Multimed Tools Appl
DOI 10.1007/s11042-013-1601-y

Synthesizing English emphatic speech for multimodal corrective feedback in computer-aided pronunciation training

Fanbo Meng · Zhiyong Wu · Jia Jia · Helen Meng · Lianhong Cai

Abstract Emphasis plays an important role in expressive speech synthesis in highlighting the focus of an utterance to draw the attention of the listener. We present a hidden Markov model (HMM)-based emphatic speech synthesis model. The ultimate objective is to synthesize corrective feedback in a computer-aided pronunciation training (CAPT) system. We first analyze contrastive (neutral versus emphatic) speech recording. The changes of the acoustic features of emphasis at different prosody locations and the local prominences of emphasis are analyzed. Based on the analysis, we develop a perturbation model that predicts the changes of the acoustic features from neutral to emphatic speech with high accuracy. Further based on the

F. Meng · J. Jia (✉) · L. Cai
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
e-mail: jjia@tsinghua.edu.cn

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

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

F. Meng · J. Jia · L. Cai
Tsinghua National Laboratory for Information Science and Technology (TNList), Beijing, China

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

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

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

Z. Wu · J. Jia · H. Meng · L. Cai
Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China

Published online: 03 August 2013

perturbation model we develop an HMM-based emphatic speech synthesis model. Different from the previous work, the HMM model is trained with neutral corpus, but the context features and additional acoustic-feature-related features are used during the growing of the decision tree. Then the output of the perturbation model can be used to supervise the HMM model to synthesize emphatic speeches instead of applying the perturbation model at the backend of a neutral speech synthesis model directly. In this way, the demand of emphasis corpus is reduced and the speech quality decreased by speech modification algorithm is avoided. The experiments indicate that the proposed emphatic speech synthesis model improves the emphasis quality of synthesized speech while keeping a high degree of the naturalness.

Keywords Emphasis · Feature analysis · Emphatic speech perturbation · Emphatic speech synthesis · HMM

1 Introduction

Multimodal information processing plays an important role in computer-aided pronunciation training (CAPT), which uses speech technologies to facilitate pronunciation training for language learners. Pronunciation training should emphasize both perceptual training (i.e., developing the learner's skills to perceive and discriminate different sounds of the language) and productive training (i.e., training the learner's ability to produce speech and providing feedback on the learner's pronunciation). In this regard, multimodality plays a significant role in enhancing the learner's speech perception to assist with speech production.

It has been shown that the availability of corrective feedback in CAPT is very effective in reducing pronunciation errors [16]. Text-to-audio-visual speech (TTAVS) synthesis technologies have much to offer for the multimodal corrective feedback [5, 26]. For example, for the segments that are easily to be confused with other sounds, emphatic speech can be generated in both audio and visual modalities with the objective of highlighting such important segments to draw the attention of the learner. As of the two modalities, audio (or acoustic speech) serves more direct information for the pronunciation training in providing discriminations between different sounds of the language. This work targets the main communicative function of emphasis and tries to synthesize acoustic emphatic speech that could be integrated in the multimodal corrective feedback for CAPT.

State-of-the-art speech synthesis technologies can synthesize speech with high degree of naturalness. However effective human-computer interaction needs the generation of expressive speech to properly convey the message, e.g., synthesizing emphasis to highlight important words [18]. Emphasis is necessary for the expression of spoken language and emphasis synthesis can be useful in many human-computer interaction scenes, e.g., Computer-Aided Pronunciation Training system [13].

Emphasis is an important feature of prosody. It has been studied for a long time in phonetics. Many acoustic features, such as pitch variables (maximum, minimum, range and contour), intensity, speech rate and pause have already been analyzed [21, 22]. And it has been found that the acoustic features of the emphatic speeches are affected by many factors, such as the location of emphasis [9, 20, 28], the relationship between the acoustic features [17, 27], the intonation [11] and so on.

There are also some implementations in emphasis generation. With the framework of waveform concatenation synthesis, Li [8] analyzed the duration pattern of emphasis and proposed a rule-based emphasis synthesis model. Zhu [31] proposed an acoustic feature prediction model with decision tree and Gaussian mixture model to supervise the process of

unit selection. However, the emphasis quality and the speech quality of the synthesized speeches are restricted by the corpus in waveform concatenation synthesis. Some basic works [12, 15] added emphasis-related questions in traditional HMM framework [24] to synthesize emphasis. To improve the performance of the HMM-based emphasis synthesis system, Yu [29] proposed the methods of the two-pass decision tree and the factorial decision tree. Yu also proposed an HMM adaptation model, in which the corpus was partitioned into emphasis and non-emphasis regions and the former regions were used to adapt a neutral HMM model to the emphasis HMM model. As there are only a few words in a sentence, the data limitation remains one of the major problems for HMM-based emphatic speech synthesis. Some other emphatic speech synthesis systems are realized by adding emphatic speech perturbation models at the back-end of neutral speech synthesis systems [10, 30]. Bou-Ghazale [1, 2] used the linear prediction model and the hidden Markov model to model the differences between the features of the neutral speeches and the emphatic speeches. Li [9, 10] analyzed the acoustic features of emphasis at different prosody boundaries and built a rule-based linear modification model. The previous works on the perturbation model of emphasis haven't made full use of the contributions of the analysis of data, e.g., post-focus pitch suppression, decreasing the emphasis quality of the converted speeches. And the modification amplitude is oversize sometimes, decreasing the speech quality and the naturalness of the generated speeches.

This paper seeks to realize the emphasis synthesis with the HMM framework. Different from the previous work, we try to synthesize emphasis with the HMM model which is trained with neutral corpus. The basic idea is based on the local prominence characteristics of emphasis [23]. That is, syllables with f0, duration and energy greater than their neighboring syllables are likely to be emphasized, even if their values are not large on average. And further if a syllable is perceived as emphasis may be different when it is put in different contexts. Hence, emphasis could be generated from neutral corpus considering different contexts. The problem could be divided into two parts: 1) How to predict the acoustic features of the emphatic speeches from the texts with emphasis annotations? 2) How to supervise the HMM model to synthesize the speeches with the target acoustic features?

For the first sub-problem, we analyze the changes of the acoustic features from neutral speech to emphatic speech considering three factors separately: 1. The location of the syllables relative to emphasis; 2. The prosody location of emphasis; 3. The local prominences of the features of neutral speech. Based on the analysis, we develop a perturbation model which predicts the feature changes from neutral to emphatic speech. The training data including emphatic speeches and the corresponding neutral speeches are first clustered according to the locations of emphasis and the prosody locations of the syllables in the sentence. As the data of emphasis are much less than those of non-emphasis, to make full use of the data and avoid the data sparseness problem, a decision tree is adopted to cluster the data using the questions with the best discriminations. And then to solve the problem of the oversize prediction, the data of each leaf node are modeled considering the local prominence of emphasis and the relationship between the acoustic features to improve the prediction accuracy of the models. We use this perturbation model and a neutral speech synthesis HMM model to predict the parameters of emphatic speech. In relation to the second sub-problem, we discretize the acoustic features of the neutral corpus. New acoustic-feature-related labels and corresponding questions on the discretized acoustic features are added for HMM modeling. In the synthesis stage, the features predicted by the perturbation model are converted to the labels. And the labels are used to supervise the HMM model to synthesize the speeches with the features similar to the emphatic speeches. We hope to incorporate the proposed emphatic speech synthesis model in automatic feedback generation on a CAPT platform.

The rest of the paper is: Section 2 presents the corpora used for data analysis and model training. Section 3 does the feature analysis of English emphatic speech. Section 4 builds a perturbation model from neutral to emphatic speech based on the analysis. Section 5 gives details of the model for emphatic speech synthesis. Section 6 describes the perceptual evaluations of the outputs of the models. Finally, Section 7 lays out conclusions.

2 Corpora

2.1 Emphasis corpus with contrastive speech recordings

We design a set of 350 text prompts for recording the emphasis corpus. These text prompts are carefully designed by considering the factors affecting the expression of emphasis. Each text prompt may contain one or more emphasized words, with each emphasized word located at different positions in the sentences. These words may be mono- or polysyllabic, with the primary stressed syllables at different places. Furthermore, the phones with all kinds of pronunciation mechanisms are covered by the text prompts. The contexts of the phones are also covered as many as possible. One example text prompt is shown as follows (with emphasized words capitalized):

"*I have met **PETERSON** on one **OCCASION**.*"

Two contrastive speech utterances are recorded for each text prompt – one with neutral intonation throughout the utterance and the other with emphasis placed on the emphasized words. A female speaker with a high level of English proficiency is invited to record in a sound proof studio. Hence we have 700 recorded utterances, saved in the wav files (16 bit mono, sampled at 16 kHz).

