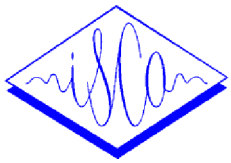

Speech Prosody 2006
Dresden, Germany
May 2-5, 2006

Perceptual Optimization of the Chinese Accent-Index Detector

Weibin Zhu

Institute of Information Science
Beijing Jiaotong University, Beijing 100044 China
zwb@computer.njtu.edu.cn

Abstract

For a TTS system， only if a large size of corpus annotated with AI (Accent Index) is available, could it be practicable to build an AI-supported prosody module in a data-driven method. An approach had been proposed to label Chinese AI automatically. Although preliminary experiments showed its effectiveness and efficiency of the approach, there are still certain issues left unsolved: the evaluation and the optimization of the AI detector. A small size of sub-corpus has been labeled with AI manually, which is expected to be as a reference for evaluating the performance. And a measure CC (Correlative-Coefficient), the CC between the auto-detected and the manual-annotated AI set, is proposed as the criteria for optimizing the detector. Thanks to the use of CC, the detector has not only been refined and optimized, but also the auto-detected AI has been assigned with prosody meaning subjectively.

1. Introduction

Currently, it is still a great challenge to synthesis speech with varying accent for Chinese TTS systems which are based on the unit-selection method [1, 2, 3]. To generate accented speech, a possible approach is to employ a prosody module with the function for accent prediction, and the module could be trained with an AI annotated corpus in the statistical method as well as what has been done in the former system [1, 4]. However, it raises the problem of AI annotation, i.e., how to label the corpus with accent-index more efficiently and effectively while the corpus is normally with a huge size [5].

An automatic detector for Chinese AI was purposed to annotate those recorded speech, by which it was with higher consistency and quicker speed than by human's perception. The main idea of the approach is: at first, a non-AI prosodic module is trained with a BI (Break Index) annotated corpus; second, the prosodic parameters predicted by such a module could be regarded as the ones of the neutral intonation; and then the differences of prosodic parameters between the real speech and the predicted one, should be mainly caused by accents varying. A measure, extracted from those differences, is designed to represent the AI thus [6].

There are two issues left unsolved. The first one is how to evaluate the performance of the detector, and the second one to optimize the detector. A solution is presented in this paper. The general idea is as follows: using the subjective annotation as the reference, a measure CC, the correlative coefficient between subjective results and detected ones, is chosen as the criteria to evaluate the detector. Moreover, the detector could be optimized while adjusting the weights used in the detector to maximize the CC.

The paper is organized as follows: Section 2 introduces the method of Chinese AI detection. Section 3 shows the procedures for the optimization of the detector. And Section 4 presents the examining results. Conclusions are presented in Section 5.

2. Automatically detecting Chinese AI

The term AI refers to a varying degree of those prosody features which are perceived as accented or unaccented. From the linguistic point of view, AI is the relevancy of the semantic emphasis or focus in a sentence. From the acoustic point of view, it is the manifestation as varying prominences of acoustic prosody parameters. The approach of automatic AI detector is to leverage those acoustic parameters and assign each prosody unit, such as syllable or prosody word, with an appropriate AI.

2.1. With AI – to describe prosody more deeply

Considering the application purpose, AI is defined to cross two layers in prosody structure: intonational phrase and prosody word [5], and as follows,

At phrase layer, AI is conveyed by prosody word, and is scaled in 3 levels.

- A2, accented, the highest level of accent, generally corresponding to the semantic focus in the phrase, and perceived as the emphasis and/or prominent part in the whole intonation.
- A1, normal level, generally corresponding to the normal syntactic constitutions in the phrase, which could be perceived as the normal articulation strength.
- A0, lightened, usually corresponding to the adjunct part in the phrase, which could be perceived as the lightened articulation strength.

At prosody word layer, it is word stress pattern with 3 levels,

- S2, stressed, generally corresponding to the most accented syllable in an A2 word.
- S1, normal, it is the accent level between S2 and S0, and could be in a word at any accent level.
- S0, lightened, the syllable with neutral tone in a word at each accent level, or the unstressed syllable in a structured word at any accent level.

2.2. Acoustic realization of AI

The fact that we have to face is that, the acoustic features lying in the surface of speech signal are not only relevant to the accent functions, but also to various pragmatic, emotive functions, and to lexical tones which are the especial ones in Chinese. To identify the accent distribution in a speech signal, what should be understood essentially is that, how accents and other prosodic events, including prosody structure and intonational modality and lexical tone, affect the acoustic features respectively.

