Generating Emphasis from Neutral Speech using Hierarchical Perturbation Model by Decision Tree and Support Vector Machine

Fanbo Meng[1], Zhiyong Wu[1,2,3], 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

In a computer-aided pronunciation training (CAPT) system, corrective feedback is desired to provide contrastive comparisons between user's and canonical pronunciations. This paper presents a hierarchical perturbation model to generate emphasis for English by modifying acoustic features of neutral speech to highlight such important speech segments. Synthesis of emphasis needs to be realized hierarchically at word, syllable and phone layers. A two-pass decision tree is constructed to cluster acoustic variations between emphatic and neutral speeches. The questions for decision tree construction are designed according to the above layers. The questions related to word and syllable layers are used to construct the main tree and then the questions related to phone layer are used to expand the leaves of main tree (deriving a set of sub-trees). Support vector machines (SVMs) are used to predict acoustic variations for all the leaves of main tree (at word and syllable layers) and sub-trees (at phone layer). Gradient descent algorithm and cross validation is used to estimate the parameters of SVMs. Experiments indicate that the proposed hierarchical perturbation model can generate emphatic speech with high quality for both naturalness and emphasis.

1. Introduction

CAPT system uses speech technologies to help language learners in pronunciation training. The availability of corrective feedback is very effective in reducing pronunciation errors [1]. Our goal is to provide corrective feedback with emphasis to highlight important speech segments that should draw the attention of the learner in an English CAPT system.

Emphasis is an important feature of prosody [2,3], which is expressed by different acoustic variations at word, syllable and phone layers. At word layer, the acoustic features relating to emphasis are affected by emphasized words and their locations in the utterance. Chen [4] found that f0s and energy after emphasized word decrease in English. Post-pitch suppression exists for double focus in English statements [5]. Barbosa's work showed that the nearer the word is to emphasized word, the longer is the duration [6]. At syllable layer, the acoustic features of emphasis are affected by stressed syllables in a word. Jong [7] found that, from neutral to emphatic speech, the duration variations of stressed syllables of emphasized words are bigger than those of other syllables. At phone layer, the acoustic features relating to emphasis are affected by the pronunciation mechanism of phones. Acoustic variations may reveal different patterns for different phones. Costa [8] analyzed the pitch and durations of vowels and consonants from emphatic speech, in comparison with neutral speech, and found that the durations of high vowels were shorter than for low vowels, and the pitch values were higher. Hence, hierarchical information is important for the modeling and generation of emphasis.

There are typically two kinds of methods [9,10] for generating prosody parameters (e.g. f0) for the target expressive speech. The first one is to build a context-dependent statistical model based on expressive corpus and use the model to directly generate target prosody parameters for speech generation [9]. The second one is to build a mapping model between neutral speech and expressive speech and predict prosody parameters of expressive speech from neutral speech at generation stage [10]. The first method often results in stereotyped prosody that does not reflect the natural variability of prosody features. Furthermore, as has been introduced,

our goal is to provide corrective feedback for CAPT system where the user's input speech is modified by emphasizing important speech segments with wrong or improper pronunciation. Hence, in this work, we focus on the second method to generate emphasis from neutral speech.

In converting neutral speech to expressive one, several methods were proposed. In [10], the conversion of f0 feature was realized at sentence, prosodic word and syllable layers using the Gaussian mixture models and classification and regression trees. [11] modeled f0 contour with discrete cosine transform coefficients at syllable layer and considered temporal correlations between syllables at phrase layer. In these methods, to convert prosody features, the same model parameters were used for all the units belonging to the same layer. This does not match the fact that the prosody features of a unit will affect the features of its neighbors and such impact tends to decrease for far neighbors [6].

Our previous work [12] classified phones into 6 categories based on the relative location of the phone in relation with the nearest emphasized word and its stressed syllables. A rule based perturbation model was proposed where a set of fixed values were used to modify prosody features of neutral speech to generate emphatic speech.

This paper attempts to analyze the acoustic features of emphasis at word, syllable and phone layers, and proposes a hierarchical perturbation model to generate emphasis from neutral speech by taking into account different acoustic variations at different layers. This work extends our previous work in two aspects.

