corpus is used to train the HMM models for generating speech with a high degree of naturalness.

The corpus contains 1132 phonetically balanced utterances recorded by an US female speaker, and stored in the 16 bit mono format as wav files with 16 kHz sampling rate. The corpus is automatically annotated by FestVox (http://www.cstr.ed.ac.uk/projects/festival/). 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. To ensure the accuracy of data analysis, the f0s of contrastive speech recordings of the emphasis corpus are manually checked and corrected before being used to train HMM models.

2.3 Acoustic analysis of the emphasis corpus

In emphatic speech, emphasized 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 [16]. 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. Figure 1 illustrates this method of phone classification. "PETERSON" and "OCCASION" are the emphasized words in the sentence.

The phones in the syllables with each emphasis category are further classified into 9 types according to their pronunciation mechanisms, as shown below. The phones in the same phone type share similar parameter representation of acoustic features. While the acoustic variations for the phones from different types are different even in the same emphasis category. For example, the duration changes of the vowels are larger than those of the consonants when comparing emphatic speech with neutral one.

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

Fig. 1 An example of phone classification (emphasis category) based on the location of stressed syllables in emphasize words

- **Type 1**: long vowel and diphthong, e.g. [i:], [ei];
- **Type 2**: mono vowel, e.g. [i];
- **Type 3**: plosive, e.g. [p];
- **Type 4**: nasal, e.g. [m];
- **Type 5**: fricative, e.g. [z];
- **Type 6**: retroflex liquid, e.g. [r];
- **Type 7**: lateral liquid, e.g. [l];
- **Type 8**: glide, e.g. [y];
- **Type 9**: affricate, e.g. [tʃ].

The data (i.e. phones) from the emphasis corpus of all 350 text prompts are used to perform the statistics of acoustic features (f0 and duration). The average f0s and durations of the phones with different emphasis categories and different phone types are shown in Table 1.

3 Modeling emphatic speech with HMM

3.1 Framework for emphatic speech synthesis

Figure 2 illustrates the diagram of the proposed model for emphatic speech synthesis with hidden Markov model (HMM). Firstly, a decision tree is constructed using the training data from both neutral and emphasis corpora, which clusters the training data (i.e. phones) into different leaf nodes according to a set of context questions without considering emphasis information (i.e. "non-emphasis-related questions" in Fig. 2).

After the above process, each leaf node of the decision tree may contain phones with different emphasis attributes, e.g. from emphasized or non-emphasized word. The HMMs

Table 1 Statistics of the average duration (*D*, in ms) and average f_0 (f_0, in Hz) of the phones from different emphasis category (EC) and phone type (PT)

PT	EC											
	I-P-E		B-P-E		A-P-E		N-B		N-A		O-R	
	D	f_0	D	f_0	D	f_0	D	f_0	D	f_0	D	f_0
Long vowel and diphthong	79	207	72	190	61	182	55	189	47	179	49	188
Mono vowel	65	217	33	192	39	183	39	189	37	177	33	187
Plosive	127	191	92	185	100	183	78	179	74	159	70	180
Nasal	40	188	38	191	27	182	35	181	36	167	30	182
Fricative	76	189	53	183	55	215	60	175	63	168	53	183
Retroflex liquid	68	192	58	188	70	189	45	193	39	180	46	185
Lateral liquid	61	195	52	188	41	184	36	179	32	171	44	184
Glide	155	187	125	181	89	193	115	177	121	173	118	177
Affricate	103	188	70	192	48	190	107	224	58	172	60	189

trained using the data from such leaves can generate speech with high naturalness but with low emphasis quality. To address the problem, we classify the phones into 6 emphasis categories by considering their relations with stressed syllables in emphasized words. The data in each leaf node are then classified into different sub-nodes according to the emphasis category (i.e. "emphasis-related questions at word and syllable layer" in Fig. 2). The phones of the same emphasis category are grouped into one sub-node and are used to train one HMM for emphatic speech synthesis.

However, a leaf node of the decision tree may also contain no data for a specific emphasis category. Based on the statistics of the acoustic (f0 and duration) differences between different emphasis categories at phone layer, a cost function is designed to select a most appropriate HMM to generate parameters for later speech synthesis (i.e. "HMM selection" in Fig. 2). But this HMM is trained using the data with another emphasis category, and this emphasis category is different from the target emphasis category. To address the acoustic differences between the selected emphasis category and the target emphasis category, a compensation model is further proposed to adjust the parameters generated by the HMM. The adjusted parameters are finally sent to speech synthesis module to generate the final emphatic speech.