2.2 Neutral corpus

To obtain well trained HMM models for generating speech with a high degree of naturalness, the CMU US ARCTIC clb corpus [6] with neutral speech recordings is used as the neutral corpus. It has 1132 phonetically balanced utterances recorded by an US female speaker, stored in the 16bit mono format as wav files with 16 kHz sampling rate.

2.3 Data preprocessing

Both emphasis corpus and neutral corpus are automatically annotated by FestVox [4]. 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. The fundamental frequencies (i.e., f0s) of the corpora are extracted by STRAIGHT [23]. To ensure the accuracy of data analysis, the f0s of contrastive speech recordings of the emphasis corpus are manually checked and corrected before data analysis.

3 Acoustic analysis of emphasis

This section provides the analysis of acoustic correlations of emphasis based on the contrastive speech recordings of the emphasis corpus. To perform the analysis, we first extract seven acoustic features related to fundamental frequency, intensity and speaking rate.

Detailed analysis is then provided about the correlations between the acoustic feature variations from neutral to emphatic speech and three kinds of contexts, 1) the location of the syllables in relation with the stressed syllables in emphasized word, 2) the position of the syllables in prosody phrase and word, and 3) the local prominences of the features in neutral speech.

3.1 Extraction of acoustic features

The objective is to analyze how emphatic words are realized in acoustic speech signal. Acoustic features that are commonly associated with emphasis include fundamental frequency (f0), intensity and speaking rate. Hence we choose to extract the following acoustic features to capture the acoustic correlations of emphasis:

- maximum f0 (P_{Max}, in Hz),
- minimum f0 (P_{Min}, in Hz),
- f0 range (P_{Range}, in Hz),
- mean f0 (P_{Mean}, in Hz),
- absolute value of f0 slope (P_{S}, in Hz/ms),
- duration per phone (D, in ms), and
- mean of RMS energy (E, in dB).

Measurements are taken from the contrastive recordings (neutral versus expressive) of each prompt in the emphasis corpus for the above acoustic features.

We first compute the ratio (in %) between the measurements of the corresponding emphasized and neutral syllable units, and the variances of the ratios. Let $F_{i,\mathrm{neu}}$ be the measurement of a certain feature of syllable i of the neutral speech recording, and $F_{i,\mathrm{emp}}$ be the measurement of the feature of the corresponding syllable of the emphatic speech recording. Let n be the number of the syllables. The change ratio ΔF of the acoustic feature F is then calculated as:

$$\Delta F = \frac{1}{n}\sum_{i}\frac{F_{i,emp}}{F_{i,\mathrm{neu}}} \tag{1}$$

Additionally, we also compute the local prominence (LP) of the acoustic feature for a particular syllable in the neutral speech recording. The LP of a particular syllable is defined as the ratio (in %) between the acoustic measurements of that syllable and the average acoustic measurements of all the syllables in the same prosody phrase. Let $F_{i,\mathrm{neu}}$ be the measurement of a certain feature of syllable i of the neutral speech recording, and syllables $i_1 \ldots i_k$ are in the same prosody phrase to which syllable i belongs. Let $\widehat{F}_{i,\mathrm{neu}}$ be the LP of the feature F for syllable i. $\widehat{F}_{i,\mathrm{neu}}$ is calculated as:

$$\widehat{F}_{i,neu} = \frac{F_{i,neu}}{\frac{1}{k}\sum_{j\in\{i_1\ldots i_k\}} F_{j,neu}} \tag{2}$$

3.2 Acoustic analysis of emphasis for the syllables at different locations in relation with stressed syllables in emphasized word

In emphatic speech, emphasized words will often effect the changes of the acoustic features of their neighboring words. For example, the speaker tends to decrease the f0s of the post-

emphasized words [3]. In this section, we classify the syllables into six classes based on the location of the syllable in relation with the nearest emphasized word and its stressed syllables:

- Class 1: the **Primary** stressed syllable of an **Emphasized** word (denoted by P-E)
- Class 2: syllables **Before** the **Primary** stressed syllable of an **Emphasized** word (denoted by B-P-E)
- Class 3: syllables **After** the **Primary** stressed syllable of an **Emphasized** word (denoted by A-P-E)
- Class 4: syllables in the **Neutral** word **Before** the emphasized word (denoted by N-B)
- Class 5: syllables in the **Neutral** word **After** the emphasized word (denoted by N-A)
- Class 6: all other (**Remaining**) syllables (denoted by R).

A syllable is assigned the class with the lowest class number if it falls into more than one class. Figure 1 illustrates this method of syllable classification. "PETERSON" and "OCCASION" are the emphasized words in the sentence.

Table 1 shows the changes of the acoustic features from neutral speech to emphatic speech for the above different class of syllables. For each syllable class, the first row shows the change ratios of the acoustic features from neutral speech to emphatic counterpart, and the second row shows the variances of the change ratios.

For the primary stressed syllables of the emphasized words (P-E), the maximum f0 increases substantially. However, the f0 minimum and energy remain largely the same. The slope and duration both increase substantially.

For the syllables before the primary stressed syllables of the emphasized words (B-P-E), the f0 maximum and the slope decrease. The energy stays largely the same. And the duration increase much. Because most syllables of B-P-E are unstressed syllables, e.g., the first syllable of the word "apartment", the speaker tends to reduce the f0 and increase the duration to highlight the latter stressed syllables.

For the syllables after the primary stressed syllables of an emphatic word (A-P-E), almost all the features increase, especially for f0 maximum, f0 range, f0 slope and duration.

The features of the syllables of the words before and after the emphatic words (N-B and N-A) don't change much. The only difference is that the f0 of N-B is a bit higher than that of N-A caused by post-pitch suppression.

For the syllables of all other words (R), the f0 increases slightly, while the f0 slope decrease much, leading to the f0 envelope plat and the speech sounds plain to highlight the emphasis.

3.3 Acoustic analysis of emphasis at different prosody positions

The measurements of the acoustic features of the syllables at different prosody positions in the neutral speech recordings are different. For example, there will be duration lengthening for the last syllables of the prosody phrases. Pitch resets are also recognized at prosody phrase boundaries. Furthermore, the changes of the acoustic features from neutral to emphatic speech are also different at different prosody positions. In this section, we classify the syllables of emphasized words into 3×3 classes according to their prosody positions at prosody phrase and word layer.

I	have	met	PE	TERSON	on	one	OC	CASION.
6	4	1	3		5	4	2	1 3

Fig. 1 An example of syllable classification based on the location of stressed syllables in emphasize words

Table 1 Changes of the acoustic features from neutral to emphatic speech for the syllables at different locations in relation with the nearest emphasized word and its stressed syllables, where Ratio (%) denotes the change ratio of the acoustic feature between emphatic and neutral speech, and Var denotes the variances of the change ratios

		ΔP_{Max}	ΔP_{Min}	$\Delta P_{\mathrm{Range}}$	ΔP_{Mean}	ΔP_{S}	ΔE	ΔD
P-E	Ratio(%)	111	97	271	103	350	104	150
	Var	0.02	0.02	5.12	0.01	91.24	0.00	0.13
B-P-E	Ratio(%)	95	98	229	96	92	102	153
	Var	0.32	0.04	38.70	0.03	39.09	0.01	0.39
A-P-E	Ratio(%)	108	104	284	104	228	104	118
	Var	0.04	0.04	18.49	0.03	34.84	0.00	0.86
N-B	Ratio(%)	99	96	144	98	109	101	111
	Var	0.02	0.02	16.11	0.03	34.88	0.00	0.44
N-A	Ratio(%)	96	95	101	95	99	100	109
	Var	0.04	0.02	17.23	0.01	22.89	0.01	0.94
R	Ratio(%)	97	96	138	96	179	100	103
	Var	0.05	0.02	19.33	0.03	28.95	0.01	0.85

At prosody phrase layer:

- Class 1: The syllables are in the **First** prosody **Phrase** in the sentence (FP).
- Class 2: The syllables are in the prosody **Phrase** in the **Middle** of the sentence (MP).
- Class 3: The syllables are in the **Last** prosody **Phrase** in the sentence (LP).

At word layer:

- Class 1: The syllables are in the **First Word** in the prosody phrase (FW).
- Class 2: The syllables are in the **Word** in the **Middle** of the prosody phrase (MW).
- Class 3: The syllables are in the **Last Word** in the prosody phrase (LW).

We use "$L_{\mathrm{Phrase}}\_L_{\mathrm{Word}}$" to represent the classes of syllables. For instance, the class "LP-FW" means the syllables are in the first word of the last prosody phrase.

Table 2 shows the changes of the acoustic features from neutral to emphatic speech and their corresponding variances for the **stressed** syllables of the emphasized words.

