Synthesizing Expressive Speech to Convey Focus using a Perturbation Model for Computer-Aided Pronunciation Training

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

[1]Key Laboratory of Pervasive Computing, Ministry of Education
Tsinghua National Laboratory for Information Science and Technology (TNList)
Department of Computer Science and Technology
Tsinghua University, 100084 Beijing, 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, 518055 Shenzhen, China
mfb03@mails.tsinghua.edu.cn, {hmmeng, zywu}@se.cuhk.edu.hk, clh-dcs@tsinghua.edu.cn

Abstract

We present a perturbation model that can modify the acoustic features of neutral speech in order to synthesize focus for certain words. In doing so, we can generate expressive speech output that highlights important speech segments to attract the listener's attention. The ultimate objective is to synthesize corrective feedback in a computer-aided pronunciation training (CAPT) system. This work involves the design and collection of a speech corpus, whose text prompts contain focus words. Each prompt is recorded twice – a neutral production followed by an expressive one where specific words are highlighted with focus. The phones in these recordings are modeled in six different classes, based on their relations with stressed syllables in focus words. Phone boundaries are obtained automatically by forced alignment with an automatic speech recognizer. Acoustic features of the phones, relating to f0, energy and duration, are extracted. Features that have highest correlation with the phone classes, as well as low variances, are incorporated into the perturbation model. The model is applied to neutral recordings of 20 test sentences. Results from a listening test show that the 13 subjects can identify the focus words with an accuracy of over 98%. The perceived degree of focus in the identified words achieves a mean score of 4.5 in a five-point Likert scale.

Index Terms: focus, expressive, computer-aided pronunciation training

1. Introduction

Computer-Aided Pronunciation Training (CAPT) uses speech technologies to facilitate pronunciation training for language learners. Pronunciation training may contain two aspects, including *perceptual training* that develops learner's skills to discriminate different sounds of the language as well as *productive training* that elicits speech from learner and provides feedback on the pronunciation. Studies suggested that discriminative perceptual training is able to improve the production of the phones [1, 2, 3], and the availability of corrective feedback is very effective in reducing pronunciation errors [4].

Our long term goal is to provide *corrective feedback* in CAPT using speech synthesis technologies. As an initial step towards this goal, this work targets the main communicative function of *focus*, which is supported primarily by prosodic features. The focus words will be segments in the generated speech responses that should draw the attention of the learner, constituting corrective feedback from the system.

Previous research has shown that changes in pitch and phone durations contribute much to the expression and perception of focus [5, 6]. Compared with neutral speech, the pitch and intensity of focus words generally increase, while the same features of post-focus words tend to decrease in some languages [7]. Costa [6] analyzed the pitch and durations of vowels and consonants from neutral speech, in comparison with focused speech. He found that the durations of high vowels were shorter than for low vowels, and the pitch values were higher. Barbosa et al. [8] analyzed the durations considering the distance between different phones and focus words. Results show that the closer the phone is to focus word, the longer is the duration. Liu [9] investigated the acoustic realization of single versus double focus in statements and yes/no questions in American English. She found that post-focus pitch range suppression occurs in both single- and double-focused statements.

In this study, we attempt to analyze the prosodic features of phones, based on their relative locations with respective to the stressed syllables of focus words. We believe that such analysis will help us identify the necessary parameters that can be used to perturb the acoustic features of neutral speech for transformation into expressive speech that convey focus. We hope to incorporate such a synthesis technique in automatic feedback generation on a CAPT platform.

The rest of this paper is organized as follows: Section 2 presents a corpus that is designed with contrastive neutral and expressive recordings to support our experimentation. It is divided into the training and testing sets. Section 3 describes the analysis of acoustic features relating to focus, based on the training set. Section 4 details the parameters of our perturbation model, where selected acoustic features from neutral speech input are modified to generate expressive speech output. Section 5 describes our experimental design and a perceptual evaluation of the outputs of the perturbation model. Finally, Section 6 lays out our conclusions and possible future directions.

