Hierarchical English Emphatic Speech Synthesis Based on HMM with Limited Training Data

Fanbo Meng[1], Zhiyong Wu[2,3], Helen Meng[2,3], Jia Jia[1,3] and Lianhong Cai[1,3]

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

mfb03@mails.tsinghua.edu.cn, {zywu, hmmeng}@se.cuhk.edu.hk, {jjia, clh-dcs}@tsinghua.edu.cn

Abstract

Emphasis is an important form of expressiveness in speech. Hidden Markov model (HMM) based speech synthesis has shown great flexibility in generating expressive speech. This paper proposes a hierarchical model based on HMMs aiming at synthesizing emphatic speech of both high emphasis quality and high naturalness with the limited amount of data. 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 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. Experiments show that the proposed hierarchical model can synthesize emphatic speech with high quality for both naturalness and emphasis, using limited amount of training data.

Index Terms: emphatic speech synthesis, hidden Markov model (HMM), hierarchy, compensation model

1. Introduction

State-of-the-art speech synthesis technologies can generate synthetic speech with a 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.

There are two typical methods for emphatic speech synthesis – one is to concatenate units from recorded speech that carries emphasis [1]. The other method is parametric speech synthesis, e.g. using HMM [2]. The former method requires a large amount of recorded speech that carries emphasis. However it is difficult to acquire such data, as a typical recorded sentence usually contains few emphasized words. The latter approach, specifically, HMM-based speech synthesis, provides a data-driven framework with flexible control of expressiveness. It groups the training data into different clusters by means of DTs, with each cluster sharing the same distribution of acoustic features. However in the case of emphatic speech synthesis, the data for emphatic speech are much less than those for non-emphatic. The imbalanced data distribution decreases the probability for emphasis-related questions to be used in DTs. Hence, the HMMs cannot train the associated acoustic models sufficiently, leading to low emphasis quality of the synthetic speech. To address this issue, Yu [3] proposed the two-pass DT method. The main DT was constructed using the emphasis-related questions at the word layer (e.g. is the current word emphasized?) using all the data, and then the leaves of the main DT were extended using the non-emphasis-related questions (e.g. is the current phone [ax]?). Due to the imbalanced distribution of emphasis data, further splitting in leaf nodes of the main DT into sub-trees with non-emphasis-related questions means that the ultimate leaf nodes do not give well-trained acoustic models. Yu then devised the factorial DT approach [3]. The general DT is constructed with non-emphasis-related questions using all the data. The emphasis DT is constructed with emphasis-related questions at the word layer using all the data. Then the emphasis DT is appended to each leaf node of the general DT to further split the data clusters. With this method, there may be no data in some of the leaves of the general DT for speech of such emphasis contexts.

As we can see, data sparseness in the corpus is an important limitation for emphatic speech synthesis. This paper proposes a hierarchical English emphatic speech synthesis model based on HMM, aiming to synthesize speech with both high emphasis quality and high naturalness, despite having limited amount of training data. To model emphasis better, more emphasis-related questions are designed for the word and syllable layers. We use non-emphasis-related questions to construct a general DT. Then the data in each leaf are further split into sub-tree of 6 emphasis categories according to the emphasis-related questions. The data of each emphasis category (in each leaf node of the sub-tree) are used to train an HMM. There may be no data in some emphasis categories. To address the problem, we select an HMM that is trained on data from other leaf nodes of the sub-tree according to a cost function. The cost function is based on phonetic broad classes (which we refer as *phone types*). Furthermore, a compensation model is proposed to adjust the f0 and duration generated by the selected HMM in an attempt to improve both the quality of emphasis and naturalness of the synthesized speech.

2. Corpora

2.1. The corpus of neutral speech (neutral corpus)

We use the CMU US ARCTIC clb corpus. It has 1,132 utterances recorded by an US female speaker, stored in the 16Bit mono format as wav files with 16kHz sampling rate. The corpus is automatically annotated by FestVox [5].