In Xu [7], those surface features are recognized as the indirect reflections of the prosody events, which could be

10.21437/SpeechProsody.2006-135

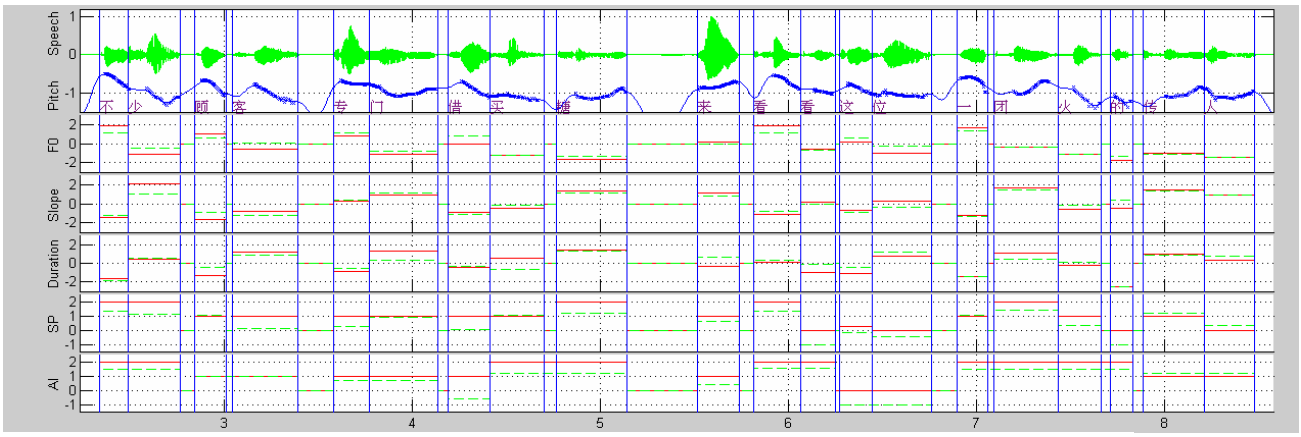

Figure 1, *AI detected with F0, Slope and Duration. Predicted parameters and detected AIs are in green dotted*

identified from, a) articulatory implementation, including articulatory constraints and articulatory strength; b) target assignment, tone and accent target are assigned by separated functional components. According to this method, during articulation, the accents are presented through two aspects mainly:

- Articulatory Strength, the amount of physical effort determines how effectively a pitch target has been implemented, which could be estimated as how sufficiently the tone is approached from the surface feature - F0. There are some adjunctive variances while the strength is changed, including the intensity and the duration varying, and the pause being inserted sometime also.
- Prominence, the differences between the local unit and its neighbors, which could be estimated with the gap of pitch range, the shift of pitch register, the change of rhythm, and the insertion or deletion of pause. Obviously, those differences between the local unit and its neighbors could be equally detected by the comparison between the features of a real speech signal with the ones of a 'neutral' speech.

In study [8], it is shown that there are some other acoustic correlates of accent and/or stress, such as spectral balance. But in this paper, the efforts are focus on the super-segment features, i.e. the prosody parameters.

2.3. AI detecting

According to the analysis on the manifestation of accent, both articulatory strength and prominence are of relative properties. Only compared with the normal or neutral one, could the articulatory strength and the prominence of any prosody unit be determined and distinguished. In the previous work [6], an acoustic prosody predictor was trained with a non-AI annotated corpus [5]. Therefore, the parameters predicted by the prosody model could be regarded as the prosodically acoustical ones of a neutral speech, i.e. the predicted ones are equal to the neutral ones. And moreover, the differences between the real parameters and the 'neutral' ones should be strongly related to AIs, and it is reasonable to retrieve the AIs from the differences.

There are 4 tones in Chinese, and each tone has its particular target, including two static pitch targets [high] and [low], and two dynamic targets [rise] and [fall]. They are associated with the four lexical tones: H (High), L (Low), R (Rising), and F (Falling), respectively. One syllable's articultory strength is mainly presented through the approximation to its target.

For a syllable with tone 1, if its pitch is higher and duration is longer comparing with its 'neutral' reference, reasonably its articulatory strength should be larger. For tone 2, the strength will be enlarged if its pitch slope is sharper than the reference. Therefore, each tone should use its specific criteria to measure the strength.