2. Corpus

2.1. Design of text prompts

We designed a set of text prompts (51 in all), each containing two focus words, which include both monosyllabic and polysyllabic words. These two words are also located at different places in the sentence. Hence, our text prompts cover phones that are in the stressed and unstressed syllables of the focus words, as well as phones that lie before or after,

near or far away from the focus words. The text prompts also include declarative and interrogative sentences to cover different intonations, such as (with focus words in boldface): *"Fighting **thirst** is the **first** thing to be done in this country."* and *"**How** large is the **hall** in the school?"*

2.2. Speech recordings

Two utterances are recorded for each text prompt – one with neutral intonation throughout the utterance and the other with expressive intonation to convey the location of the focus words in the sentence. A female speaker with a high level of English proficiency was invited to record in a studio. Hence we have 102 recorded utterances, saved in the wav format as sound files (16 bit mono, sampled at 16 kHz). Phone boundaries are located automatically by means of forced alignment with an automatic speech recognizer that is trained on the TIMIT database [13]. Pitch tracking is done by Praat [10]. Smoothing is performed in the f0 trajectory and phone segments with obvious errors (amounting to about 5%) are excluded from subsequent analysis.

We randomly select 82 utterances in the corpus for training and the remaining 20 is used for testing. We conducted careful analysis of the training data to select parameters for the perturbation model, as described in the following.

3. Acoustic analysis of focus features

3.1. Classification of phones of corpus

To analyze acoustic features relating to focus, we categorized the phones in the speech corpus into 6 classes, based on the location of the phone in relation with the nearest focus word and its stressed syllable(s). The classes include:

- For a focus word with a syllable carrying primary stress:
  Class 1: Phones in the stressed syllable
  Class 2: Phones before the stressed syllable
  Class 3: Phones after the stressed syllable
- For words without focus:
  Class 4: Phones in the word before the focus word
  Class 5: Phones in the word after the focus word
  Class 6: All other (remaining) phones.

A phone is assigned the class with the lowest class number if it falls into more than one class. Figure 1 illustrates this method of phone classification. "Peterson" and "occasion" are the focus words in the sentence.

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

Figure 1: *An example of phone classification based on the location of stressed syllables in focus words.*

3.2. Extraction of acoustic features

Our objective is to analyze how focus words are realized in the acoustic speech signal. Acoustic features that are commonly associated with prosody include fundamental frequency (f0), intensity and speaking rate. We choose to extract the following acoustic features to capture focus:

- 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 recordings (neutral versus expressive) of each prompt. We compute the ratio (in %) between the measurements of the corresponding expressive and neutral phone units, and variances of the ratios.

3.3. Analysis of acoustic features of focus

This section provides an analysis of acoustic features for each of the 6 classes of phones described in section 3.1. Recall that the classification is related to the location of phones relative to focus words. Let $F_{i,\,\mathrm{neu}}$ be the value of one feature for the ith phone in neutral speech and $F_{i,\mathrm{focus}}$ be its counterpart in focus speech. Let n be the number of the phones in the phone class (i.e. with a total of 6 classes as listed in section 3.1). The average ratio of this feature F for the class is calculated as shown in Equation 1. As will be seen, a general rule of thumb adopted in this study is that a change in ratio of over 5% will be regarded as a major change.

$$R = \frac{1}{n}\sum_{i}\frac{F_{i,\mathrm{focus}}}{F_{i,\mathrm{neu}}} \qquad (1)$$

- **Class 1** (phones in the stressed syllables of focus words): We have 135 phones for class 1 in neutral speech recordings and correspondingly the same number in expressive speech recordings, leading to 270 phones in all. Table 1 shows their comparative acoustic measurements. Both voiced (V) and unvoiced (U) phones are analyzed separately.

For voiced phones, the maximum f0 increases substantially as we go from neutral speech (Neu) to expressive speech (Foc). However, the f0 minimum and energy remain largely the same. The slope and duration both increase substantially. The variances of the ratios of f0 range and f0 slope are relatively large.

For unvoiced phones, the energy remains largely the same, while the duration is lengthened.