The feature changes of P_{s} and P_{Range} increase in all cases, but the variances are very large. For P_{Max}, P_{Min} and P_{Mean}, the feature changes become larger when the syllables are more close to the end of the phrase. For instance, the changes of the P_{Mean} of the syllables in class "FP-MW" (syllables in the word in the middle of the first prosody phrase) are larger than those of the syllables in class "FP-FW" (syllables in the first word of the first prosody phrase), but lower than those of the syllables in class "FP-LW" (syllables in the last word of the first prosody phrase). This is mainly due the pitch declination. The closer the syllables are to the end of the phrase, the smaller the f0s of the syllables are. Hence, to realize emphasis close to the end of the prosody phrase, the speaker has to increase f0s more.

In addition, there are no significant differences between the features changes of the syllables in the same word locations of different phrase. For instance, the changes of the mean f0s of the syllables in class "FP-FW" (syllables in the first word of the first prosody phrase) are similar to those of the syllables in class "LP-FW" (syllables in the first word of the last prosody phrase). This is because the f0s are reset at the boundaries of the prosody phrases and the perception of emphasis is mainly due to the feature differences between the

Table 2 Changes of the acoustic features from neutral to emphatic speech and their corresponding variances for the **stressed** syllables of emphasized words

		ΔP_{Max}	ΔP_{Min}	ΔP_{Range}	ΔP_{Mean}	ΔP_S	ΔD	ΔE
FP-FW	Ratio(%)	102	90	157	103	160	165	104
	Var	0.00	0.04	1.63	0.02	2.90	0.21	0.01
FP-MW	Ratio(%)	109	93	194	104	343	151	109
	Var	0.03	0.05	2.80	0.03	81.44	1.05	0.01
FP-LW	Ratio(%)	111	102	165	111	298	131	112
	Var	0.03	0.04	2.10	0.02	69.17	0.56	0.01
MP-FW	Ratio(%)	106	94	284	102	577	162	102
	Var	0.00	0.07	3.43	0.02	21.68	0.04	0.00
MP-MW	Ratio(%)	113	100	214	109	265	129	105
	Var	0.03	0.07	5.12	0.04	13.06	1.20	0.01
MP-LW	Ratio(%)	117	106	178	114	354	110	105
	Var	0.02	0.03	1.61	0.02	96.72	0.21	0.01
LP-FW	Ratio(%)	105	100	174	105	230	183	97
	Var	0.02	0.13	1.23	0.02	6.89	1.69	0.01
LP-MW	Ratio(%)	110	95	201	106	411	167	101
	Var	0.03	0.04	2.76	0.02	104.11	0.92	0.01
LP-LW	Ratio(%)	111	107	144	110	312	125	107
	Var	0.02	0.05	0.86	0.02	74.51	0.36	0.01

emphasized words and the nearby words. Hence, feature changes of the syllables in different prosody phrases are similar.

The changes of duration show opposite pattern compared to f0. This is because the f0 cannot increase unlimitedly due to the physical limitation. When the intrinsic f0s are high, the speaker could not increase f0 much and the speaker tends to increase the durations for emphasis generation. The changes of energy are similar to those of f0.

Table 3 shows the changes of the acoustic features from neutral to emphatic speech and their corresponding variances for the **unstressed** syllables of the emphasized words.

The feature changes of unstressed syllables of emphasized words from neutral to emphatic speech are similar to those of the stressed syllables. It should be noted that as the intrinsic features of stressed syllables are higher than those of unstressed syllables, similar feature changes will make the differences between the features of stressed syllables and the features of unstressed syllables increase.

3.4 Correlation analysis of feature changes of emphasized words and local prominences of neutral speech

The f0s of the syllables in neutral speech are not only affected by their prosody positions, but also the intrinsic f0 of the voiced phones, the intonations and so on. In this section, we focus on the local prominence (*LP*) of the syllables in neutral speech, and try to analyze the pattern of the influences of different *LP*s of the features on the changes of the acoustic features of emphasis from neutral to emphatic speech.

Table 4 shows the correlations between the changes of the acoustic features of the stressed syllables of the emphasized words from neutral to emphatic speech and the *LP*s of the features

Table 3 Changes of the acoustic features from neutral to emphatic speech and their corresponding variances for the **unstressed** syllables of emphasized words

		P_{Max}	P_{Min}	P_{Range}	P_{Mean}	P_{S}	D	E
FP-FW	Ratio(%)	99	91	134	97	160	144	99
	Var	0.01	0.01	2.23	0.02	5.70	0.41	0.02
FP-MW	Ratio(%)	102	96	206	100	323	130	104
	Var	0.05	0.06	23.17	0.05	59.31	0.72	0.01
FP-LW	Ratio(%)	104	101	149	103	160	131	105
	Var	0.05	0.08	1.40	0.06	2.92	0.74	0.01
MP-FW	Ratio(%)	104	98	146	101	170	128	107
	Var	0.00	0.00	0.48	0.00	1.71	0.02	0.00
MP-MW	Ratio(%)	111	105	141	107	280	118	105
	Var	0.03	0.04	1.03	0.03	52.40	1.09	0.01
MP-LW	Ratio(%)	116	111	189	114	316	117	109
	Var	0.10	0.36	37.54	0.27	28.71	0.66	0.01
LP-FW	Ratio(%)	104	99	113	102	94	114	104
	Var	0.00	0.00	0.12	0.00	0.18	0.05	0.00
LP-MW	Ratio(%)	111	109	157	110	253	137	106
	Var	0.04	0.07	49.41	0.06	25.68	1.16	0.01
LP-LW	Ratio(%)	113	111	177	113	222	128	108
	Var	0.10	0.11	2.91	0.10	9.76	0.30	0.01

of neutral speech. For a certain feature, the feature change has negative correlation with the corresponding *LP*. For instance, the correlation between ΔP_{Max} and $\widehat{P}_{\mathrm{Max}}$ is −0.71. This indicates that the changes of the acoustic features from neutral to emphatic speech are negatively correlated to the *LP*s of the features in the neutral speech. The higher the *LP*s are, the lower the feature changes are. Besides, the changes of a certain feature are also correlative to the *LP*s of other features. For example, ΔP_{Max}, ΔP_{Min} and ΔP_{Mean} have positive correlations with $\widehat{D}$, while ΔD have positive correlations with $\widehat{P}_{\mathrm{Max}}$, $\widehat{P}_{\mathrm{Max}}$ and $\widehat{P}_{\mathrm{Mean}}$. This indicates that when the f0s of neutral speech are high, the speaker tends to increase the durations more to generate emphasis, which is consistent to the analysis in Section 3.3.

Table 5 shows the correlations between the changes of the acoustic features of the unstressed syllables of the emphasized words from neutral to emphatic speech and the *LP*s of the features of neutral speech. The correlation pattern of unstressed syllables is similar to that of stressed syllables. The main difference is that ΔP_{Max}, ΔP_{Min} and ΔP_{Mean} have no significant correlation with $\widehat{D}$. This is because the intrinsic f0s of unstressed syllables are low, and the speaker could increase f0s as required by generating emphasis and do not need to increase the durations additionally.

The *LP*s are important to the perception of emphasis. The *LP*s of the features of the emphatic speech could be calculated according to the changes of the acoustic features from neutral to emphatic speech and the *LP*s of the features of the neutral speech. Hence, the *LP*s of the features of the neutral speech should be involved in the perturbation model from neutral to emphatic speech. For example, the neutral speech is "there is a star in the bar." The *LP*s of the acoustic features of "star" are higher than those of "in" in the neutral speech. The modification amplitude of the acoustic features of "star" to be emphasized will be higher than those of "in" to be emphasized.

Table 4 Correlations between the changes of the acoustic features (ΔF) of the **stressed** syllables of emphasized words from neutral to emphatic speech and the local prominences (LP, $\widehat{F}$) of the features of neutral speech, where F is the measurement of a certain acoustic feature in Section 3.1

	$\widehat{P}_{\text{Max}}$	$\widehat{P}_{\text{Min}}$	$\widehat{P}_{\text{Range}}$	$\widehat{P}_{\text{Mean}}$	$\widehat{P}_{\text{S}}$	$\widehat{D}$	$\widehat{E}$
ΔP_{Max}	−0.71	−0.69	0.36	−0.69	0.17	0.58	−0.15
ΔP_{Min}	−0.86	−0.87	0.47	−0.91	0.08	0.47	−0.38
ΔP_{Range}	0.14	0.59	−0.92	0.46	−0.80	−0.64	0.98
ΔP_{Mean}	−0.76	−0.85	0.59	−0.85	0.26	0.65	−0.49
ΔP_{S}	−0.27	0.19	−0.73	0.04	−0.74	−0.27	0.80
ΔD	0.65	0.65	−0.37	0.65	−0.11	−0.64	0.24
ΔE	−0.44	−0.52	0.38	−0.48	0.17	0.74	−0.30

4 Decision tree based perturbation model from neutral to emphatic speech

Based on the above acoustic analysis, a perturbation model based on decision tree is proposed in this section. This perturbation model captures the above correlations between the acoustic feature variations from neutral to emphatic speech and the three contexts, and can be used to generate acoustic features for emphatic speech synthesis given the contexts and the acoustic features of the neutral speech.