1) To model hierarchical characteristics of emphasis, a two-pass decision tree is constructed which models acoustic variations between emphatic and neutral speeches at not only word and syllable layer but also phone layer. The main tree is first constructed by considering emphasis-related questions at word and syllable layer. For each leaf of the main tree, a sub tree is further constructed by considering emphasis-related questions at phone layer. Emphasis is expressed by different acoustic feature variations at different layers. Different acoustic features are considered for distance calculation when constructing the decision trees.

2) Instead of using fixed values for acoustic feature modification, a set of SVMs are built for all the leaves of both the main tree and sub trees. These SVMs are used to predict the acoustic variations for the leaves of the main tree at word and syllable layer, and the leaves of the sub trees at phone layer.

The rest of the paper is organized as follows: Section 2 presents the corpus designed to support our experimentation. Section 3 describes the analysis of acoustic features related to emphasis at word, syllable and phone layers. Section 4 details the hierarchical perturbation model for emphasis generation. Section 5 details the realization of our hierarchical perturbation model. Section 6 describes the perceptual evaluations of the outputs of the model. Finally, Section 7 lays out conclusions and possible future directions.

2. Corpus

To generate corrective feedback with exaggerated emphasis that highlights important speech segments to draw learner's attention, a set of text prompts are carefully designed and contrastive speech utterances are recorded for the analysis and modeling of emphasis.

2.1. Design of text prompts

The text prompts (350 in all) are carefully designed by considering the factors affecting the expression of emphasis at word, syllable and phone layers. For word layer, one or more emphasized words are contained in each text prompt, with each emphasized words located at different positions in the sentences. For syllable layer, the words are monosyllabic and polysyllabic, with the primary stressed syllables at different places. For phone layer, the phones with all kinds of pronunciation mechanisms are covered by the text prompts. The contexts of the phones are also covered as many as possible.

Two example text prompts are shown as follows (with focus words in boldface and underlined):

"*Fighting **thirst** is the **first** thing to be done in this country*." and "*I have met **Peterson** on one **occasion***."

2.2. Contrastive speech recordings

Two contrastive utterances are recorded for each text prompt – one with neutral intonation throughout the utterance and the other with expressive intonation to emphasize emphasized words in the sentence. A female speaker with a high level of English proficiency is invited to record in a studio. We have 700 recorded utterances, saved in the wav format as sound files (16 bit mono, sampled at 16 kHz). The corpus is annotated by FestVox [13] using the text transcription of prompts. The pitch contours and the phone, syllable and word boundaries are then derived from the annotation result.

From the 350 text prompts, 20 prompts (and 40 related utterances) are randomly selected as the test set for experimentation, all the other prompts (and related utterances) are used as the training set.

3. Acoustic analysis of emphasis

3.1. Extraction of acoustic features

Acoustic features associated with prosody include fundamental frequency (f0), intensity and speaking rate. The following acoustic features are extracted to capture the acoustic correlations of emphasis:

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

Measurements are taken from the contrastive speech recordings of each prompt. We compute the ratio (in %) between the measurements of the corresponding emphasized and neutral units, and variances of the ratios.

3.2. Acoustic analysis of emphasis at word and syllable layer

Following the scheme in our previous work [12], we classify the syllables into 6 classes at word and syllable layer, based on the location of the syllable in relation with the nearest focus word and its stressed syllables:

- *S-E*: the primary ***S***tressed syllable of a ***E***mphasized word;
- *B-S-E*: syllable ***B***efore the primary ***S***tressed syllable of a ***E***mphasized word;
- *A-S-E*: syllable ***A***fter the primary ***S***tressed syllable of a ***E***mphasized word;
- *N-B*: syllable in the ***N***eutral word ***B***efore a focus word;
- *N-A*: syllable in the ***N***eutral word ***A***fter a focus word;
- *O-R*: All ***O***ther ***R***emaining syllables.

Table 1 shows the feature variations from neutral speech to emphatic speech at word and syllable layer. It indicates that the feature variations of emphasized words are significantly higher than those of non-emphasized words, while the features of the stressed syllables of emphasized words change the most. Generally the nearer to the stressed syllables, the higher the feature changes are. In addition, clear lengthening of pause durations are observed between emphasized words and non-emphasized words.