The above relates to the overall change in stressed syllables of focus words – a consistent f0 minimum (cf. the neutral case) leads to an increased f0 slope, together with an increased f0 maximum. This is accompanied by a longer duration, all to convey the presence of focus.

Table 1. *Changes in acoustic features between neutral and expressive speech, based on phones (Class 1) in the stressed syllable(s) of focus word(s). Ratio (%) denotes the average ratio between expressive and neutral speech. Var denotes the variances of the ratios.*

		Max	Min	R	Mean	S	E	D
V	Neu	236	201	35	215	282	65	140
	Foc	259	194	65	222	405	68	220
	Ratio (%)	111	97	271	103	350	104	150
	Var	0.02	0.02	5.12	0.01	91.24	0.00	0.13
U	Neu	-	-	-	-	-	51	99
	Foc	-	-	-	-	-	50	127
	Ratio (%)	-	-	-	-	-	96	141
	Var	-	-	-	-	-	0.02	0.53

- **Class 2** (phones before stressed syllables of focus words): There are a total of 348 phones in this class, across both neutral and expressive recordings. Table 2 shows that for voiced (V) phones, the f0 maximum increases and its slope decrease, the energy remains largely stable and duration is lengthened. For unvoiced (U) phones, the energy decreases and durations are lengthened. Most of the phones in this class belong to unstressed syllables, e.g. the first syllable of "apartment". The speaker has a tendency to lengthen its duration to highlight the subsequent stressed syllable.

- **Class 3** (phones after stressed syllables of focus words*)*: There are in total 400 phones in this class. Table 3 shows that the almost all measurements in increase as we migrate from neutral to focused speech, especially for f0 maximum, range and slope for voiced phones, as well as durations for all phones.

Table 2. *Changes in acoustic features between neutral and expressive speech, based on phones (Class 2) before the stressed syllable(s) of focus word(s).*

		Max	Min	R	Mean	S	E	D
V	Neu	233	212	21	222	479	55	133
	Foc	223	207	16	213	332	56	187
	Ratio (%)	95	98	2.29	96	92	102	153
	Var	0.32	0.04	38.70	0.03	39.09	0.01	0.39
U	Neu	-	-	-	-	-	56	95
	Foc	-	-	-	-	-	55	117
	Ratio (%)	-	-	-	-	-	91	120
	Var	-	-	-	-	-	0.02	0.48

Table 3. *Changes in acoustic features between neutral and expressive speech, based on phones (Class 3) after the stressed syllable(s) of focus word(s).*

		Max	Min	R	Mean	S	E	D
V	Neu	236	218	18	227	363	57	104
	Foc	242	206	36	222	379	59	142
	Ratio (%)	108	104	284	104	228	104	118
	Var	0.04	0.04	18.49	0.03	34.84	0.00	0.86
U	Neu	-	-	-	-	-	53	123
	Foc	-	-	-	-	-	49	160
	Ratio (%)	-	-	-	-	-	99	172
	Var	-	-	-	-	-	0.00	3.39

- Class 4 to 6 (phones in words without focus): These classes include phones in words that are in the immediate vicinity (i.e. before or after) of focus words, as well as in other positions.

Table 4. *Changes in acoustic features (ratios) between neutral and expressive speech, based on phones in words that precede (Class 4) and follow (Class 5) focus words, as well as words in other locations (Class 6). Average pause durations before and after focus words increase dramatically as we move from neutral to expressive speech.*

Class		Max	Min	R	Mean	S	E	D
4	V (%)	98	97	172	97	119	101	121
	U (%)	-	-	-	-	-	100	97
5	V (%)	95	95	102	94	97	101	112
	U (%)	-	-	-	-	-	93	126
6	V (%)	97	96	138	96	179	100	103
	U (%)	-	-	-	-	-	101	104
Pause (%)		-	-	-	-	-	-	902

There are a total of 750 phones in Class 4 (i.e. words preceding focus words). Major changes include increases in f0 range (172%), slope (119%) and durations (121%) of voiced phones.