3.2 HMM training for emphatic speech synthesis

This section explains our method on how to train the HMM models from the corpora (with both emphasis corpus and neutral corpus). That is how to derive the "HMMs" as shown in Fig. 2.

3.2.1 Decision tree construction with non-emphasis-related questions

In the framework of HMM based speech synthesis, the training data are clustered into different groups using the decision tree with the data in each group (i.e. leaf node of the decision tree) sharing the same set of contexts and similar acoustic parameters.

In this work, we construct the general decision tree with non-emphasis-related questions using the training data from the neutral corpus with the minimum description length (MDL) criterion [12]. This grows a general decision tree according to 1488 standard context questions (i.e. non-emphasis-related questions in Fig. 2) from the official HTS toolkit [13]. These non-

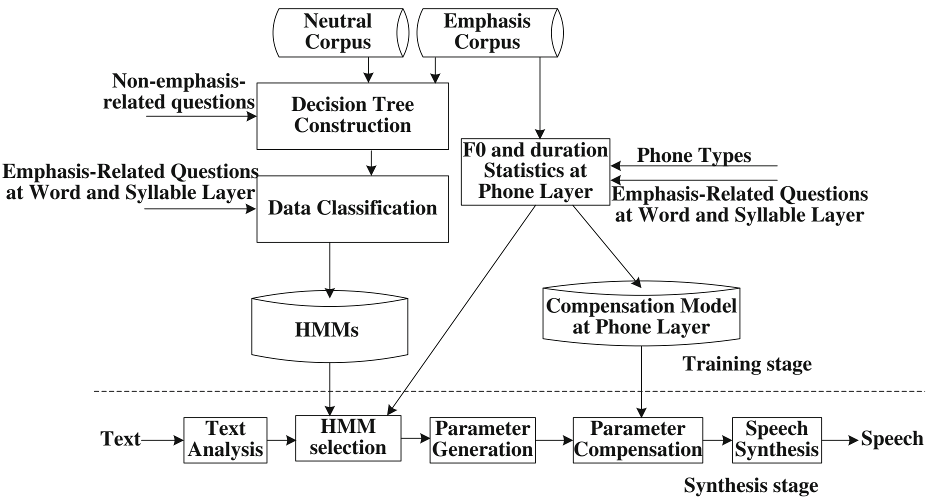


Fig. 2 Diagram of the model for emphatic speech synthesis with hidden Markov model

emphasis-related context questions are related to phones, positions, syllables, words, lexical stress, pitch accent, etc. For example, "Is the current phone [ey]?", "Does the number of the syllables in the next word equal to 1?", and so on.

The general decision tree is then used to group the phones of the emphasis corpus into different clusters (i.e. leaf nodes). The data in each leaf node of the decision tree satisfy the target non-emphasis-related context requirements of the input text. The HMMs trained using the data from such leaves can ensure the naturalness of the synthetic speech.

However, *two problems* exist for the above procedure. First, in the leaves of the decision tree, the data from emphasized and non-emphasized words may be mixed. Second, there might be no emphasis related data in some leaves. The two problems decrease the emphasis quality of the synthetic speech.

To address the problems, we perform data classification with emphasis-related questions for the first problem and propose a method to select HMM using cost function and to refine parameters with compensation model for the second problem.

3.2.2 Data classification with emphasis-related questions

To classify the data in each leaf of the decision tree, additional emphasis-related questions are introduced. In deriving the emphasis-related questions, we follow our phone categorization scheme as proposed in Section 2.3 to classify the phones into 6 categories according to phones' relative locations and distances to the nearest emphasized words and the primary stressed syllables at word and syllable layers. These categories are used in composing 6 emphasis-related context questions (i.e. "emphasis-related questions at word and syllable layer" in Fig. 2) in the form of "Does the current phone belong to category *i*?" where *i* is one of the 6 emphasis categories listed above.

The data in each leaf node of the decision tree are classified into different sub-nodes according to emphasis-related questions. The data of each sub-node are then used to train one HMM. As each sub-node only contains the data of the same emphasis category, the emphasis quality of synthetic speech can be improved.