4.1 Feature selection for modeling

We observe that the variances of acoustic feature P_{Range} and P_{S} are approximately 100 times of the other features in Tables 1, 2, and 3. These two features are not stable. Regardless of these two features, feature P_{Max} changes the most at emphasized words. Hence, we choose P_{Max} and P_{Min} to control f0 range. In addition, feature D and E are also chosen for modeling.

4.2 Decision tree for feature clustering

As analyzed in Section 3, the changes of the acoustic features from neutral to emphatic speech are affected by different contexts. Data clustering is necessary to improve the accuracy of the model to predict the values of the acoustic features for emphatic speech from neutral speech. Decision tree provides an efficient way to associate the contexts with

Table 5 Correlations between the changes of the acoustic features (ΔF) of the **unstressed** syllables of emphasized words from neutral to emphatic speech and the local prominences (LP, $\widehat{F}$) of the features of neutral speech

	$\widehat{P}_{\text{Max}}$	$\widehat{P}_{\text{Min}}$	$\widehat{P}_{\text{Range}}$	$\widehat{P}_{\text{Mean}}$	$\widehat{P}_{\text{S}}$	$\widehat{D}$	$\widehat{E}$
ΔP_{Max}	−0.56	−0.66	0.16	−0.62	−0.14	0.03	−0.48
ΔP_{Min}	−0.62	−0.71	0.14	−0.68	−0.18	0.06	−0.49
ΔP_{Range}	−0.33	−0.52	0.44	−0.45	−0.32	−0.11	−0.21
ΔP_{Mean}	−0.59	−0.68	0.18	−0.64	−0.09	0.00	−0.55
ΔP_{S}	−0.11	−0.32	0.55	−0.26	−0.16	0.05	−0.02
ΔD	0.57	0.36	0.62	0.46	0.12	−0.60	0.59
ΔE	−0.82	−0.82	−0.14	−0.84	−0.61	−0.03	−0.43

clusters and can select most discriminative context questions to split data clusters. Hence, decision tree is used in this work for data clustering.

We design 12 questions for decision-tree-based data clustering. These questions are classified into four classes and there are three questions for each class, as shown in Table 6. As there are only a few emphasized words in a sentence, the data of emphatic speech are much less than those of non-emphatic in the training data of the emphasis corpus. Due to this reason, the discriminative powers of emphasis-related questions are lower than those of non-emphasis-related questions. To avoid clustering the data of emphatic speech and those of non-emphatic into the same leaf node, the emphasis-related questions are used prior to the non-emphasis-related questions.

The distance measurement used for decision tree in splitting nodes is represented by the average Euclidean distance between all the data in the node and the center of the data in the node. Let $\mathbf{V}_i$ be the feature vector composited of the changes of the features of syllable i from neutral to emphatic speech and the *LPs* of the features of the corresponding syllable of the neutral speech. $\mathbf{V}_i$ is represented as:

$$\mathbf{V}_i = \left[\begin{array}{ccccccc} \Delta P_{\mathrm{Max},i} & \Delta P_{\mathrm{Min},i} & \Delta D_i & \Delta E_i & \widehat{P}_{\mathrm{Max},i} & \widehat{P}_{\mathrm{Min},i} & \widehat{D}_i \quad \widehat{E}_i \end{array} \right] \tag{3}$$

Let $\mathbf{L}$ be the current node, and the syllable indices in the node be $l_1,l_2,\ldots,l_n$. n is the number of syllables in the current node. Then the distance of node $\mathbf{L}$ is calculated as:

$$d(\mathbf{L}) = \frac{1}{n}\sum_{i=l_1}^{l_n} f\left(\mathbf{V}_i, \frac{1}{n}\sum_{j=l_1}^{l_n}\mathbf{V}_j\right) \tag{4}$$

where f(.) denotes the algorithm of Euclidean distance. Let $\mathbf{Q}$ be the question set used for decision tree clustering. Let $\mathbf{L}_{ql}$ and $\mathbf{L}_{qr}$ be the sub-nodes of the node $\mathbf{L}$ split by question q. The question q_0 which decreases the distance the most is then used to split the current node:

$$\begin{aligned} \Delta d_q &= \left\{ d\left(\mathbf{L}_{ql}\right) + d\left(\mathbf{L}_{qr}\right) - 2d(\mathbf{L}) \right\} \\ q_0 &= \arg\min_{q\in\mathbf{Q}} \left(\Delta d_q\right) \end{aligned} \tag{5}$$

Two conditions are used to stop the data clustering process: 1) there aren't any questions which could decrease the distance; or 2) the number of the data in the current node is below a threshold value. Figure 2 shows the top part of the decision tree for feature clustering. The

Table 6 The question set for growing decision tree

Question classes	Questions	Answers
The questions about the relative positions between the current word and the emphasized words	If the current word is emphasized word/ before emphasized word/after emphasized word?	Yes/no
The questions about the relative positions between the current syllables and the stressed syllables within the same word	If the current syllable is stressed syllable/ before stressed syllable/after stressed syllable?	Yes/no
The questions about the positions of the current word (where the current syllables located) in the phrase	If the current word is the first word/the middle word/the last word in the prosody phrase?	Yes/no
The questions about the positions of the current phrase (where the current syllables located) in the sentence	If the current phrase is the first phrase/the middle phrase/the last phrase in the sentence?	Yes/no

questions about the relative positions between the current word and the emphasized words are firstly used to cluster the data and then other questions related to the prosody positions of pronunciation units are used for further splitting nodes.

4.3 Linear perturbation model of the changes of the acoustic features from neutral to emphatic speech

As the questions with the minimum distances are used during data clustering, the feature vectors in the same leaf node are similar to each other. A basic prediction method is to calculate the average values of the feature changes in the same leaf node, and use the average values as the prediction results from the contexts of the leaf node. But without considering the *LPs* of the features of the neutral speeches, the emphasis quality and the naturalness of the generated speeches will decrease as has been detailed in Section 3.4. To improve the prediction accuracy of the model, we assume that there are linear relations between the changes of the acoustic features from neutral to emphatic and the *LPs* of the features of neutral speech. Then we have:

$$\begin{aligned} \Delta P_{\mathrm{Max},i} &= a_1 \widehat{P}_{\mathrm{Max},i} + b_1 \\ \Delta P_{\mathrm{Min},i} &= a_2 \widehat{P}_{\mathrm{Min},i} + b_2 \\ \Delta D_i &= a_3 \widehat{D}_i + b_3 \\ \Delta E_i &= a_4 \widehat{E}_i + b_4 \end{aligned} \tag{6}$$

As described in Section 3.4, the changes of a certain feature may be affected by different *LP*s of other features. Considering such correlations between the changes of different acoustic features, the above formula can be extended as:

$$\begin{matrix} \mathbf{R} \\ \begin{bmatrix} \Delta P_{\mathrm{Max},i} \\ \Delta P_{\mathrm{Min},i} \\ \Delta D_i \\ \Delta E_i \end{bmatrix} \end{matrix} = \begin{matrix} \mathbf{A} \\ \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \end{matrix} \begin{matrix} \mathbf{T} \\ \begin{bmatrix} \widehat{P}_{\mathrm{Max},i} \\ \widehat{P}_{\mathrm{Min},i} \\ \widehat{D}_i \\ \widehat{E}_i \end{bmatrix} \end{matrix} + \begin{matrix} \mathbf{B} \\ \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} \end{matrix} \tag{7}$$

where $a_{ii}(i=1,2,3,4)$ represent the relations between the feature changes from neutral to emphatic speech and the *LP* of the corresponding features of neutral speech, and $a_{ij}(i \neq j)$ represent the relations between the features changes from neutral to emphatic speech and the *LPs* of other features of neutral speech. When **A** is equal to identity matrix and **B** is the average vector of the changes of the acoustic features in the leaf node, the model degenerates to the statistics model which uses the average values as the predicted values.

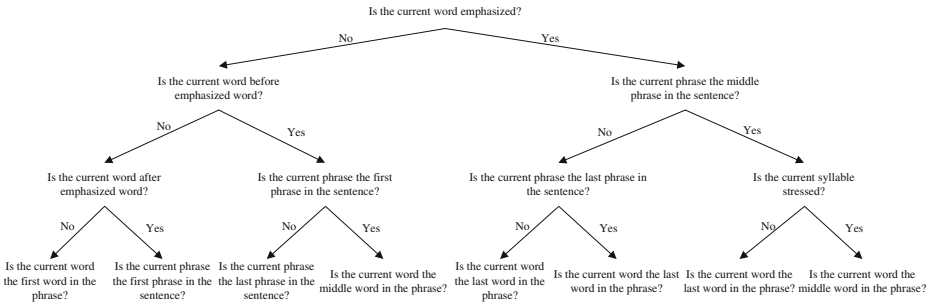

Fig. 2 The top part of the decision tree or feature clustering

In our work, nonlinear least squares regression is used to estimate the parameter matrixes **A** and **B** for each leaf node of the decision tree.