Table 1 Feature variations from neutral speech to emphatic speech at word and syllable layer (%), and "*Pause*" variations between emphasized and non-emphasized words

		Max	*Min*	*R*	*Mean*	*S*	*E*	*D*
S-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-S-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-S-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
O-R	Ratio(%)	97	96	138	96	179	100	103
	Var	0.05	0.02	19.33	0.03	28.95	0.01	0.85
Pause	Ratio(%)	-	-	-	-	-	-	867

3.3. Acoustic analysis of emphasis at phone layer

To analyze the feature variations of different phones, we further group English phones into 9 types according to their pronunciation mechanism:

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

Table 2 shows the duration (*D*) variations and Table 3 shows mean f0 (*Mean*) variations of phones from neutral speech to emphatic speech. As the phones in the syllables with class "*O-R*" (all other remaining syllables) are far from emphasized words, their features are little affected by emphasized words. The acoustic variations of the phones for all other classes are computed against the variations of the phones with class "*O-R*". For the emphasized words (*S-E*, *B-S-E*, *A-S-E*), it shows that the duration variations of long vowel, diphthong and plosive are bigger than average, while the duration variations of fricative and glide are smaller than average. The variations of mean f0 are much smaller than those of duration. For the emphasized words (*S-E*, *B-S-E*, *A-S-E*), it shows that in *S-E* the mean f0 variations of vowels are higher than

average; while in *A-S-E*, the mean f0 variations of vowels are all lower than average, and the mean f0s of the syllables of *A-S-E* in table 1 increase. This indicates that the increase of the f0s of *A-S-E* is mainly due to the consonants.

Table 2 Duration (*D*) variations of phones from neutral speech to emphatic speech (%)
(SC: Syllable Class, PT: Phone Type)

PT \ EC	*S-E*	*B-S-E*	*A-S-E*	*N-B*	*N-A*	*O-R*
Long vowel and diphthong	160	146	123	111	96	100
Mono vowel	198	100	119	120	112	100
Plosive	183	132	144	112	106	100
Nasal	133	127	88	116	118	100
Fricative	144	101	104	114	120	100
Retroflex liquid	150	126	154	99	86	100
Lateral liquid	140	118	94	83	73	100
Glide	131	106	76	97	102	100
Affricate	171	116	79	178	97	100
Average	154	118	105	114	101	100

Table 3 Mean f0 (*Mean*) variations of phones from neutral speech to emphatic speech (%)
(SC: Syllable Class, PT: Phone Type)

PT \ EC	*S-E*	*B-S-E*	*A-S-E*	*N-B*	*N-A*	*O-R*
Long vowel and diphthong	110	101	97	100	95	100
Mono vowel	116	103	98	101	95	100
Plosive	106	103	102	99	88	100
Nasal	103	105	100	99	92	100
Fricative	103	100	117	95	92	100
Retroflex liquid	104	102	102	105	97	100
Lateral liquid	106	102	100	97	93	100
Glide	106	102	109	100	98	100
Affricate	100	102	100	119	91	100
Average	106	102	103	102	93	100

4. Hierarchical perturbation model to generate emphasis

4.1. Feature selection for different layers

We observe that the variances of acoustic feature *R* and *S* are approximately 100 times of other features in table 1. Regardless of the two features, the feature *Max* changes the most at the emphasized words. Addition, *Max* and *Min* are the features for a certain length of speech segment, they will be unstable at phone layer. Hence, we choose *Max* and *Min* to control the f0 range at word and syllable layer. The mean f0 (*Mean*) of different phone types have some special pattern, e.g., the f0 increase of *A-P-E* is mainly due to the consonants phones. And the durations of syllables could be estimated by the durations of phones. Hence we use *Mean* and *D* at phone layer. Besides, as *E* is very stable in table 1, we use *E* at word and syllable layer as a minor feature.

4.2. Two-pass decision tree for feature clustering

A two-pass decision tree is constructed which models acoustic variations between emphatic and neutral speeches not only at word and syllable layer but also at phone layer.

To construct the two-pass decision tree, a set of emphasis-related questions are designed according to the hierarchical characteristics for generating emphasis, as shown in Table 4. The main decision tree is first constructed by considering the emphasis-related questions at word and syllable layer. The emphasis-related questions related to phone layer are then used to expand the leaves of the main tree which leads to a set of sub trees. Each sub tree corresponds to one leaf of the main tree.