A syllable's prominence is mainly presented through its pitch register and duration. Except tone 3, if pitch is higher, being compared with 'neutral' one, the syllable should be with larger prominence. The case for tone 3 is converse. For all 4 tones, if duration is enlarged, the syllable should be also with larger prominence.

Considering these facts synthetically, a measure is defined to present syllable's AI,

$$As(S_i) = \sum_j W_{t,j}(Tone(S_i)) \cdot Diff(C_{real,j}(S_i), C_{pred,j}(S_i)) \quad (1)$$

Where $As(S_i)$ is the AI value of the *i-th* syllable S_i of the sentence; $Tone(S_i)$ is the tone function that outputs the tone value of syllable S_i; $C_{pred,j}$ indicates the *j-th* element of predicted parameters and $C_{real,j}$ indicates the *j-th* element of parameters of real speech; *Diff*() is the function that calculates the difference between real and predicted parameters. $W_{t,j}$ refers to the weight for tone *t* and the difference of the *j-th* element of parameters.

Word's AI is defined as:

$$Aw(PW_i) = W_s \bullet Max(As(S(PW_i))) + W_{f0} \bullet Diff(C_{f0}(PW_i)) + W_{dur} \bullet Diff(C_{dur}(PW_i)) \quad (2)$$

Where *Max*() is the function that outputs the largest value of syllable AI of *As*() among all syllables *S*() in the whole word PW_i; W_i refers to the weight for the max syllable AI value; W_{f0} refers to the weight for pitch difference of the word and W_{dur} refers to the weight for duration difference of the word.

3. Optimization of AI detector

Essentially, the term AI refers to a prosody feature in the perception area. Consequently, the most credible determination of AI should be a perceptive result. It is a natural and reasonable thought to use the manually annotated AI as reference, with which to evaluate the performance of the AI detector and optimize the detector itself. Certainly, much closed to the manual-annotated AI, more credible the auto-detected one is.

3.1. CC as the criteria

CC is used to measure the similarity between two sets of AI annotation which are for the same utterance but labeled in different methods. Corresponding to AI definition, CC_sylable and CC_word, are calculated to measure the AI similarity respectively for syllable and word. The equations are as follows:

$$CC\_syllable = \frac{\sum_{i}^{M} (As_1(S_i) - \overline{A}s_1) \cdot (As_2(S_i) - \overline{A}s_2)}{M \cdot \overline{A}s_1 \cdot \overline{A}s_2} \quad (3)$$

$$CC\_word = \frac{\sum_{i}^{N} (Aw_1(W_i) - \overline{A}w_1) \cdot (Aw_2(W_i) - \overline{A}w_2)}{N \cdot \overline{A}w_1 \cdot \overline{A}w_2} \quad (4)$$

where $As(S_i)$ is the AI value of the *i-th* syllable S_i of the sentence; $\overline{A}s$ is the average of all syllables' AI values of the entire sentence. Similarly, $Aw(W_i)$ is the AI value of the *i-th* word W_i of the sentence; $\overline{A}w$ is the average of all words' AI values of the entire sentence. And the sub-index 1 or 2 of *As* or *Aw* indicates the different annotating method.

3.2. Manually labeling AI

All utterances have been annotated with BI, so that AI manually labeling could be implemented on the prosody-structure-annotated data directly. The principle method is 'to label what you hear' as well as what had been used during BI annotating [5].

The perceptive cues for syllable AI judgment:

- Approximation to tone target
- Sufficiency of articulation
- Prominence of pitch register

Word AI labeling is implemented after syllable AI being annotated, and perceptive cues used there are:

- Prominence of pitch register of the whole word
- Max AI value of syllable within the word

60 cells of sentence have been annotated with AI at both syllable and word layer by one phonetician. Since these sentences with AI are used as 'standard' ones, they have been labeled totally for four times successively to ensure the annotations with enough high confidence. CCs between the latest two times AI are listed as follows:

CC_word = 0.78
CC_syllable = 0.81

The consistency between two times of annotation is significantly great, and the latest time one is used as the reference then.