There are a total of 550 phones in Class 5 (i.e. words following focus words). We observe that the f0 maximum, minimum and mean both decrease to 95%, which seems to indicate post-focal pitch compression that is mentioned in previous work [9]. Unvoiced phones also increase in duration (126%).

There are a total of 1126 phones in Class 6 (i.e. all other words without focus). Major changes include increase in f0 range and slope, 138% and 179% respectively.

In addition, we also observe clear lengthening of pause durations before and after focus words. This amounts to a ratio of 902% in comparison with corresponding pause durations in neutral speech.

3.4. Feature selection

To examine the validity of our method of phone classification, we compute the correlation between the acoustic features and the phone classes. Let R_i be the ratio of one feature of the ith phone. Let C_i be the class number of the ith phone ($1 \leqq C_i \leqq 6$). Let n be the number of the phones. Then the correlation is calculated as shown in Equation 2.

$$\frac{\sum_{i=1}^{n}(R_i C_i) - \frac{\sum_{i=1}^{n} R_i \sum_{i=1}^{n} C_i}{n}}{\sqrt{(\sum_{i=1}^{n} R_i^2 - \frac{\left(\sum_{i=1}^{n} R_i\right)^2}{n})(\sum_{i=1}^{n} C_i^2 - \frac{\left(\sum_{i=1}^{n} C_i\right)^2}{n})}} \quad (2)$$

As shown in Table 5, the f0 maximum, duration and energy have high negative correlations with the phone classes. The correlation between f0 minimum and the phone class is low. We also observe large variances in the ratios of f0 slope and range. Hence, as we develop the perturbation model that aims to convert neutral speech to expressive speech carrying focus words, we decide to include four model parameters – namely, the ratios of f0 maximum and minimum, duration and energy.

Table 5. *The correlation between acoustic features and phone classes for voiced phones.*

Features	Max	Min	R	Mean	S	E	D
Relevance	-0.25	-0.03	-0.17	-0.16	-0.13	-0.22	-0.35

4. Perturbation model to synthesize focus

Table 6 shows the parameters of the perturbation model, which are used to convert the acoustic measurements from neutral speech to synthesize expressive speech conveying focus. The f0 measures and energy of class 5 are reduced, to demonstrate post-focus suppression.

Table 6. *Parameters of the perturbation model that transforms neutral speech to expressive speech carrying focus.*

Class		Max (%)	Min (%)	D (%)	E (%)
1	V	110	97	150	104
	U	-	-	113	95
2	V	96	97	139	103
	U	-	-	116	98
3	V	103	93	153	98
	U	-	-	129	104
4	V	97	96	119	91
	U	-	-	96	94
5	V	90	90	112	93
	U	-	-	131	91
6	V	92	91	101	93
	U	-	-	98	90
Pause		-	-	867	-

4.1. Realization based on STRAIGHT

Perturbation is realized by STRAIGHT, which is developed by Kawahara et al. [11, 12], which uses pitch-adaptive spectral analysis combined with a surface reconstruction method in the time-frequency region, together with an excitation source designed based on phase manipulation. We realize the perturbation model through four steps:

1. Modification of energy: Let $\mathbf{s}_i(n)$ be the ith phoneme waveform of the neutral speech, which begins at time step b_i and ends at time step e_i and $\mathbf{s}'_i(n)$ be the ith phoneme waveform of the target speech. Let R_{Energy} be the perturbation ratio for energy. Then energy of $\mathbf{s}_i(n)$ is adjusted by scaling with R_{Energy}. $\mathbf{s}'_i(n)$ is further smoothed by a Hamming window $\mathbf{W}_i(n)$ [14].

$$\mathbf{S}'_i(n) = \mathbf{S}_i(n) R_{energy} \mathbf{W}_i(n), n \epsilon [b_i, e_i] \quad (3)$$

$$\mathbf{W}_i(n) = 0.54 - 0.46 \cos\left(\frac{2\pi(n - b_i)}{e_i - b_i}\right), n \epsilon [b_i, e_i] \quad (4)$$