Emphasis is expressed by the variation of different acoustic features at different layers. Hence, different acoustic features are considered for calculating the distances when constructing the main tree and sub trees. For the main tree, the maximum f0 (*Max*), minimum f0 (*Min*) and energy (*E*) are used for distance calculation. While for the sub trees, the mean f0 (*Mean*) and duration (*D*) are used for distance calculation.

Table 4 Emphasis-related questions and answers for decision tree construction at different layers

	Question	Answer
Word layer	Current word is focus word	1/0
	Next word is focus word	1/0
	Previous word is focus word	1/0
Syllable layer	Current syllable is primary stressed syllable	1/0
	Next syllable is primary stressed syllable	1/0
	Previous syllable is primary stressed syllable	1/0
Phone layer	Current phone belongs to type *i* (*i* = 1, …, 9)?	1/0

4.3. SVM for feature prediction

4.3.1. The nu-SVR model

We use the nu-SVR [15] model as the regression model to predict the acoustic variations for each leaf of the main tree and sub trees.

Let (Y_i, X_i, i=1,2,…,l) be a set of data, where X_i is the input and y_i is the output. The regression function is as formula (1):

$$f(X)=\sum_{i=1}^{l}(\alpha_i^*-\alpha_i)K(X_i,X)=\sum_{i=1}^{l}(\alpha_i^*-\alpha_i)\phi(X_i)+b \quad (1)$$
$$=W\phi(X)+b$$

where $\alpha_i^*>0$, $\alpha_i>0$, i=1,2,…,l. The X_i which is not zero is the support vector. K is the kernel function. The problem could be solved by maximize formula (2), where the constraint condition is formula (3).

$$W(\alpha,\alpha^*)=-\varepsilon\sum_{i=1}^{l}(\alpha_i+\alpha_i^*)+\sum_{i=1}^{l}(\alpha_i^*-\alpha_i)y_i \quad (2)$$
$$-\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_i)K(X_i,X_j)$$

$$\sum_{i=1}^{l}(\alpha_i^*-\alpha_i)=0,\alpha_i\in\left[0,\frac{C}{l}\right],\alpha_i^*\in\left[0,\frac{C}{l}\right], \quad (3)$$
$$\sum_{i=1}^{l}(\alpha_i^*+\alpha_i)\le C\cdot\mu$$

The optimal α_i^*, α_i, i=1,2,…,l can be calculated by maximizing formula (2). Formula (1) is used for data prediction.

In this paper, for training the SVMs for the leaves of the main decision tree, *Max*, *Min* and *E* of neutral speech are used as the input, while the variations of the three features from neutral speech to emphatic speech are used as the output. For training the SVMs for the leaves of the sub trees, *Mean* and *D* of neutral speeches are used as the input, while the variations of the two features from neutral to emphatic speech are used as the output. The SVMs are then trained using the following cross validation and gradient descent algorithm.

4.3.2. Optimization of initial parameters for SVMs

In this paper, Gaussian Radial Basis Function is used as the kernel function of SVM, as formula (4).

$$K(X_i,X)=\exp(-|X_i-X|/2\sigma^2) \quad (4)$$

Cross validation method and gradient descent algorithm are used to determine the best initial values for the parameters the cost (C) in formula and constant nu (μ) in formula (3) and the σ in formula (4) for improving the performance of the SVMs.

The cross validation method partitions the data into N folds. Of the N folds, 1 fold of the data is used as the validation data for testing the model, the other N-1 folds are used as the training data. The cross validation process will repeat N times, with each of the N folds used exactly once as the validation data. This technique can make full use of all training data when training data is limited. In this paper, N is set to 10 and mean squared error (MSE) is used as the evaluation function which should be minimized while optimizing the initial parameters (C, μ, σ) for SVMs.

There are 5 steps for initial parameter optimizing using the gradient descent algorithm based on MSE:

(1) Initialize the parameter (C, μ, σ) and the update step (ΔC, $\Delta\mu$, $\Delta\sigma$). Let I be the total number of iterations and F be the number of continuous iterations where MSE is not improved. Set I to 1, and F to 0.

(2) Calculate the MSE of SVM based on the current parameter (C, μ, σ), denoted as MSE(C, μ, σ).

(3) Calculate MSE(C+ΔC, μ, σ), (C-ΔC, μ, σ), (C, μ+$\Delta\mu$, σ), (C, μ-$\Delta\mu$, σ), (C, μ, σ+$\Delta\sigma$) and (C, μ, σ-$\Delta\sigma$), denoted by MSE_{C+}, MSE_{C-}, $MSE_{\mu+}$, $MSE_{\mu-}$, $MSE_{\sigma+}$ and $MSE_{\sigma-}$.