4.4 Realization of the perturbation model

The inputs of the perturbation model are the measurements of the acoustic features (maximum f0, minimum f0, duration and energy) of the neutral speech, and the outputs are the measurements of corresponding acoustic features of the emphatic speech. During generation process, text analysis is first performed and the parameter matrixes **A** and **B** of the syllables are got from the decision tree according to the contexts. At the meantime, the acoustic features of the neutral speech are extracted and the *LP*s are calculated. Then the changes of the acoustic features could be calculated according to the *LP*s and the parameter matrixes. The f0s and durations of the neutral speeches are modified by the perturbation model and the energies are adjusted and smoothed by Hamming window.

Assume that there are N syllables in the neutral speech. Let $\mathbf{P}_i(n)$, $\mathbf{E}_i(n)$ and $\mathbf{D}_i(n)$ be the f0 vector, energy vector and the corresponding time vector of the ith syllable, which begins at time step b_i and ends at time step e_i.

The *LP*s of the neutral speech are calculated according to the result of the text analysis and the extracted features using the formula (2). Then the changes of the acoustic features ($\Delta P_{\mathrm{Max},i}$, $\Delta P_{\mathrm{Min},i}$, ΔD_i and ΔE_i) of the syllables could be calculated with the *LP*s and the parameter matrixes **A** and **B** using formula (7).

Predicting the f0s and the durations: The target f0 vector $\mathbf{P}_i^{'}(n)$ are calculated as follows:

$$P_{\mathrm{Min},i}^{'} = P_{\mathrm{Min},i} \times \Delta P_{\mathrm{Min},i} \tag{8}$$

$$P_{\mathrm{Max},i}^{'} = P_{\mathrm{Max},i} \times \Delta P_{\mathrm{Max},i} \tag{9}$$

$$\mathbf{P}_i^{'}(n) = P_{\mathrm{Min},i}^{'} + \frac{P_{\mathrm{Max},i}^{'} - P_{\mathrm{Min},i}^{'}}{P_{\mathrm{Max},i} - P_{\mathrm{Min},i}} \times \left(\mathbf{P}_i(n) - P_{\mathrm{Min},i}\right), n \in [b_i, e_i] \tag{10}$$

$$\mathbf{D}_i^{'}(n) = b_i + (\mathbf{D}_i(n) - b_i) \times \Delta D_i, n \in [b_i, e_i] \tag{11}$$

Predicting the energies: the energy vector of $\mathbf{E}_i(n)$ are adjusted with ΔE_i and smoothed by Hamming window $\mathbf{H}_{i,k}(n)$ of which window length is L, window shift is $M/2$.

$$\mathbf{E}_{i,k}^{'}(n) = \mathbf{E}_i(n)\mathbf{H}_{i,k}(n)\Delta E_i, k \in \left[0, 2\left\lfloor \frac{e_i - b_i}{M} \right\rfloor\right] \tag{12}$$

$$\mathbf{H}_{i,k}(n) = \begin{cases} 0.54 - 0.46\cos\left(\dfrac{2\pi\left(n - b_i - {}^{kM}/_{2}\right)}{e_i - b_i}\right), n \in \left[b_i + {}^{kM}/_{2}, b_i + \left({}^{k}/_{2} + 1\right)M\right] \\ 0, n \notin \left[b_i + {}^{kM}/_{2}, b_i + \left({}^{k}/_{2} + 1\right)M\right] \end{cases} \tag{13}$$

$$\mathbf{E}_i^{''}(n) = \sum_{k=0}^{2\left\lfloor \frac{e_i - b_i}{M} \right\rfloor} \mathbf{E}_{i,k}^{'}(n), n \in [b_i, e_i] \tag{14}$$

5 Two-stage HMM-based English emphatic speech synthesis

In Section 3, we analyze the correlations between the changes of the acoustic features from neutral to emphatic speech and three different kinds of contexts. Based on the analysis, a decision tree based perturbation model is proposed in Section 4 that can be used to generate acoustic features for emphatic speech given the contexts and acoustic features of neutral speech. In this section, we propose a two-stage HMM-based emphatic speech synthesis method to ensure both the naturalness and quality of the synthetic emphatic speech.

5.1 Two-stage HMM-based emphatic speech synthesis

There are two typical methods for emphatic speech synthesis. One method is to synthesize a neutral speech first followed by modifying the acoustic features of the synthetic neutral speech according to the perturbation model, and then to regenerate the emphatic speech according to the perturbed acoustic features. The problem of this method is that the speech quality may degrade a lot when the perturbation ratios are larger than a threshold. The other method is parametric speech synthesis, for example, using HMM. However in the case of emphatic speech synthesis, as there are only a few emphasized words in one sentence, the data for emphatic speech are much less than those for non-emphatic. It remains the biggest problem on how to derive a well-train HMM with very limited amount of emphatic speech data for HMM-based emphatic speech synthesis.

Considering the above issues, a two-stage HMM-based emphatic speech synthesis method is proposed, as shown in Fig. 3. Different from the previous work, two HMMs are involved in our work at different stages. The first HMM is called the neutral HMM model (N-HMM), and the second HMM is called the emphatic HMM model (E-HMM). The two HMMs will take effect sequentially at two stages in our method to synthesize emphatic speech.

During training, the N-HMM is trained from the neutral speech data (i.e., the speech recordings from both neutral corpus and emphasis corpus) with the standard context questions (details will be elaborated in Section 5.2). The trained N-HMM is then used to synthesize neutral utterances for all the text prompts from both neutral and emphasis corpora. The synthetic neutral speech utterances from the text prompts of emphasis corpus, together with the emphasis speech recordings of the emphasis corpus, form a new pseudo corpus and are used to train the perturbation model (details elaborated in Section 5.3). Instead of modifying the synthetic neutral speech with the perturbation model directly, a new E-HMM is trained from the synthetic neutral speeches with additional acoustic-feature-related labels for decision tree growing. The additional acoustic-feature-related labels are extended by the prediction of the perturbation model. These extended labels ensure the variations of the desired acoustic features of the emphatic speech are captured by the E-HMM (details in Section 5.4).

During synthesis, three steps are involved to predict the acoustic features of the emphatic speech. In the first step, the input text is first converted to emphasis-related labels and non-emphasis-related labels by text analysis module. The latter are provided to the N-HMM model to predict the acoustic features of the neutral speech. In the second step, the acoustic features of the neutral speech and the emphasis-related labels are then provided to the perturbation model to predict the acoustic features of the emphatic speech. The acoustic features of the emphatic speech are then descretized and converted to the additional acoustic-feature-related labels and added to the non-emphasis-relate labels. Details on how perturbation model functions can be found in Section 4.4. Finally, the E-HMM model is used to predict the acoustic features according to the new extended labels. These predicted acoustic features are generated from the E-HMM models trained from the large amount of intermediate synthetic speech, and can

ensure the naturalness of the synthetic result. Furthermore, the values of predicted acoustic features are also similar to those of the target emphatic speech, due to the introduction of the additional acoustic-feature-related labels. In this way, the quality of the emphasis can also be affirmed. Finally the emphatic speech is synthesized with the predicted parameters.

5.2 Training the neutral N-HMM model

As our neutral corpus and emphasis corpus are recorded by two different speakers. The train of the neutral N-HMM model in fact involves two steps. The basic HMM model is first trained using the speech data of the neutral corpus. This basic HMM model is then adapted with the neutral speech recordings of the emphasis corpus while ignoring the emphasis labels to derive the final neutral N-HMM model. The standard maximum likelihood linear regression (MLLR) [7] is used for the adaptation.

For both training of the basic HMM model and adaptation to derive the N-HMM model, the 1,488 standard context questions are used for growing decision trees. These context questions are extracted from the official HTS toolkit [25], and are related to phones, positions, syllables, words, lexical stress, pitch accent, etc. Examples include: “Is the current phone [ey]?”, “Is the number of the syllables in the next word equal to 1?”, etc.