3.2.3 HMM training for emphatic speech synthesis

To train the HMMs for emphatic speech synthesis, following steps are involved.

1) The general decision tree is used to group the phones of the neutral corpus into different leaf nodes. The neutral HMMs are trained using the data from each leaf node of the general decision tree.
2) The same general decision tree is used to group the phones of the emphasis corpus into different leaf nodes. As stated in Section 2, the emphasis and neutral corpora are recorded by two different speakers. Maximum likelihood linear regression (MLLR) [4] is used to adapt the parameters of the above neutral HMMs (as in step 1) using the data from the emphasis corpus for each leaf node of the general decision tree.
3) The emphasis-related sub-trees are further used to divide the data in each leaf node of the general decision tree into sub-nodes. The phones of each sub-node belong to the same emphasis category and are used to adapt the HMM (as in step 2) from the parent leaf node of the general decision tree with MLLR [4] to get the final HMMs for emphatic speech synthesis.

However, due to the limited amount of emphasis data in the corpus, there might be no data in some sub-nodes, therefore no HMM can be trained for these sub-nodes. For example, about

55 % of the leaf nodes of the general decision tree are found to contain no data in the "I-P-E" emphasis category when they are further split based on the emphasis-related sub-trees. This leads to the problem that no HMM of that specific emphasis category can be trained for parameter generation.

To solve the issue, a cost function is designed to select the most appropriate HMM from other leaves of the same emphasis-related sub-tree. As the selected HMM is derived from the same leaf of the general decision tree, with the non-emphasis-related contexts, the naturalness of the synthetic speech can be maintained.

3.3 HMM selection for parameter generation

In selecting the most appropriate HMM, a cost function is designed based on the analysis of the f0 and duration differences between different emphasis categories at phone layer.

If there is no data for the emphasis category in the current leaf node of the general decision tree, the cost function is designed to select a most appropriate HMM of the other sub-node coming from the same leaf node of the general decision tree. The HMMs with the least cost are used for parameter generation for speech synthesis.

As the decision trees for f0 and duration are constructed separately, the process of selecting HMMs for generating f0 and duration are also carried out respectively, whilst the process are the same. The following illustration takes f0 as example.

Suppose we are going to synthesize speech for a phone whose emphasis category is e and phone type is p. The leaf node L of the general DT satisfies the non-emphasis-related contexts of this target phone. For the sub-node K of the leaf node L, let the emphasis category of the data in this sub-node be m. The cost for using the HMM trained from this sub-node K to generate f0 is calculated as follows.

If the emphasis category e of the target phone is the same as m, the cost is 0. Otherwise, suppose there are N phones in the sub-node K. Let n_t be the number of the phones whose phone type is t in the sub-node K, and $N=n_1+n_2+\ldots+n_9$. Here, 9 is the total number of phone type as in Section 2.3. If there is no data whose phone type is t in the sub-node K, $n_t=0$. Then the cost function is defined as:

$$C = \begin{cases} 0 & , \text{ if } e = m \\ \left| 1 - \frac{1}{f_{0e,p}} \frac{1}{N} \sum_{t=1}^{9} n_t f_{0m,t} \right|, & \text{if } e \neq m \end{cases} \quad (1)$$

where $f_{0e,p}$ is the average f0 for all the phones whose emphasis category is e and phone type is p; and $f_{0m,t}$ is the average f0 for all the phones whose emphasis category is m and phone type is t. These statistical values are all taken from Table 1.

For example, let the sentence to be synthesized be "**take** it please", where "take" is the emphasized word. Let the current phone to be synthesized be the second phone of "take", which is [ey]. Part of the decision tree for generating f0 is show in Fig. 3. As the current phone is [ey] and the number of the syllables of the next word "it" is 1, the data in the leaf node annotated by "*" will be used for generating f0. Therefore the target phone is [ey] for the emphasis category "I-P-E". However, data is available for only two emphasis categories "B-P-E" (K_1) and "A-P-E" (K_2) in the leaf node of the general DT. To generate f0 for the diphthong [ey] with emphasis category "I-P-E" (whose average f0 in Table 1 is 207), let C_1 and C_2 be the cost of using the HMM trained with the data in K_1 and K_2 respectively. To calculate C_1, the emphasis category of the data in K_1 is "B-P-E", the average f0 for the diphthongs [ey] and [ow] in Table 1 is 190; the average f0 for the mono vowel [ae] is 192; and the average f0 for the affricate [ch] is 192. To calculate C_2, the emphasis category of the data in K_2 is "A-P-E", the average f0 for the nasals [m] and [n] is 182. The costs are then calculated as Eq. (2) and the HMM trained by the data in K_1 is selected for generating f0.