(4) Let MSE_{min} be the minimum of MSE_{C+}, MSE_{C-}, $MSE_{\mu+}$, $MSE_{\mu-}$, $MSE_{\sigma+}$ and $MSE_{\sigma-}$. If $MSE_{min}<$ MSE(C, μ, σ), update the parameter of (C, μ, σ) to the one where MSE_{min} comes from and let F=0. Otherwise, let $\Delta C=\Delta C/2$, $\Delta\mu=\Delta\mu/2$, $\Delta\sigma=\Delta\sigma/2$ and $F=F+1$.

(5) $I=I+1$. If $I>I_{max}$ or $F>F_{max}$, finish. Otherwise go to step (2) and repeat.

4.4. Hierarchical perturbation model to for emphasis generation

Fig.1 shows the diagram of the proposed hierarchical perturbation model with a two-pass decision tree and a set of SVMs. Firstly, the questions at word and syllable layer are used to construct the main tree, and maximum f0 (*Max*), minimum f0 (*Min*) and energy (*E*) are used for distance calculation. The data in each leaf node of the main tree are used to train a SVM for predicting the variation ratios of *Max*, *Min* and *E* from neutral speech to emphatic speech. Secondly, the questions at phone layer are used to expand the leaves of the main tree, while the mean f0 (*Mean*) and duration (*D*) are used for distance calculation. This second pass constructs a set of sub trees. The data in each leaf node of the sub tree are used to train a SVM for predicting the variation ratios of *Mean* and *D* from neutral speech to emphatic speech.

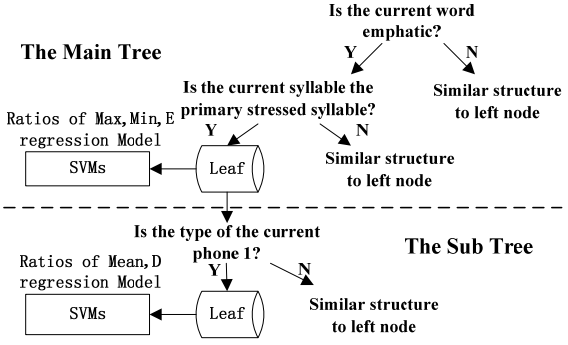

Fig. 1 The diagram of the two-pass decision tree and SVMs

5. Realization of hierarchical perturbation model based on STRAIGHT

The perturbation is realized by STRAIGHT, which is developed by Kawahara et al [14]. To generate emphasis, the variation ratios of the acoustic features are predicted by the two-pass decision tree and SVMs. The energy of the input neutral speech is first modified according to the predicted variation ratios for each syllable. Thereafter STRAIGHT is used to extract f0s. And then f0s and duration are modified at word, syllable and phone layer according to the perturbation model. And finally STRAIGHT is used to generate the emphatic speech.

Step 1: Energy modification at syllable layer: Assume that there are N syllables in the neutral speech. Let $S_i(n)$ be the waveform of the ith syllable, which begins at time step b_i and ends at time step e_i, and $S_i^{'}(n)$ be the ith syllable waveform of the target speech. Let $R_{\text{energy},i}$ be the energy perturbation ratio of the ith syllable. Then the energy of $S_i(n)$ is adjusted with $R_{\text{energy},i}$ and further smoothed by Hamming window $H_i(n)$ of which window length is L, window shift is $L/2$.

$$S'_{i,k}(n) = S_i(n) H_{i,k}(n) R_{\text{energy},i}, k \in \left[0, 2\left\lfloor \frac{e_i - b_i}{L} \right\rfloor\right] \quad (5)$$

$$H_{i,k}(n) = \begin{cases} 0.54 - 0.46\cos\left(\dfrac{2\pi(n - b_i - kL/2)}{e_i - b_i}\right), n \in \left[b_i + kL/2, b_i + (k/2+1)L\right] \\ 0, n \notin \left[b_i + kL/2, b_i + (k/2+1)L\right] \end{cases} \quad (6)$$

$$S_i^{'}(n) = \sum_{k=0}^{2\left\lfloor \frac{e_i - b_i}{L} \right\rfloor} S_{i,k}^{'}(n), n \in [b_i, e_i] \quad (7)$$