2.2. The corpus of emphatic speech (emphasis corpus)

350 text prompts are carefully designed by considering the factors affecting the expression of emphasis at the word, syllable and phone layers. For the word layer, one or more emphasized words are contained in each text prompt, with each emphasized

word located at a different position in the sentences. For the syllable layer, the words may be monosyllabic or polysyllabic, with the primary stressed syllables at different places. For the phone layer, we strive to attain complete phone coverage and broad phonetic coverage in our corpus. Examples include (with emphasized words in boldface):

"Fighting **thirst** is the **first** thing to be done in this country."

Each text prompt is recorded twice – once with neutral intonation throughout the utterance and the other with emphasis placed on the selected words. A female speaker with a high level of English proficiency was invited to record in a studio. Hence we have 700 recorded utterances, saved in the wav files (16Bit mono, sampled at 16kHz). This corpus is also automatically annotated by FestVox using the raw text transcription of prompts.

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.

3. Modeling emphatic speech with HMM

3.1. Growing a general decision tree and emphasis-related sub-trees

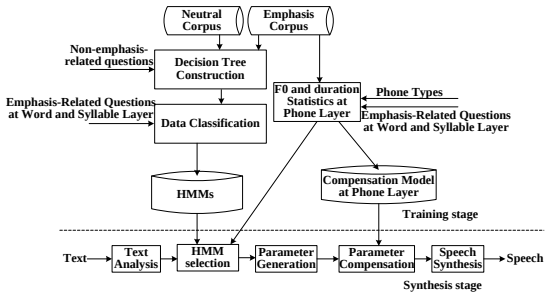

Figure 1: *The diagram of the hierarchical model for emphatic speech synthesis combining HMM with compensation model*

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

Figure 2: *An example of the 6 phone categories*

Figure 1 shows the diagram of the proposed hierarchical model for emphatic speech synthesis. First, a general DT is constructed with the minimum description length (MDL) criterion [4] using the training data from the neutral corpus. This grows a DT according to 1,488 standard context questions (non-emphasis-related) from the official HTS toolkit [2]. The context questions 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.

The general DT is used to group the phones of the emphasis corpus into different clusters (i.e. leaf nodes). Each leaf node of the general DT may contain phones with different emphasis attributes, e.g. from emphasized or non-emphasized words. The HMMs trained using the data from such leaves can generate speech with high naturalness but with low emphasis quality. To address the problem, the phones in the leaf nodes are classified into 6 emphasis categories, using emphasis-related questions at the word and syllable layers. The 6 questions or categories are:

(1) **I-P-E**: Is the phone *In* the *Primary* stressed syllable of an *Emphasized* word?
(2) **B-P-E**: Is the phone *Before* the *Primary* stressed syllable of an *Emphasized* word?
(3) **A-P-E**: Is the phone *After* the *Primary* stressed syllable of an *Emphasized* word?
(4) **N-B**: Is the phone in the *Neutral* word *Before* an emphasized word?
(5) **N-A**: Is the phone in the *Neutral* word *After* an emphasized word? and
(6) **E-P**: Is the phone *Excluded* from the *Previous* 5 categories?

Figure 2 illustrates the method of this phone categorization, where "PETERSON" and "OCCASION" are emphasized words. This categorization further splits each leaf node of the general DT into an emphasis-related sub-tree. The leaf nodes of this sub-tree are defined as the *sub-nodes* of the leaf node of the general DT.

3.2. HMM training for emphatic speech synthesis

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

1) The general DT 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 DT.

2) The same general DT 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) [6] is used to adapt the parameters of the above HMMs using the data from the emphasis corpus for each leaf node of the general DT.

3) The emphasis-related sub-trees are further used to divide the data in each leaf node of the general DT into sub-nodes. The phones of each sub-node belong to the same emphasis category and are used to adapt the HMM from the parent leaf node of the general DT with MLLR [6] to get the final HMMs for emphatic speech synthesis.