5.3 Training the perturbation model

As the features of the neutral speech utterances synthesized by the neutral N-HMM model may be different from the neutral recordings, to improve the prediction accuracy, the perturbation

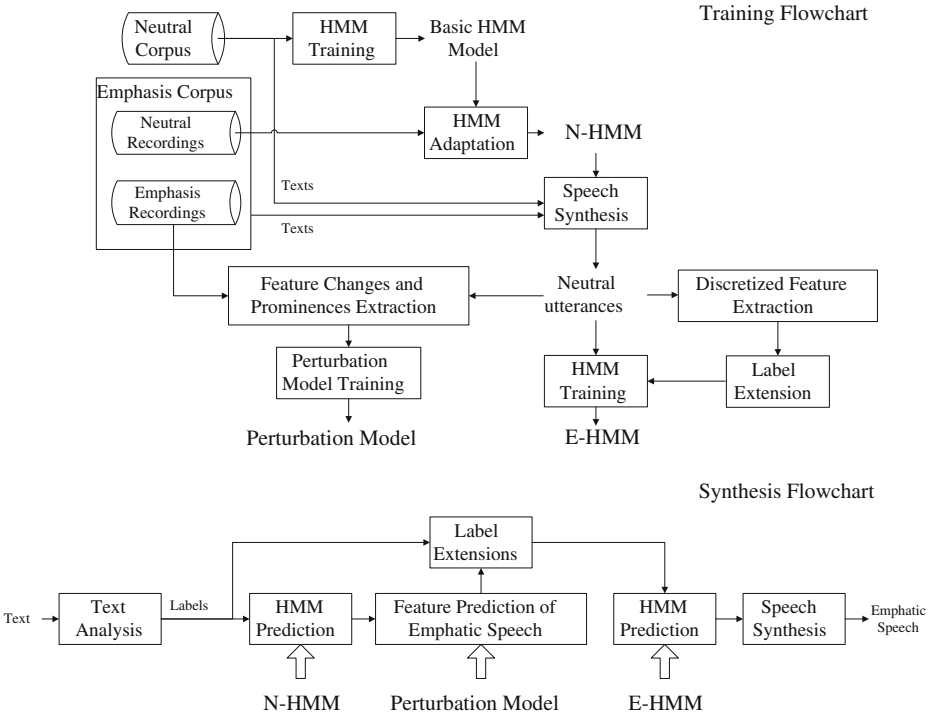

Fig. 3 The training process of the emphatic speech synthesis model based on HMM

model should be built by taking into account the new acoustic feature changes from the synthetic neutral speech to the original emphatic speech recordings. In doing so, the text prompts of the emphasis corpus are used to generate the neutral speech utterances with the N-HMM model. The changes of the acoustic features from the synthetic neutral speeches to the emphatic speech recordings are then calculated. The local prominences (*LP*s) of the acoustic features of the synthetic neutral speeches are also computed. These parameters are then used to train the perturbation model. The training process is detailed in Section 4.

5.4 Training the emphatic E-HMM model

As has been explained, generating emphatic speech from neutral speech by directly perturbing the acoustic features may degrade the speech quality a lot when the perturbation ratios are large. Hence, instead of modifying the features of the neutral speeches with the perturbation model directly, we choose to build another emphatic HMM model (E-HMM) which is supervised by the perturbation model. The purpose of the E-HMM model is to generate speech with acoustic features similar to the features of the emphatic speech predicted by the perturbation model. This is done by adding additional labels for the growing of the decision tree of E-HMM. We calculate the discretized acoustic features of the phones in the training corpus and add the discretized acoustic features to the labels of their corresponding phones. New questions related to the discretized acoustic features are then designed and added to the context question set. Then the corpus with the extended labels is used to grow the decision tree of E-HMM.

There are three steps for training E-HMM:

5.4.1 Preparing the neutral utterances

As there are only 350 neutral utterances in our emphasis corpus, the HMM will not be sufficiently trained with such small amount of data. All the text prompts from both the neutral corpus and the emphasis corpus (without emphasis label) are used to synthesize the neutral speech utterances with the N-HMM model. The synthetic neutral utterances have the same timbre as the emphasis corpus (i.e., sounds like from the same speaker). These synthetic speech utterances are used to train the E-HMM model later.

5.4.2 Preparing the labels and questions

In the process of traditional HMM-based speech synthesis, the inputs are the labels of the target phones, and the generated acoustic features cannot be controlled. To supervise the HMM model to generate acoustic features as required by emphatic speech (predicted by the perturbation model), additional labels and questions on acoustic features, including maximum f0, minimum f0 and durations, are designed and used to grow the decision trees.

For a certain feature (maximum f0, minimum f0 or duration), we extract the features of all the phones in the corpus. For the phone whose acoustic feature value is F, let F_{Min} be the minimum value of the feature of all the phones in the corpus. The value of the acoustic feature is discretized to generate the label L for this acoustic feature:

$$L = \left\lfloor \frac{F - F_{\mathrm{Min}}}{w} \right\rfloor \tag{15}$$

where w is the width of the discretization. It should be noted that the labels of maximum f0 and minimum f0 for voiceless phones are fixed to be 0. For instance, if the maximum f0 of a

phone are 235Hz, the width of the discretization of the f0 maximum is 10Hz and the minimum value of the feature maximum f0 (excluding voiceless phones) in our corpus is 60, then the generated label for the maximum f0 is $\lfloor(235-60)/10\rfloor=17$. These acoustic-feature-related labels are added to the original context-related labels of the corpus.

A set of questions is also added for each acoustic feature. The questions for a certain feature are to ask if the label for the feature of the current phone is equal to a special value. For example, the set of questions for maximum f0 is shown in Table 7.

5.4.3 Training the E-HMM model

Finally, the E-HMM model is trained with all the speeches synthesized by N-HMM using all labels and all the questions. Figure 4 shows the top part of the decision tree of the E-HMM model for durations.

As emphatic speech recordings of the emphasis corpus are only used during the training of the perturbation model, which predicts the changes of the acoustic features of the syllables from neutral to emphatic speech, the requirement for emphatic corpus of our model is much less than that of the traditional HMM model for emphatic speech synthesis.

6 Experiments and discussions

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. These experiments include two objective experiments and four subjective experiments. Three experiments of them are used to evaluate the emphatic speech perturbation model and others are used to evaluate the emphatic speech synthesis model.

For subjective evaluations, we invite ten participants. All of them are Ph.D or Master candidates in Tsinghua University.

Next, we first introduce our datasets used in the experiments. Then we evaluate the emphatic speech perturbation model, by comparing the mean absolute errors (*MAE*) and the root of the mean squared errors (*RMSE*) of the models with different parameters, and also two subjective experiments on the emphasis intensity and the naturalness of the converted speeches of the models. Finally, we evaluate the emphatic speech synthesis model, by comparing the prediction accuracy of the models with different discretization widths. Additionally two subjective experiments are carried to evaluate the emphasis intensity and the naturalness of the synthesized speeches of different models. The experimental results validate the effectiveness of our approach.

Table 7 The extended acoustic-feature-related question set for growing decision tree of E-HMM (take maximum f0 as the acoustic feature example)

Questions	Answers
Is the label for maximum f0 of the current phone 0?	Yes/No
Is the label for maximum f0 of the current phone 1?	Yes/No
Is the label for maximum f0 of the current phone no more than 1?	Yes/No
Is the label for maximum f0 of the current phone 2?	Yes/No
Is the label for maximum f0 of the current phone no more than 2?	Yes/No
…	…

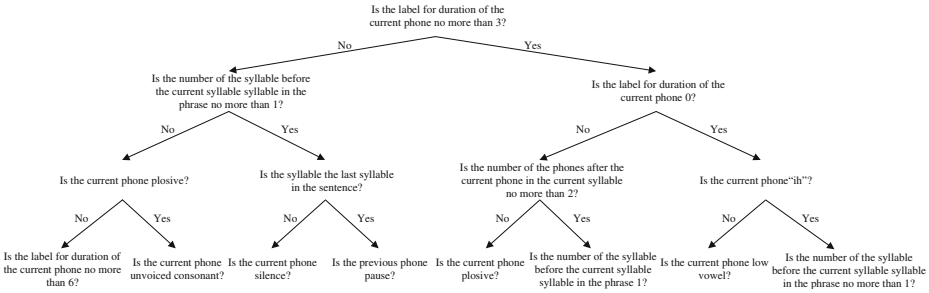

Fig. 4 The top part of the decision tree of the E-HMM model for durations

6.1 Datasets

The emphasis corpus, detailed in Section 2.1, is used to evaluate the emphasis perturbation model, among which 20 texts and the 40 corresponding utterances are used for testing and the left are used for training.

For the evaluation of the emphatic speech synthesis model, the neutral corpus, detailed in Section 2.2, is used for HMM training. And the neutral speeches in the emphasis corpus are used for HMM adaptation.

6.2 Evaluations of the emphatic speech perturbation model

To demonstrate the effectiveness of our proposed approach to emphatic speech perturbation model, three experiments are conducted in this section. As the input of these experiments are the neutral speechs. The STRAIGHT algorithm is used to extract the 39 Mel-frequency cepstral coefficients, log F0 and aperiodic components. The maximum f0, minimum f0, duration and energy of the neutreal speech are then derived from these parameters.