2. Modification of f0 maximum and minimum: Let $\mathbf{P}_i(n)$ be the pitch sequence of the ith phoneme. Let $\mathbf{D}_i(n)$ be the corresponding time sequence. Let $P_{\text{Max},i}$ and $P_{\text{Min},i}$ be the f0 maximum and f0 minimum of the phone; R_{Max}, R_{Min} and R_{Duration}

be the perturbation ratios for f0 maximum, f0 minimum and duration respectively. Then the f0 sequence $\mathbf{P}_i^{'}(n)$ and its time sequence $\mathbf{D}_i^{'}(n)$ of the focused speech are calculated as shown in Equations (5)-(8).

$$P_{\mathrm{Min},i}^{'} = P_{\mathrm{Min},i} \times R_{\mathrm{Min}} \tag{5}$$

$$P_{\mathrm{Max},i}^{'} = P_{\mathrm{Max},i} \times R_{\mathrm{Max}} \tag{6}$$

$$\mathbf{P}_i^{'}(n) = P_{\mathrm{Min},i}^{'} + \frac{P_{\mathrm{Max},i}^{'} - P_{\mathrm{Min},i}^{'}}{P_{\mathrm{Max},i} - P_{\mathrm{Min},i}} \times \left(\mathbf{P}_i(n) - P_{\mathrm{Min},i}\right), n\epsilon\left[b_i, e_i\right] \tag{7}$$

$$\mathbf{D}_i^{'}(n) = \mathbf{D}_i(n) \times R_{\mathrm{Duration}}, n\epsilon\left[b_i, e_i\right] \tag{8}$$

3. Modification of neutral speech: Let $\mathbf{S}_i^{''}(n^{'})$ be the expressive speech. After obtaining the f0 sequence from the perturbation model, together with the corresponding time sequence, we apply STRAIGHT algorithm to $\mathbf{S}_i^{'}(n)$.

$$\mathbf{S}_i^{''}(n^{'}) = \mathrm{f}\left(\mathbf{S}_i^{'}(n), \mathbf{P}_i^{'}(n), \mathbf{D}_i^{'}(n)\right), n\epsilon\left[b_i, e_i\right], n^{'}\epsilon\left[b_i^{'}, e_i^{'}\right] \tag{9}$$

where $\mathbf{f}(\bullet)$ represents the synthesis process of the STRAIGHT algorithm.

4. Finally, the entire expressive focus speech utterance is generated by concatenating waveforms of the *N* phonemes.

$$\mathbf{S}^{''}(n) = \left\{\mathbf{S}_1^{''}(n), \ldots, \mathbf{S}_i^{''}(n), \ldots, \mathbf{S}_N^{''}(n)\right\} \tag{10}$$

5. Experimental results

The perturbation model is applied to neutral speech recordings from a disjoint set of 20 test sentences (previously mentioned in Section 2.3). We ran a listening test where each subject is shown the raw text of each sentence (with no annotations) as he/she listens to the output of the perturbation model. Each subject is asked to identify the two focus words in each utterance. They are also asked to indicate the degree of focus perceived in each of the identified focus words, based on a five-point Likert scale, i.e.:

'1' (unclear); '2' (slight focus); '3" (focus); '4' (strong focus) and '5' (exaggerated focus)

13 subjects participated in the listening test. The recall of focus words is 98.5% and the precision is 98.8%. We also computed the mean opinion score (MOS) by averaging the five-point Likert scale over all the subjects. The MOS is 4.5 over all the correctly identified focus words and 2.7 over the falsely identified words.

Analysis shows that synthetic focus words tend to achieve higher MOS with higher f0 values. For example, when we analyze the words "table" and "stable" in "*He put the **table** in the **stable***", the average MOS of both focus words are 4.8. On the other hand, when we analyze the words in "stairs" and "more" in "*Please use the **stairs more***", the average MOS of the focus words are 4.2 and 4.1 respectively. We note that the words "table" and "stable" generally have higher f0 than "stairs" and "more". In addition, for the sentence "*The poster outside the **school** is **cool***", the former and latter focus words have average MOS of 3.8 and 4.5 respectively. We observe that no pause exists between "the" and "school", but 55ms pause is found between "is" and "cool". This observation suggests that pause insertions before and after focus words are critical for achieving high MOS.