3.3. Function refinement

At the very beginning, in equation (1) and (2), *Diff*() is a linear function, which results in that its output is often over-dominated by those parameters with too prominent value. To constrain the effectiveness of each parameter, a non-linear function *NonLinearTran*() is included here. And then, *Diff*() is changed in the format as follows:

Diff() = *NonLinearTran*(*diff_param*) (5)

diff_param = w_syl(real_param - pred_param)* (6)
*- w_phr*diff_phrase - w_word*diff_word*

Where definition of *NonLinearTran*() is as follows:

aa = NonLinearTran (*bb*) (7)
if bb >= 0, aa = 1.0 - exp(-bb);
else aa = -1.0 + exp(bb);

The response curve of the function *NonLinearTran*() is shown in Fig. 2.

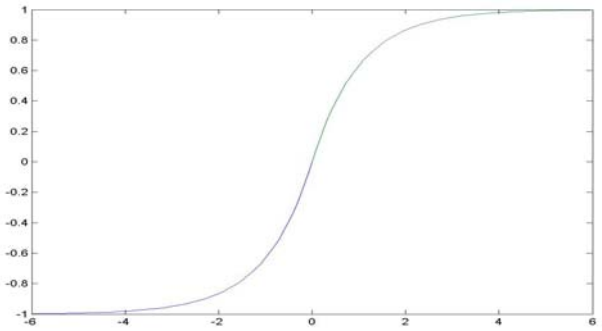

Figure 2, *The response curve of NonLinearTran()*

3.4. Optimizing the weights

Once the parameters of real speech and its corresponding 'neutral' ones are obtained, the value of AI will only depend upon the AI detector's weights. In the previous work [6], those weights were assigned and/or adjusted only according to the observation on several samples of utterance manually. Obviously, this method could not ensure the optimization of the detector.

As what has been mentioned above that manual-annotated AI could be as standard or reference for auto-detected AI, the value of CC between two sets of AI could be as criteria to evaluate the detector's performance. The greater value of the CC, the more close to the manual-annotated AIs for the auto-detected ones, and then, the higher the performance of the detector. Therefore, to enlarge the CC value through adjusting those weights is equal to optimize the detector.

The function, fminsearch, in the Matlab Optimization Toolbox, which searches for the minimum of an unconstrained multivariable function, could be used to optimize the free parameters. In this case, those weights to be adjusted are the free parameters while negative *CC_syllable* and negative *CC_word* are the unconstrained multivariable functions. Whenever the minimum of the functions are found, the largest *CC_syllable* and *CC_word* will be achieved, and those weights will be optimized accordingly.

Since only 60 utterances with AI are available, the number of free parameters in each run of optimization should not be too large. Thus, the optimization is separated into many different runs. In each run, only some of the weights are optimized and the others are fixed at their original values.

The weights could be separated by its scope as follows:

Parameter related
- weight_F0;
- weight_slope;
- weight_duration;

Tone related:
- weight_tone1;
- weight_tone2;
- weight_tone3;
- weight_tone4

Diff_param related
- weight_syllable;
- weight_word;
- weight_phrase;