The waves of the syllables are concatenated to get the speech with energy adjusted.

$$\mathbf{S}^{'}(n) = \left\{\mathbf{S}_1^{'}(n), \ldots, \mathbf{S}_i^{'}(n), \ldots, \mathbf{S}_N^{'}(n)\right\} \quad (8)$$

Step 2: Parameter extraction: We use STRAIGHT to extract f0s. Let $\mathbf{W}_i(n)$, $\mathbf{P}_i(n)$ and $\mathbf{D}_i(n)$ be the spectrum, f0 vector and corresponding time vector of the ith syllable, from b_i to e_i. Let $P_{\text{Max},i}$ and $P_{\text{Min},i}$ be the maximum and minimum of f0; $R_{\text{Max},i}$, and $R_{\text{Max},i}$ be the perturbation ratio of maximum and minimum f0 from SVMs. We don't modify the spectrum. Hence:

$$\mathbf{W}'_i(n) = \mathbf{W}_i(n) \quad (9)$$

Step 3: Feature modification at syllable layer: The target f0 vector $\mathbf{P}_i^{'}(n)$ are calculated as formula 10-12:

$$P_{\text{Min},i}^{'} = P_{\text{Min},i} \times R_{\text{Min},i} \quad (10)$$

$$P_{\text{Max},i}^{'} = P_{\text{Max},i} \times R_{\text{Max},i} \quad (11)$$

$$\mathbf{P}'_i(n) = P'_{\text{Min},i} + \frac{P'_{\text{Max},i} - P'_{\text{Min},i}}{P_{\text{Max},i} - P_{\text{Min},i}} \times \left(\mathbf{P}_i(n) - P_{\text{Min},i}\right), n \in [b_i, e_i] \quad (12)$$

Step 4: Feature modification at phone layer: Let $S^{'}(n)$ consist of $N^{'}$ phones, from b_j to e_j, $\mathbf{P}_j^{'}(n)$ be the f0 vector of jth phone and $\mathbf{D}_j^{'}(n)$ be the corresponding time vector. Let $P_{\text{mean},j}$ be the mean f0 of jth phone before the modification of maximum and minimum f0 and $P^{'}_{mean,j}$ be the mean f0 after the modification of maximum and minimum f0; and $R_{\text{Mean},j}$, and $R_{\text{D},j}$ be the perturbation ratio of mean f0 and duration from SVMs. The f0 and the duration are computed as formula 13-14:

$$\mathbf{P}''_j(n) = \mathbf{P}'_j(n) - P'_{\text{Mean},j} + P_{\text{Mean},j} \times R_{\text{Mean},j}, n \in [b_j, e_j] \quad (13)$$

$$\mathbf{D}''_j(n) = \mathbf{D}'_j(n) \times R_{\text{D},j}, n \in [b_j, e_j] \quad (14)$$

Step 5: Generating waveforms for phones: Let $S_j^{''}(n^{'})$ be waveforms of jth phone. STRAIGHT is used to generate the modified waveforms with the target f0 vector and duration vector.

$$S''_j(n') = \mathrm{f}\left(\mathbf{W}'_j(n), \mathbf{P}''_j(n), \mathbf{D}''_j(n)\right), n \in [b_j, e_j], n' \in [b'_j, e'] \quad (15)$$

where $\mathbf{f}(\bullet)$ is the generation algorithm.

Step 6: Generating entire emphatic speech: Finally, the whole emphatic speech is generated by concatenating the waveforms of $N^{'}$ modified phones.

$$\mathbf{S}^{''}(n) = \left\{\mathbf{S}_1^{''}(n), \ldots, \mathbf{S}_i^{''}(n), \ldots, \mathbf{S}_N^{''}(n)\right\} \quad (16)$$

6. Experiments and discussion

6.1. Experiment on model prediction accuracy

Mean squared errors (MSE) are used to evaluate the prediction accuracy of the models. Three models are compared in the experiment.

- S1: SVM models are trained directly with all features as training instances and the ratios of the features from neutral speech to emphatic speech as the target labels.
- S2: The data is firstly clustered by the two-pass decision tree detailed in section 4.2 and one SVM is trained for each leaf of the tree without initial parameter optimization.
- S3: The data is firstly clustered by the two-pass decision tree and one SVM is trained for each leaf of the tree using the optimizing method detailed in section 4.3.