6.2.1 The experiment on the prediction accuracy of the perturbation model

This experiment is designed to compare the prediction accuracy of three models. The first model uses the decision tree to cluster the training data and then uses the Average of the Feature Vectors of each leaf node as the predicted values of the contexts (*AFV*). The second model uses the decision tree to cluster the training data and then uses the Local Prominences of the features of the neutral speech (*LP*) to predict the features of the emphatic speech using formula (6). Based on the second model, additional parameters, the Correlations between the Changes of Acoustic Features (*CCAF*), are considered in the prediction process of the third model, using formula (7).

MAE and *RMSE* are use to evaluate the prediction accuracy of the models. Let ΔF_i^j be the changes of feature j from neutral to emphatic speech of the ith sample, where $j \in \{P_{\mathrm{Max}}, P_{\mathrm{Min}}, E, D\}$, and ΔF_i^j be the predicted changes of feature j from neutral to emphatic speech of the ith sample. *MAE* and *RMSE* are calculated as:

$$MAE = \frac{\sum_{j \in \{P_{\mathrm{Max}}, P_{\mathrm{Min}}, \mathrm{D}, \mathrm{E}\}} \sum_{i=1}^{N} \left| \Delta F_i'^j - \Delta F_i^j \right|}{4N} \tag{16}$$

$$RMSE = \sqrt{\frac{\sum_{j\in\{P_{\mathrm{Max}},P_{\mathrm{Min}},\mathrm{D},\mathrm{E}\}}\sum_{i=1}^{N}\left(\Delta F_i'^{j}-\Delta F_i^{j}\right)^2}{4N}} \tag{17}$$

where N is the number of the samples.

Table 8 shows the results of the experiment. *MAE* and *RMSE* of different models for both training set and testing set are shown. The prediction errors (*MAE* and *RMSE*) of the model using *LP* are significantly lower than those of the model using *AFV*. This is because the data distribution of the features changes from neutral to emphatic speech in a leaf node is related to the local prominences of the features in the corresponding neutral speech. The modeling of LP describes this special relation and improves the prediction accuracy. The prediction errors of the model using *LP* and *CCAF* are a bit lower than those of the model using only *LP*, which indicates that the changes of a certain feature from neutral to emphatic speech are related with other features and involving this effect could improve the accuracy of the models.

6.2.2 The experiment on the emphasis intensity of the generated speech by the perturbation model

This experiment is designed to evaluate the ability of generating emphasis of the models. Ten neutral speeches from testing set are provided to two models. Each prompt contains one or more emphasized word(s). One is the model using *AFV* to predict features and the other is the model using *LP* and *CCAF* to predict features. The 20 converted speeches together with the ten corresponding emphatic recordings and the raw texts are presented to the subjects. 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:

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

Ten subjects participated in the experiment. Table 9 shows the results, 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 rate of the converted speeches of the model using *AFV* is 85 %, while that of the converted speeches of the model using *LP* and *CCAF* is 97 %, which is equal to the accuracy rate of the recordings. Besides, the rate of "False Positive" of the converted speeches of the model using *LP* and *CCAF* is 5 %, a bit higher than recordings, while the rate of "False Negative" is 3 %, a bit lower than recordings. The

Table 8 The prediction errors of the models with different modeling parameters

Modeling parameters	Training set		Testing set	
	MAE	*RMSE*	*MAE*	*RMSE*
AFV	0.16	0.14	0.19	0.17
LP	0.10	0.09	0.14	0.13
LP and *CCAF*	0.08	0.09	0.12	0.12

Table 9 The experiment results of the emphasis intensity of the converted speeches

Speech set	Accuracy		False positive		False negative	
	Rate	SC level	Rate	SC level	Rate	SC level
Recordings	97 %	4.7	2 %	4.5	3 %	–
AFV	85 %	3.5	6 %	2.5	15 %	–
LP and *CCAF*	97 %	4.3	5 %	3.6	3 %	–

results indicate that the model using *LP* and *CCAF* has stronger ability to generate emphasis than the model using *AFV*. It is because when the local prominences of the features in neutral speech are low, the model using *LP* and *CCAF* will give larger predicted changes than the model using *AFV* and the converted speech would be perceived more emphasized.

6.2.3 The experiment on the naturalness of the generated speech by the perturbation model

This experiment is designed to evaluate the naturalness of the converted speeches of the models. Another ten neutral speeches from testing set are provided to the same models in Section 6.2.2. Each sentence contains one or more emphasized word(s). The 20 converted speeches together with the ten corresponding emphatic recordings and the texts with emphasis annotations are presented to the subjects. The subjects are asked to give a 5-scaled MOS score according to the naturalness of the speech.

Ten subjects participated in the experiment. The mean MOS scores of the recordings, the converted speeches of the model using *AFV* and the converted speeches of the model using *LP* and *CCAF* are 4.8, 3.8 and 4.4. The experiment results shows that when the local prominences of the features of the neutral speech is high, the predicted changes of the model using *LP* and *CCAF* and lower than those of the model using *AFV*, avoiding oversize modification and increasing the naturalness of converted speech.

6.3 Evaluations of the two-stage HMM-based emphatic speech synthesis model

To demonstrate the effectiveness of our proposed emphatic speech synthesis model, three experiments are conducted in this section. The systems for the experiments are built with the multi-space density HMMs (MSDHMM) provided by the HTS toolkit [25] using different ways for HMM modeling. The static feature set includes 39 Mel-frequency cepstral coefficients, log F0 and aperiodic components extracted by the STRAIGHT speech analysis system. The speech parameters are modeled by 7-state left-to-right HMM. Four models are built for the experiments:

The first model is the traditional HMM adaptation model denoted by “adapt”. For emphasis speech synthesis, we add six emphasis-related questions according to the data analysis in Section 3. The six questions are:

(1) Is the phone In the Primary stressed syllable of an Emphasized word?
(2) Is the phone Before the Primary stressed syllable of an Emphasized word?
(3) Is the phone After the Primary stressed syllable of an Emphasized word?
(4) Is the phone in the Neutral word Before an emphasized word?
(5) Is the phone in the Neutral word After an emphasized word? and
(6) Is the phone Excluded from the Previous five categories?

Basic HMMs are first trained with all of the non-emphasis-related and emphasis-related questions using both neutral and emphasis corpora. MLLR [7, 19] is then used to adapt the parameters of the basic HMMs with the emphasis corpus to get the final HMMs for emphatic speech synthesis.

The second model is a hierarchical model based on HMM according to our previous work [14], denoted by "hierarchical". In the model, the training data are clustered by a two-pass decision tree, in which the non-emphasis-related questions are used for tree growing first. Based on the HMM model, we use a method based on cost calculation to select suitable HMM to predict parameters, and additional a compensation model is used to adjust the predict parameters.

The third model is the proposed model detailed in Section 5, denoted by "convert-model".

The fourth model is to add an emphatic speech perturbation model at the back end of a neutral speech synthesis model, denoted by "model-convert". We build the neutral speech synthesis model with the neutral corpus and the emphasis perturbation model using the method detailed in Section 4.

6.3.1 The experiment on the prediction accuracy of the emphatic speech synthesis models

This experiment is designed to compare the prediction accuracy of the models. Ten texts of the testing set are selected and provided to the four models. The features (P_{Max}, P_{Min} and D) of the syllables of the emphasized words in the 40 synthesized sentences are compared with those in the corresponding emphatic recordings.

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

$$A = 100 \times \left(1 - \frac{\sum_{i=1}^{N} \left(F_i'^j - F_i^j \right) \Big/ F_i^j}{N} \right) \% \tag{18}$$

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

The experiment results are shown in Table 10. The prediction accuracy of the model "hierarchical", "convert-model" and "model-convert" are significantly higher than that of the model "adapt". The accuracy of the model "convert-model" is a bit higher than that of the model "hierarchical", while that of the "model-convert" is the highest. It is because the model "hierarchical" uses the parameter distribution of global data to evaluate the parameter distribution of local data, involving in errors, while the model convert-model uses the perturbation model with high prediction accuracy to supervise the HMM model to predict parameters, increasing the prediction accuracy. But as the HMM model clusters the samples

Table 10 The prediction accuracy of different emphatic speech synthesis models

Models	A of P_{Max} (%)	A of P_{Min} (%)	A of D (%)
Adapt	83	81	63
Hierarchical	89	88	72
Convert-model	90	90	79
Model-convert	91	92	83

with similar features (among the discretization width) into one leaf node, the predicted value cannot be exactly the same as the target. Hence, the prediction accuracy of the model "convert-model" is lower than that of the model "model-convert".