However, due to the limited amount of emphasis data in the corpus, there may 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 DT are found to contain no data in the I-P-E category when they are further split based on the emphasis-related sub-trees.

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 DT, 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 the phone layer.

Table 1. *Statistics of average durations (D, in ms) and average f0s (f_0, in Hz) of the phones from different emphasis categories (EC) and phone types (PT)*

EC / PT	I-P-E		B-P-E		A-P-E		N-B		N-A		E-P	
	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

3.3.1. Statistics from the emphasis corpus

Recall that the phones in the emphasis corpus are first classified into 6 emphasis categories. The phones in each emphasis

category are further classified into 9 broad classes / phone types:
(1) long vowels and diphthongs, e.g. [iy], [ey], [ow];
(2) mono vowels, e.g. [ih], [ae];
(3) plosives, e.g. [p];
(4) nasals, e.g. [m], [n];
(5) fricatives, e.g. [z];
(6) retroflexed liquids, e.g. [r];
(7) lateral liquids, e.g. [l];
(8) glides, e.g. [y]; and
(9) affricates, e.g. [ch].

The average f0s and durations of the phones in different emphasis categories and phone types are shown in Table 1.

3.3.2. Cost function for HMM selection

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

Suppose we are going to synthesize speech for a target 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 $N=n_1+n_2+\ldots+n_9$. 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:

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

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.

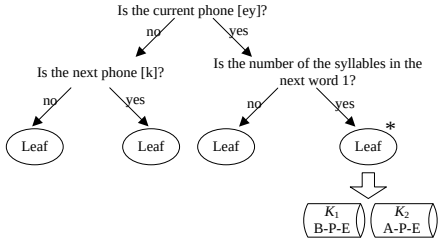

Figure 3: *Part of the decision tree for generating f_0. Data is available for only two emphasis categories B-P-E (K_1) and A-P-E (K_2) in the leaf node annotated by "*". There are four phones in K_1: [ey][ow][ae][ch] and two phones in K_2: [m][n]*

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 DT for generating f0 is show in Figure 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 Equation (2) and the HMM trained by the data in K_1 is selected for generating f0.

$$C_1=\left|1-\frac{1}{207}\frac{1}{4}(2\times190+192+192)\right|=0.08, C_2=\left|1-\frac{1}{207}182\right|=0.12 \quad (2)$$

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 DT 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:

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

where R_{f0} is the compensation factor for f0, which can be computed as Equation (4) using the statistic information from the emphasis corpus as shown in Table 1.

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

where the notations for $f_{0e,p}$, $f_{0m,t}$ and *e*, *p*, *m*, *t* are all the same as those in Equation (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 $R_{f0}=1$.

Recall the example in section 3.3.2, 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:

$$R_{f_0}=\frac{207}{(2\times190+192+192)/4}=1.08 \quad (5)$$

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 [2] to generate the synthetic emphatic speech.

4. Experiments and discussion

The systems for the experiments are built with the multi-space density HMMs (MSDHMM) provided by the HTS toolkit [2] using different ways to construct DTs. 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. Three systems are built for the experiments:

The first system is the traditional HMM adaptation system, denoted by "adapt". Basic HMMs are first trained with all of the

non-emphasis-related and emphasis-related questions using both neutral and emphasis corpora. MLLR [6] 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 system is the two-pass DT system by Yu, denoted by "2-pass-Yu". We construct the main DT with emphasis-related questions using both neutral and emphasis corpora, and then extend the leaves of the main DT with non-emphasis-related questions.

The third system is the proposed hierarchical system, denoted by "hierarchical", which is detailed in section 3.

4.1. Evaluating emphasis quality

10 prompts from the test set were 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, were presented to the subjects in random order. Each subject was asked to listen to the sentence and identify which word(s) are emphasized. The subject was 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).