Table 5 shows the MSE of the three models in the testing set. The MSE of S2 is 5.4% lower than that of S1 and the MSE of S3 is 9.5% lower than that of S1. This indicates that clustering the data with two-pass decision tree before SVM modeling can increase the accuracy. The cross validation and gradient descent

algorithm for initial parameter optimization of SVMs further improves the accuracy.

Table 5 The MSE of different models

Systems	S1	S2	S3
MSE	0.0297	0.0281	0.0269

6.2. Subjective experiments on generating emphasis with hierarchical perturbation model

To evaluate the performance of the proposed hierarchical perturbation model, two subjective perceptual experiments were conducted by comparing the following three systems:

- A1: Natural speech recordings with expressive intonation to highlight emphasized words in the sentence;
- A2: The system using non-hierarchical model to generate emphatic speech based on the rules of our previous work [12] from neutral speech recording;
- A3: The system using the proposed hierarchical perturbation model to generate emphatic speech from neutral speech recording.

6.2.1. Experiment on emphasis intensity

This experiment was conducted to evaluate if the methods can properly generate emphatic speech. 10 natural recorded emphatic speech files were directly selected from A1. The corresponding 10 neutral speech recordings were provided to A2 and A3 to generate emphatic speech files respectively. Each recorded or generated speech file contains one or more emphatic words. The speech files together with corresponding text prompts with emphasis annotation were provided to the subjects. After listening, the subjects were asked to give the order of the 3 speech files according to the perceived intensity of the emphatic words. Equality is permitted if it is difficult to distinguish the intensity of the emphatic words between two or three speech files.

15 subjects participated in the experiment, and results are shown in Fig. 2. "A1>=A3" represents that the emphasis intensity of 64% speech files from A1 is considered to be higher than or equal to that of the speech files from A3, while the emphasis intensity of 36% speech files from A3 is higher than that from A1. This indicates the emphasis intensity of the speech files generated by our hierarchical perturbation model is comparable to the natural speech recordings. It should be noted that most emphatic words whose emphasis intensity from A3 is considered to be better than that of A1 are at the end of speech utterance. This is due to the pitch declination in natural speech, which causes the decrease of the emphasis intensity at the end of speech utterance. Furthermore, the emphasis intensity of 72% samples from A3 are equal to A2, while 23% are higher, indicating the proposed hierarchical model is better than non-hierarchical model for generating emphasis.

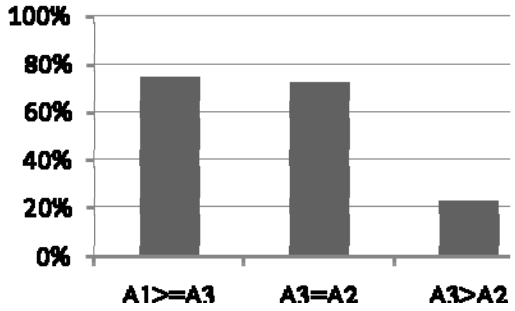

Fig. 2 The result of the experiment of the emphasis intensity

6.2.2. Experiment on naturalness

This experiment was conducted to evaluate if the methods can generate speech with proper naturalness. Another 10 natural recording emphatic speech files were directly selected from A1. The corresponding 10 neutral speech recordings were provided to A2 and A3 to generate emphatic speech files respectively. The speech files together with corresponding text prompts with emphasis annotation were provided to the subjects, and asked them to give the order of the 3 speech files according to the naturalness of the speeches. Again, equality is permitted if it is difficult to distinguish the naturalness between the two or three sentences. The ordering results were converted to a score representing the naturalness according to the following criterion:

If $x>y>z$, then $S(x)=5$, $S(y)=3$, $S(z)=1$;
If $x=y>z$, then $S(x)=5$, $S(y)=5$, $S(z)=1$;
If $x>y=z$, then $S(x)=5$, $S(y)=3$, $S(z)=3$;
If $x=y=z$, then $S(x)=5$, $S(y)=5$, $S(z)=5$.

where x, y, z can be any one of the system A1, A2 or A2, and $S(.)$ is the naturalness score.