6.3.2 The 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. Ten prompts from the test set were provided to each system. Each prompt contains one or more emphasized word(s). The resulting 40 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:

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

Ten subjects participated in the experiment. Table 11 shows the results, 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 model "adapt" has the lowest "Accuracy" and the highest "False Positive" and "False Negative". The results of the model "hierarchical" are similar to those of the model "convert-model" and a bit higher than those of the model "model-convert". This is because the features (P_{Max}, P_{Min} and D) of the emphasized words of the synthesized sentences of the model "model-convert" are similar to those of the model "convert-model", but the spectrums are not adjusted when the pitches and durations are modified, lowering the naturalness of the synthesized speeches, affecting the perception of emphasis.

6.3.3 The experiment of the naturalness of synthesized speech

This experiment is designed to evaluate the naturalness of the synthesized speeches of the models. Another ten prompts from testing set are provided to the four models. Each prompt contains one or more emphasized word(s) The 40 synthesized speeches together with the texts with emphasis annotations are presented to the subjects. The subjects are asked to give a 5-scaled MOS score according to the naturalness of the speech.

Ten subjects participated in the experiment. The average MOS scores of different models are shown in Table 12. The synthesized speeches of the model "adapt" have the highest

Table 11 The experiment results of the emphasis intensity of the synthesized speeches

Models	Accuracy		False positive		False negative	
	Rate	SC level	Rate	SC level	Rate	SC level
Adapt	70 %	2.8	15 %	2.6	30 %	–
Hierarchical	98 %	4.1	6 %	3.4	2 %	–
Convert-model	98 %	4.1	5 %	3.3	2 %	–
Model-convert	96 %	3.8	8 %	3.3	4 %	–

Table 12 The results of the experiment on the naturalness of the synthesized speeches

Models	MOS
Adapt	4.5
Hierarchical	4.3
Convert-model	4.3
Model-convert	3.9

MOS score, while those of the model "model-convert" have the lowest MOS score. The MOS score of the model "hierarchical" and the model "convert-model" are the same. This is because the model "hierarchical" and the model "convert-model" involve additional emphasis/feature related questions, which will make more leaf nodes have less data. As a result, the naturalness of the synthesized speeches decreases. But as the HMM model of the model "convert-model" are trained with neutral corpus, we believe that as using more neutral speeches (which could be collected much more economic than emphatic speeches), the naturalness of the synthesized speeches of the model "convert-model" could be improved while keeping the emphasis intensity of the synthesized speeches in a high degree. As the spectrums are not fit for the pitches, the synthesized speeches of the model "model-convert" have the lowest MOS score.

7 Conclusion

In this paper, we analyzed the acoustic features changes from neutral to emphatic speeches in different prosody locations, and unveiled the relationship between the features changes and the local prominences of the features in the neutral speeches. Based on the analysis, we proposed a emphatic speech perturbation model considering the prosody locations of the syllables, the local prominences of the features in the neutral speeches, and the correlations between the changes of acoustic features. Experiments showed that the proposed perturbation model can generate emphatic speech with both high emphasis intensity and high naturalness. The collection of emphasis corpus is a big problem for emphatic speech synthesis, as there are only a few emphasized words in a sentence. Aiming at this problem, this paper proposed an emphatic speech synthesis model, in which the HMM model was trained with neutral corpus. We used the proposed perturbation model to supervise the HMM model to synthesize the emphatic speeches. Experiments show that the proposed synthesis model could generate emphatic speech with high emphasis intensity and high naturalness. We believe that as the training data (neutral speeches) increases, the synthesized speeches could be improved further.

Future work will incorporate this emphatic speech synthesis model into an interactive CAPT platform, where synthesized emphasis aims to draw the learner's attention to segments of the system's feedback.

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 (60805008 and 61003094). 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. Bou-Ghazale SE, Hansen JHL (1996) Generating stressed speech from neutral speech using a modified CELP vocoder. Speech Comm 20:93–110, Oxford University Press
2. Bou-Ghazale SE, Hansen JHL (1998) HMM-based stressed speech modeling with application to improved synthesis and recognition of isolated speech under stress. IEEE Trans Speech Audio Process 6:201–216, IEEE Press
3. Chen SW, Wang B, Xu Y (2009) Closely related languages, different ways of realizing focus. Proceedings of Interspeech
4. http://www.cstr.ed.ac.uk/projects/festival/
5. Jia J, Zhang S, Meng FB, Wang YX, Cai LH (2011) Emotional audio-visual speech synthesis based on PAD. IEEE transactions on audio, speech, and language processing 19(3):570–582
6. Kominek J, Black AW (2003) CMU ARCTIC databases for speech synthesis. Tech. Rep. CMU-LTI-03-177, Carnegie Mellon University
7. Leggetter CJ, Woodland PC (1995) Maximum likelihood linear regression for speaker adaptation of continuous density HMMs. Comput Speech Lang 9:171–186
8. Li AJ (1994) Duration charateristics of stress and its synthesis rules on standard Chinese, report of phonetic research. CASS, Beijing
9. Li Y, Lu YC, Xu XY, Tao JH (2011) Influence of rhythm and tone pattern on Mandarin stress perception in continuous speech. J Tsinghua Univ (Sci Technol) 51:1239–1243, Tsinghua University Press
10. Li Y, Pan SF, Tao JH (2011) HMM-based expressive speech synthesis with a flexible Mandarin stress adaptation model. J Tsinghua Univ (Sci Technol) 51:1171–1175, Tsinghua Unversity Press
11. Liu F (2010) Single vs double focus in English statements and yes no questions. Proceedings of Speech Prosody. ISCA Press, Chicago
12. Maeno Y, Nose T, Kobayashi T, Ijim Y, Nakajima H, Mizuno H, Yoshioka O (2011) HMM-based emphatic speech synthesis using unsupervised context labeling. Proceedings of Interspeech. Oxford University, Italy, p. 1849–1852
13. Meng H, Lo WK, Harrison AM, Lee P, Wong KH, Leung WK, Meng FB (2011) Development of automatic speech recongition and synthesis technologies to support Chinese learners of English: the CUHK experience. Proceedings of APSIPA. Cambridge University Press, Taiwan, March
14. Meng FB, Wu ZY, Meng H, Jia J, Cai LH (2012) Hierarchical English emphatic speech synthesis based on HMM with limited training data. Proceedings of Interspeech. Oxford University Press
15. Morizane K, Nakamura K, Toda T, Saruwatari H, Shikano K (2009) Emphasized speech synthesis based on hidden Markov models. Proceedsing of Speech Database and Assessments Oriental COCOSDA International Conference. IEEE Press, p. 76–81
16. Neri A, Cucchiarini C, Strik H (2006) ASR-based corrective feedback on pronunciation: Does it really work? Proceedings of Interspeech. Pittsburgh, USA
17. Plag I (2006) The variability of compound stress in English: structural, semantic and analogical factors. Engl Lang Linguist 10(1):143–172, Cambridge University Press
18. Raux A, Black AW (2003) A unit selection approach to F0 modeling and its application to emphasis, Proceedings of ASRU
19. Rump HH, Collier R (1996) Focus conditions and the prominence of pitch-accented syllables. Lang Speech 39:1–17, MIT Press
20. Selkirk EO (1980) The role of prosodic categories in English word stress. Linguist Inq 11(3):563–605, MIT Press
21. Strangert E (2003) Emphasis by pausing. Proceedings of 15th ICPhS. Cambridge University Press, Barcelona, p. 2477–2480
22. Tamburini F (2003) Automatic prosodic prominence detection in speech using acoustic features: an unsupervised system. Proceedings of Eurospeech. Oxford University, 129–132
23. Tamburini F (2003) Prosodic prominence detection in speech. Proceedings of Signal Processing and its Applications. IEEE Press, p. 385–388
24. Tokuda K, Yoshimura T, Masuko T, Kobayashi T, Kitamura T (2003) Speech parameter generation algorithms for HMM-based speech synthesis. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey, 3:1315–1318
25. 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/
26. Xie L, Liu ZQ (2007) Realistic mouth-synching for speech-driven talking face using articulatory modelling. IEEE Multimedia 9(3):500–510

27. Xu JP, Chu M, Lin HE, Lu SN (2004) The influence of Chinese sentence stress on pitch and duration. Chin J Acoust 4:335–339, Allerton Press
28. Xu Y, Xu CX (2005) Phonetic realization of focus in English declarative intonation. J Phon 33:159–197, Academic Press
29. Yu K, Mairesse F, Young S (2010) Word-level emphasis modeling in HMM-based speech synthesis. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Cambridge University Press, p. 4238–4241
30. Zhou QF, Cai LH (1996) Mandarin stress and its simulation in TTS system. Microcomputer 16(4):16–19, Microcomputer Press
31. Zhu WB (2007) A Chinese speech synthesis system with capability of accent realizing. J Chin Inf Process 21(3):122–128, Chinese Information Processing Press

Fanbo Meng received the B.E. degrees in computer science and technology from Tsinghua University, Beijing, China, in 2007. He is now a Ph.D candidate in Tsinghua University.

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.

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.

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.

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 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 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.