$$C_1 = \left|1 - \frac{1}{207}\frac{1}{4}(2 \times 190 + 192 + 192)\right| = 0.08, C_2 = \left|1 - \frac{1}{207}182\right| = 0.12 \tag{2}$$

3.4 Compensation model for emphasis synthesis

For the emphasis category having no data in the current leaf node, an HMM of the other sub-node (with a different emphasis category) from the same leaf node of the general decision tree is selected by the cost function to generate parameters (f0 or duration) for speech synthesis. This will cause the emphasis category of the data used for parameter generation to be different from the target emphasis category, which reduces the emphasis quality of the synthetic speech. To alleviate this problem, a compensation model is further proposed to adjust the f0 and duration generated by the HMM at the phone layer.

For the target phone to be synthesized, let $\mathbf{F}(n)$ be the f0 sequence generated by the HMM trained with the data in the sub-node K. Let the new f0 sequence after compensation be $\mathbf{F}'(n)$, which can be calculated as:

$$\mathbf{F}'(n) = R_{f_0} \times \mathbf{F}(n) \tag{3}$$

where R_{f0} is the compensation factor for f0, which can be computed as Eq. (4) using the statistic information from the emphasis corpus as shown in Table 1, where 9 is the total number of phone types as illustrated in Section 2.3.

$$R_{f_0} = \begin{cases} 1 & \text{, if } e = m \\ \dfrac{f_{0e,p}}{\frac{1}{N}\sum_{t=1}^{9} n_t f_{0m,t}} & \text{, if } e \neq m \end{cases} \tag{4}$$

where the notations for $f_{0e,p}$, $f_{0m,t}$ and e, p, m, t are all the same as those in Eq. (1). Especially, if the target emphasis category e is the same as the emphasis category m of the sub-node K, no compensation is needed, and R_{f0}=1.

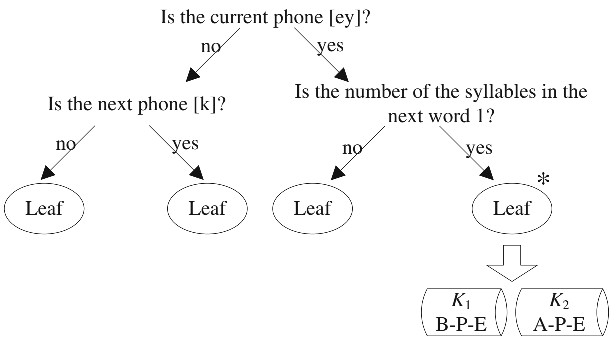


Fig. 3 Part of the decision tree for generating f0. Data is available for only two emphasis categories B-P-E (K_1) and A-P-E (K_2) in the leaf node annotated by "*". There are four phones in K_1: [ey][ow][ae][ch] and two phones in K_2: [m][n]

Recall the example in Section 3.3, the target phone is [ey] in the emphasized word "take". The HMM trained by the data in K_1 is used for generating the f0s, the compensation factor for f0 is calculated as:

$$R_{f_0} = \frac{207}{(2 \times 190 + 192 + 192)/4} = 1.08 \quad (5)$$

The method for compensating the durations is the same as that for compensating the f0s.

The new compensated f0s and durations are then feed to the official HTS toolkit [13] to generate the synthetic emphatic speech.

4 Experiments

4.1 Experimental setup

To test our proposed approach, we conduct a set of experiments on the emphatic corpus and the neutral corpus. The experimental results validate the effectiveness of our approach. Three HMM-based models for emphatic speech synthesis are compared in our experiments:

1) The first model is an HMM adaptation model denoted by "adapt". We partition all the speech recordings in the emphasis corpus into 6 parts according to the emphasis-related questions. We first train a neutral HMM model using all the data in the corpus and then adapt the parameters of the neutral HMM model with the data of every part separately. During synthesis stage, the speech segments of different emphasis categories are synthesized by the corresponding adapted HMM models.
2) The second model is the HMM model trained with the method of two-pass decision tree proposed by Yu [18] denoted by "two-pass-Yu". We first grow the main decision trees with emphasis-related questions and then expand the leaf nodes of the main decision tree with non-emphasis-related questions.
3) The third model is the method proposed in this paper, denoted by "hierarchical".