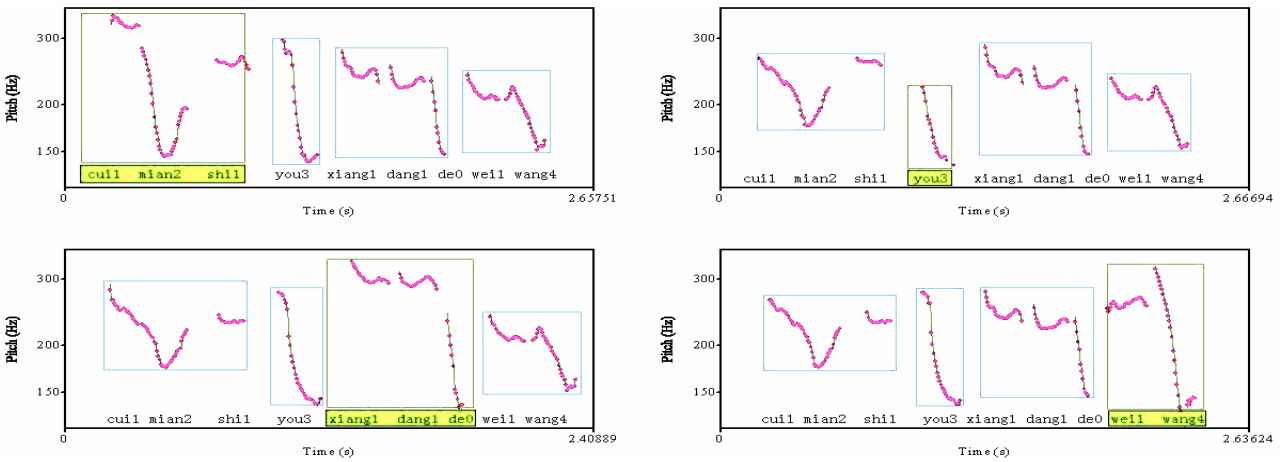

Figure 3, *four pitch contours of synthesized speech with AI varying. The script in Pinyin:* ***cui1 mian2 shi1 you3 xiang1 dang1 de0 wei1 wang4***

Unit related:

weight_As;
weight_word_F0;
weight_word_slope;
weight_word_duration;

Word AI related weights should be optimized after syllable AI related ones have been fixed. Table 1 shows the integrated effects of the weight-optimization and the function-refinement. Both CC_word and CC_sylable are improved significantly.

Table 1. *CC values before/after optimization and refinement*

	initial	latest
CC_word	0.62	0.77
CC_syllable	0.66	0.80

4. Experiments

To evaluate the reasonableness and efficiency of the AI detector, a new prosody parameter predictor has been trained with a 5k cells size of corpus which has been auto-annotated with AI. It replaces the former module in the TTS system, and certainly the TTS system becomes an AI-supported one. In a convincing experiment, manually assigning the AI annotations, the output speech was generated just with expectant accents. The perceptual evaluation showed that the accent manifestation was distinguishable and acceptable.

Fig. 3 shows four pitch contours of synthesized speech with AI varying. The script of the four sentences are all same as 催眠师有相当的威望 (in English: the hypnotist is with considerable prestige. in Pinyin: ***cui1 mian2 shi1 you3 xiang1 dang1 de0 wei1 wang4***). The focus is moved to the next prosody word one by one in these voices, and in each pitch contour the pitch range is enlarged in the area corresponding to the focused prosody word i.e. which with high AI.

5. Conclusions

The measure CC, the correlative-coefficient between the auto-detected and the manual-annotated AI set, is designed as the criteria to evaluate the performance of the AI detector. After optimizing weights used in the detector, the CC_syllable is improved from 0.66 to 0.80 and the CC_word is improved from 0.62 to 0.77. Thanks to the use of CC, the detector has not only been refined and optimized, but also the auto-detected AI is assigned with prosody meaning subjectively, since the auto-detected AIs are far more closed to the manual-annotated one after the procedure of optimizing.

6. Acknowledgement

The most part of the paper work was completed in IBM China Research Lab. The author thanks Shi Qin, Ma Xijun and Zhang Wei in CRL for their support. And I am also grateful to Prof. Cao Jianfen who provided valuable recommendations and suggestions.

7. References

[1] Ma, X., Zhang, W., Zhu, W., et al, "Probability Prosody Model for Unit Selection", *ICASSP 2004*, Montreal, Canada, 2004

[2] Chu Min, Peng Hu and Chang Eric, "A Concatenative Mandarin TTS System without Prosody Model and Prosody Modification", *SSW4* , Pitlochry, Scotland, 2001

[3] Wang Ren-Hua, Ma Zhongke, Li Wei, et al, "A Corpus-based Chinese Speech Synthesis With Contextual Dependent Unit Selection", *ICSLP 2000*, Beijing, China, 2000

[4] Shi, Q., Ma, X., Zhu, W., et al, "Statistic Prosody Structure Prediction Based on Annotated Corpus", *IEEE TTS Workshop 2002*, Santa Monica, USA, 2002

[5] Zhu, W., Shi, Q., et al, "Corpus Building for Data-Driven TTS Systems," *IEEE TTS Workshop 2002*, Santa Monica, USA, 2002

[6] Zhu, W., Zhang W., Shi Q., et al, "Automatic Detection of Chinese Accent-Index Based on Approximation-Ratio", *ISCSLP 2004*, Hong Kong, China, 2004

[7] Xu, Y. "Separation between functional components of tone and intonation and observed F0 patterns", *International Symposium on Tonal Aspects of Languages*, Beijing, China, 2004

[8] Sluijter, A. and van Heuven, V., "Spectral balance as an acoustic correlate of linguistic stress", *JASA. 100(4)*, 1996