6. Conclusions and future work

This paper presents a methodology for synthesizing expressive focus in speech. This is achieved by the development of a perturbation model that modifies acoustic features in a neutral speech utterance to generate an expressive utterance that conveys focus in selected words. We designed a corpus where each text prompt contains a couple of focus words. Each prompt is then recorded twice, as contrastive neutral versus expressive recordings. We also model the phones in six different classes, based on their relations to the stressed syllables in focus words. Acoustics features are extracted from these phones, including f0 maximum, f0 minimum, f0 range, mean f0, f0 slope, energy and duration. . Analysis shows that the acoustic features change most markedly at the stressed syllables of focus words. The features considered most descriptive of the phone class are selected based on measures such as correlation and variance. Four features are selected to form the perturbation model, which modifies the neutral speech parameters (f0 maximum, f0 minimum, duration and energy) to generate the expressive speech parameters to convey focus. Results from a perceptual test shows that the listeners are able to identify focus words with an accuracy of over 98.5%. These words also achieve an MOS of 4.5 (based on a 5-point scale)

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

7. Acknowledgements

This work was conducted when the first author was a summer intern in the Human-Computer Communications Laboratory, The Chinese University of Hong Kong (CUHK). We wish to acknowledge the CUHK OALC summer internship program. The work is jointly supported by the research funds from the Hong Kong SAR Government's Research Grants Council (CUHK4161/08), the National Natural Science Foundation of China (60928005, 60805008, 60910130), and the National High Technology Research and Development Program of China (2009AA011905).

8. References

[1] B.L. Rochet, "Perception and Production of L2 Speech Sounds by Adults," in W. Strange [Ed], *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues*, 379-410, Timonium: York Press, 1995.

[2] R. Akahane-Yamada, Y. Tohkura, A.R. Bradlow and D.B. Pisoni, "Does Training in Speech Perception Modify Speech Production?" *Proc. ICSLP*, 606-609, 1996.

[3] V. Hazan, A. Sennema, M. Iba and A. Faulkner, "Effect of Audiovisual Perceptual Training on the Perception and Production of Consonants by Japanese Learners of English," *Speech Communication*, 47(3): 360-378, 2005.

[4] A. Neri, C. Cucchiarini and H. Strik, "ASR-based Corrective Feedback on Pronunciation: Does It Really Work?" *Proc. Interspeech*, 2006.

[5] W. Yunjia, C. Min and H. Lin, "An Experimental Study on the Distribution of the Focus-related and Semantic Accent in Chinese," *Chinese Teaching in the World*, 2, 2006.

[6] F. Costa, "Intrinsic Prosodic Properties of Stressed Vowels in European Portuguese," *Proc. Speech Prosody*, 53-56, 2004.

[7] S.W. Chen, B. Wang and Y. Xu, "Closely Related Languages, Different Ways of Realizing Focus," *Proc. Interspeech*, 2009.

[8] A. Barbosa, P. Arantes and L.S. Silveira, "Unifying Stress Shift and Secondary Stress Phenomena with a Dynamical Systems Rhythm Rule," *Proc. Speech Prosody*, 49-52, 2004.

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

[10] P. Boersma and D. Weenink, "Praat: Doing Phonetics by Computer," 2003.: http://www.praat.org,

[11] H. Kawahara, "Speech Representation and Transformation using Adaptive Interpolation of Weighted Spectrum: Vocoder Revisited", *Proc. ICASSP* 1997.

[12] H. Kawahara, I. Masuda-Katsuse and 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.

[13] H. Meng, Y.Y. Lo, L. Wang and W.Y. Lau. "Deriving Salient Learners' Mispronunciations from Cross-language Phonological Comparisons", *Proc. ASRU*, 2007.

[14] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier-transform," *Proc. IEEE*, 66: 51–83, 1978.