We perform three experiments to compare the performance of the models. The first experiment is an objective experiment which evaluates the prediction accuracy of the models. And the other two experiments are the subjective experiments which evaluate the emphasis intensity and the naturalness of the emphatic speeches synthesized by the HMM models.

For subjective evaluations, we invite 10 participants. All of them are Ph.D. or master students in Tsinghua University or the Chinese University of Hong Kong.

4.2 Experiment on the prediction accuracy of the emphatic speech synthesis models

This experiment is designed to compare the prediction accuracy of the models. 10 texts from testing set are selected and provided to the three models. The features (mean f0 and duration) of the phones of the emphasized words in the 30 synthesized sentences are compared with those in the corresponding emphatic recordings.

Table 2 Prediction accuracy (%) of different emphatic speech synthesis model

Acoustic features		Adapt	Two-pass-Yu	Hierarchical
Emphasized word	mean f0	79	86	89
	duration	67	77	78
Non-emphasized word	mean f0	86	88	87
	duration	75	76	76

The prediction accuracy PA for a certain feature is calculated as:

$$PA = \left(1 - \frac{\sum_{i=1}^{N} \left(F_i^{'j} - F_i^j \right) / F_i^j}{N} \right) \times 100\% \tag{6}$$

where F_i^j is the value of feature j of the ith word of emphatic speech recordings, while $F_i^{'j}$ is the predicted value of feature j of the ith word. N is the number of the samples.

Table 2 shows the prediction accuracy of different emphatic speech synthesis model. For emphasized words, the PAs of mean f0 and duration of the model "adapt" are significantly lower than those of the other two models and the PAs of the model "hierarchical" are the highest. For non-emphasized words, the PAs of the models "adapt", "two-pass" and "hierarchical" are generally the same.

4.3 Experiment on the emphasis intensity of the synthesized speeches of the models

This experiment is designed to compare the ability of generating emphasis of the models. 5 prompts from training set and 5 prompts from testing set are provided to each system. Each prompt contains one or more emphasized word(s). The resulting 30 sentences, together with the raw text prompts without emphasis annotation, are presented to the subjects in random order. Each subject is asked to listen to the sentence and identify which word(s) are emphasized. The subject is also asked to indicate the confidence level of emphasis perceived for each of the identified emphasized word, based on five-point Likert scale:

Table 3 Evaluation of emphasis quality through an emphasis identification experiment (SC level: subjects' confidence level), as the subjects are asked to give confidence level for the identified emphasis word, no SC level for "False Negative"

Systems	Accuracy		False positive		False negative	
	Rate (%)	SC level	Rate (%)	SC level	Rate (%)	SC level
Adapt	80	3.5	3	3.2	20	–
Two-pass-Yu	96	3.6	1	3.0	4	–
Hierarchical	98	3.6	1	3.0	2	–

Table 4 Results of the experiment on the naturalness of the synthesized speeches

Models	MOS
Adapt	3.7
Two-pass-Yu	3.4
Hierarchical	3.8

'1' (unclear); '2' (slight emphasis); '3' (emphasis); '4' (strong emphasis) and '5' (exaggerated emphasis).

10 subjects participated in the experiment. Table 3 shows the results of the experiment, where "Accuracy" is the rate of correctly identified emphasized words, "False Positive" is the rate of neutral words that are falsely identified as emphasized, and "False Negative" is the rate of emphasized words that are not detected. The accuracy rates and the related confidence levels of the "two-pass-Yu" and the proposed "hierarchical" models are much higher than those of the model "adapt". The accuracy confidence level of the model "hierarchical" is equal to that of the model "two-pass-Yu". The "False Negative" rate of the model "hierarchical" is slightly lower than that of the model "two-pass-Yu". These indicate the proposed "hierarchical" model can synthesize emphatic speech with almost the same emphasis quality as the model "two-pass-Yu", and much higher than the "adapt" model.

4.4 Experiment of the naturalness of synthesized speech

This experiment is designed to evaluate the naturalness of the synthesized speeches of the models. Another 5 prompts from training test and another 5 prompts from testing set are provided to the three models. Each prompt contains one or more emphasized word(s). The 30 synthesized speeches together with the texts with emphasis annotations are presented to the subjects in random order. The subjects are asked to give a 5-scaled MOS score according to the naturalness of the speech.