15 subjects participated in the experiment. Table 2 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 DT system and the proposed hierarchical system are much higher than those of the emphasis adaptation system. The "False Positive" rate of the hierarchical system is slightly lower than that of the two-pass DT system, and the confidence levels are higher. These indicate the proposed hierarchical system can synthesize emphatic speech with almost the same emphasis quality as the two-pass DT system, and much higher than the emphasis adaptation system.

Table 2. *Evaluation of emphasis quality through an emphasis identification experiment (SC level: subjects' confidence level). As the subjects are only asked to give confidence level for the identified emphasized word, no SC level for "False Negative"*

Systems	Accuracy		False Positive		False Negative	
	Rate	SC level	Rate	SC level	Rate	SC level
adapt	70%	2.8	15%	2.6	30%	-
2-pass-Yu	98%	4.1	8%	3.2	2%	-
hierarchical	98%	4.0	6%	3.4	2%	-

4.2. Evaluating naturalness

Another 10 prompts from the test set were used in this experiment. Some prompts contain one or more emphasis word(s), while others do not. For each text prompt, 3 speech files were generated by the 3 systems. The text prompts with emphasis annotations were provided to the subjects. Each subject was asked to listen to the 3 files with the same text prompt and give the order of the 3 files according to the naturalness of speech. Equality is permitted if it is difficult to distinguish the naturalness between the 2 or 3 files.

The same 15 subjects participated in the experiment. Figure 4 shows the preference rate of naturalness between different systems and the 95% confidence interval, where the preference rate is calculated as the percentage of the speech files that are identified to have the best naturalness among the 3 files with the same text prompt. Since the files may be perceived to have equal naturalness by the subjects, the sum of the preference rates from 3 systems is larger than 1. As can be seen, the naturalness of the speech files generated by the proposed hierarchical system is slightly lower than that of the emphasis adaptation system, but much higher than that of the two-pass DT system.

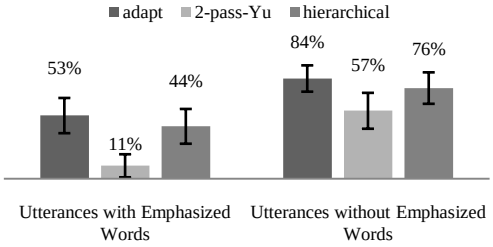

Figure 4: *Evaluating the naturalness of synthetic speech*

5. Conclusions

This paper presents an approach for synthesizing emphatic speech based on HMM. This work first constructs a general decision tree (DT) with non-emphasis-related questions using both neutral and emphasis corpora. Then the data in each leaf node of the general DT are further split into 6 emphasis categories based on emphasis-related questions. This forms the emphasis-related sub-tree. The data within the same emphasis category are grouped and used to train an HMM. Due to the limited quantities of emphasized speech data, there may be no data for some of the sub-tree's leaf nodes. Hence, no HMMs can be trained for synthesis in the given context. To address this problem, we designed a cost function with which we can select another HMM in the same sub-tree for synthesis. Furthermore, a compensation model at the phone layer is proposed to modify HMM-predicted parameters to improve synthesis quality. Experiments show that the proposed method can synthesize emphatic speech with high emphasis quality as compared with two-pass decision tree method by Yu and with high naturalness as compared with the traditional HMM adaptation method.

6. Acknowledgements

The work is jointly supported by the research funds from the Hong Kong SAR Government's Research Grants Council (CUHK4161/08), the National Natural Science Foundation of China (60928005, 60805008, 60931160443 and 61003094), and the NSFC/RGC Joint Research Scheme (project no. N_CUHK 414/09).

7. References

[1] Raux, A., Black, A.W., "A unit selection approach to F0 modeling and its application to emphasis", *Proc. ASRU*, 2003.

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

[3] Yu, K., Mairesse, F., Young, S., "Word-level emphasis modeling in HMM-based speech synthesis", *Proc. ICASSP*, 2010.

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

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

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