The same 15 subjects participated in the experiment of naturalness. The average naturalness score for each system is calculated and shown in Table 6. The naturalness of the generated speeches from A3 is higher than that of A2. This is because more acoustic features are modeled in the hierarchical emphasis model, making the prosody of the speeches generated by the hierarchical model more close to the natural speech.

Table 6 The result of the experiment of the naturalness

Systems	A1	A2	A3

Scores	4.6	2.1	3.3

7. Conclusion and future work

The synthesis of emphasis in English should be realized at word, syllable and phone layers. This paper presents a hierarchical perturbation model to generate emphasis from the neutral speech. The model generates emphasis by hierarchically modifying the acoustic features of the input neutral speech at word, syllable and phone layers using a two-pass decision tree and a set of SVMs. To construct the decision tree, a set of emphasis-related questions are designed according to the hierarchical characteristics for generating emphasis. The questions related to word and syllable layer are used to construct the main tree. The questions related to phone layer are then used to expand the leaves of the main tree which leads to a set of sub trees. Each sub tree corresponds to one leaf of the main tree. A set of SVMs are then built for all the leaves of both the main tree and sub trees. For the main tree, SVMs are used to predict variations of maximum f0 (*Max*), minimum f0 (*Min*) and energy (*E*) from neutral to emphatic speech at word and syllable layer. For the sub trees, SVMs are used to predict variations of mean f0 (*Mean*) and duration (*D*) at phone layer. Cross validation and gradient descent algorithm are used to estimate the parameters of SVMs. Experiments show that the proposed hierarchical perturbation model can generate emphasis speech with high emphasis quality and naturalness.

Future work will incorporate this hierarchical model into a CAPT platform to provide corrective feedback to the user by synthesizing emphasis for important speech segments.

8. Acknowledgment

This 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).

9. References

[1] A. Neri, C. Cucchiarini, H. Strik, "ASR-based corrective feedback on pronunciation: Does it really work?" *Proc. of Interspeech*, 1982-1985, 2006.

[2] H.H. Rump, R. Collier, "Focus conditions and the prominence of pitch-accented syllables," *Language and Speech*, 39: 1-27, 1996.

[3] Y.J. Wang, M. Chu, L. He, "An experimental study on the distribution of the focus-related and semantic accent in Chinese," *Chinese Teaching in the World*, 2006.

[4] S.W. Chen, B. Wang, Y. Xu, "Closely related languages, different ways of realizing focus," *Proc. of Interspeech*, 1007-1010, 2009.

[5] F. Liu, "Single vs. double focus in English statements and yes/no questions," *Proc. of Speech Prosody*, 2010.

[6] P.A. Barbosa, P. Arantes, L.S. Silveira, "Unifying stress shift and secondary stress phenomena with a dynamical systems rhythm rule," *Proc of Speech Prosody*, 49-52, 2004.

[7] K. de Jong, "Stress, lexical focus, and segmental focus in English: Patterns of variation in vowel duration," *Journal of Phonetics*, 32(4): 493-516, 2004.

[8] F. Costa, "Intrinsic prosodic properties of stressed vowels in European Portuguese," *Proc. of Speech Prosody*, 53-56, 2004.

[9] Z. Inanoglu, S. Young, "Intonation modeling and adaptation for emotional prosody generation," *Proc. of ACII*, 186-293, 2005

[10] C.H. Wu, C.C. Hsia, C.H. Lee, M.C. Lin, "Hierarchical prosody conversion using regression-based clustering for emotional synthesis," *IEEE Trans. Audio, Speech and Language Processing*, 18(6): 1394-1405, 2010.

[11] C. Veaux, X. Rodet, "Intonation conversion from neutral to expressive speech", *Proc. of Interspeech*, 2765-2768, 2011.

[12] F.B. Meng, H. Meng, Z.Y. Wu, L.H. Cai, "Synthesizing expressive speech to convey focus using a perturbation model for computer aided pronunciation training," *Proc. of L2WS*, 2010.

[13] The Festival speech synthesis system, http://www.cstr.ed.ac.uk/projects/festival/

[14] H. Kawahara, I. Masuda-Katsuse, A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, 27(3–4):187–207, 1999.

[15] B. Scholkipf, A.J. Smola, R. Williamson, P. Bartlett, "New support vector algorithms," *NeuroCOLT2 Technical Reports Series*: NC2-TR-1998-031, 1998.