10 subjects participated in the experiment. The average MOS scores of different models are shown in Table 4. The synthesized speeches of the model "two-pass-Yu" have the lowest MOS score, while those of the proposed "hierarchical" model have the highest MOS scores. This indicates that the generated parameters of the latter method are more accurate, making the synthesized speeches sound more natural.

5 Conclusions

This paper proposes a framework for emphatic speech synthesis based on hidden Markov model (HMM). The framework contains the general decision tree (DT) which is first grown using non-emphasis-related questions and then expanded using emphasis-related questions. The general DT maintains the naturalness of synthesized speech. To improve the ability in synthesizing emphasis out of training data, a HMM-selection method and a parameter compensation model are added in the framework. This paper analyzes the contrastive (neutral versus emphatic) speech recordings considering kinds of contexts. Based on the analysis, the HMM selection and the parameter compensation based on the cost function considering the locations of the phones in relation with the emphasized words and the phone types are proposed. Experimental results show that the proposed method can synthesize emphatic speech with both high naturalness and high emphasis intensity.

Acknowledgments This work is supported by the National Basic Research Program of China (2012CB316401 and 2013CB329304). This work is also partially supported by the Hong Kong SAR Government's Research Grants Council (N-CUHK414/09), the National Natural Science Foundation of China (61375027, 61370023 and 60805008), the National Social Science Foundation (Major Project) (13&ZD189), and Guangdong Provincial Science and Technology Program (2012A011100008). The authors would like to thank the students of the research group of Human Computer Speech Interaction in Tsinghua University, the Graduate School at Shenzhen of Tsinghua University and the Chinese University of Hong Kong, for their cooperation with the dataset setup and experiments.

References

1. Boersma P, Weenink D (2003) Praat: doing phonetics by computer, http://www.praat.org
2. Cai LH, Huang DZ, Cai R (2003) Foundation and applications of modern speech technology. Press of Tsinghua University, Beijing
3. Kominek J, Black AW (2003) CMU ARCTIC databases for speech synthesis, Tech. Rep. CMU-LTI-03-177, Carnegie Mellon University
4. Leggetter CJ, Woodland PC (1995) Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Comput Speech Lang 9(2):171–185
5. Li AJ (1994) Duration charateristics of stress and its synthesis rules on standard Chinese, Report of phonetic research
6. Maeno Y, Nose T, Kobayashi T, Ijim Y, Nakajima H, Mizuno H, Yoshioka O (2011), HMM-based emphatic speech synthesis using unsupervised context labeling, In: Proc. Annual conference of international speech communication association (INTERSPEECH), 1849–1852
7. Meng H, Lo WK, Harrison AM, Lee P, Won KH, Leung WK, Meng FB (2011) 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)
8. Meng H, Lo YY, Wang L, Lau WY (2007) Deriving salient learners' mispronunciations from cross-language phonological comparisons, In: Proc. IEEE workshop on automatic speech recognition and understanding (ASRU)
9. Meng FB, Wu ZY, Jia J, Meng H, Cai LH (2013) Synthesizing english emphatic speech for multimodal corrective feedback in computer-aided pronunciation training. Multimed Tools Appl. doi:10.1007/s11042-013-1601-y
10. Morizane K, Nakamura K, Toda T, Saruwatari H, Shikano K (2009) Emphasized speech synthesis based on hidden Markov models, In: Proc. of speech database and assessments oriental COCOSDA Int. Conf., 76–81
11. Raux A, Black AW (2003) A unit selection approach to F0 modeling and its application to emphasis, In: Proc. IEEE workshop on automatic speech recognition and understanding (ASRU)
12. Shinoda K, Watanabe T (2000) MDL-based context-dependent subword modeling for speech recognition. Acoust Soc Japan (E) 21:79–86
13. Tokuda K, Zen H, Yamagishi J, Masuko T, Sako S, Black A, Nose T (2008) The HMM-based speech synthesis system (HTS) version 2.1, http://hts.sp.nitech.ac.jp/
14. Wu ZY, Meng H, Yang HW, Cai LH (2009) Modeling the expressivity of input text semantics for Chinese text-to-speech synthesis in a spoken dialog system. IEEE Trans Audio Speech Lang Process 17(8):1567–1577
15. Xu J (2009) Parametric analysis and synthesis for emotional speech, Doctoral dissertation, Tsinghua Unverisity
16. Xu Y, Xu CX (2005) Phonetic realization of focus in english declarative intonation. J Phon 33:159–197
17. Xydas G, Kouroupetroglou G (2006) Tone-group F0 selection for modeling focus prominence in small-footprint speech synthesis. Speech Comm 48(9):1057–1078
18. Yu K, Mairesse F, Young S (2010) Word-level emphasis modeling in HMM-based speech synthesis, In: Proc. of IEEE Int. Conf. on acoustics, speech, and signal processing. (ICASSP), 4238–4241
19. Zhu WB (2007) A Chinese speech synthesis system with capability of accent realizing. J Chin Inf Process 21(3):122–128

Zhiyong Wu received the B.S. and Ph.D. degrees in computer science and technology from Tsinghua University, Beijing, China, in 1999 and 2005, respectively.

He has been Postdoctoral Fellow in the Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong (CUHK) from 2005 to 2007. He joined the Graduate School at Shenzhen, Tsinghua University, Shenzhen, China, in 2007, where he is currently an Associate Professor. He is also with the Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems. His research interests are in the areas of multimodal multimedia processing and communication, more specially, audiovisual bimodal modeling, text-to-audio-visual-speech synthesis, and natural language understanding and generation.

Dr Wu is a member of the Technical Committee of Intelligent Systems Application under the IEEE Computational Intelligence Society and the International Speech Communication Association.

Yishuang Ning received the M.S. degree in computer science and technology from Beijing University of Technology, Beijing, China, in 2013. He is now a Ph.D. candidate in Tsinghua University.

His main research interests include emotional speech conversion and expressive speech synthesis.

Xiao Zang received the B.S. degree in computer science and technology from Beijing Institute of Technology, Beijing, China, in 2012. He is now a master student in Tsinghua University.

His main research interests include spoken dialogue system and natural language processing.

Jia Jia received the Ph.D. degree from Tsinghua University, Beijing, China, in 2008.

She is currently an Associate Professor with the Department of Computer Science and Technology, Tsinghua University. She is a member of Multimedia Committee of Chinese Graphics and Image Society. She has been awarded Scientific Progress Prizes from the Ministry of Education, China. Her current research interests include affective computing, and computational speech perception.

Fanbo Meng received the B.S. and Ph.D. degrees in computer science and technology from Tsinghua University, Beijing, China, in 2007 and 2014, respectively.

He has been awarded Scientific Progress Prizes from the Ministry of Education, China. His main research interests include emotional speech conversion, and expressive speech synthesis.

Helen Meng received the S.B., S.M., and Ph.D. degrees, all in electrical engineering, from the Massachusetts Institute of Technology (MIT), Cambridge.

She has been Research Scientist with the MIT Spoken Language Systems Group, where she worked on multilingual conversational systems. She joined The Chinese University of Hong Kong (CUHK) in 1998, where she is currently a Professor and Chairman in the Department of Systems Engineering and Engineering Management. In 1999, she established the Human–Computer Communications Laboratory at CUHK and serves as Director. In 2005, she established the Microsoft-CUHK Joint Laboratory for Human-Centric Computing and Interface Technologies, which was upgraded to MoE Key Laboratory in 2008, and serves as Co-Director. She is also Co-Director of the Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems. Her research interest is in the area of human–computer interaction via multimodal and multilingual spoken language systems, as well as translingual speech retrieval technologies.

Prof. Meng has been elected IEEE Fellow in 2013 and was Editor-in-Chief of the IEEE Transactions on Audio, Speech and Language Processing. She is also an elected board member of the International Speech Communication Association.

Lianhong Cai received the B.S. degree in computer science and technology from Tsinghua University, Beijing, China, in 1970.

She is currently a Professor with the Department of Computer Science and Technology, Tsinghua University. She was Co-Director of the Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems from 2006 to 2012, Director of the Institute of Human–Computer Interaction and Media Integration from 1999 to 2004. Her major research interests include human–computer speech interaction, speech synthesis, speech corpus development, and multimedia technology. She has undertaken 863 National High Technology Research and Development Program and National Natural Science Foundation of China projects.

Prof. Cai is a member of the Multimedia Committee of Chinese Graphics and Image Society and Chinese Acoustic Society