ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

Automatic Detection of Chinese Accent-Index Based on Approximation-Ratio

Weibin Zhu, Wei Zhang, Qin Shi, Xijun Ma, and Liqin Shen

IBM China Research Lab, Beijing 100085, China
{zhuweib, zhangzw, shiqin, maxijun, shenlq}@cn.ibm.com

Abstract

For a TTS system, to synthesize speech with better prosody, accent information is expected to be involved besides other. Therefore, we defined a set of Accent Index (AI) to represent the variances of accent in Chinese speech, and proposed a novel method to automatically annotate Chinese speech with AI. In the method, a parameter, named Approximation-Ratio, was used to numerically indicate the accent of prosodic unit. And the value of AI was the discretization of Approximation-Ratio. One corpus was annotated with AI by the method. And with the corpus, a refined prosody parameter prediction model was built. The experiment results showed that prosody parameters predicted by the refined model were more close to ones of real speech than by the former model without AI. Further, a perceptual evaluation showed that the accent manifestation generated by AI-ready synthesizer was distinguishable and acceptable.

1. Introduction

There're two main components in a typical trainable TTS system. Front-end [1] is, corresponding to the input text, to generate the symbolic description about speech, i.e., what to be uttered. Back-end [2] is, based on the symbolic description, to generate acoustic targets, i.e., how to be uttered. Being measured by predicted targets, the best sequence of units could be selected from a large speech database [3], and those units are concatenated to form the appropriate speech. What being related with prosody in both components are, in front-end, to predict the prosodic events based on the text analysis, while in back-end, to generate the prosodic acoustics parameters as the targets for unit-selecting.

In such a system, both prosody event predictor and prosody parameter predictor are trained by corpora annotated with prosody description symbols. Those annotations maybe include prosodic structure, intonational modality, and accent and/or stress. Without the constraint of the accent, only monotonic intonation could be generated by such a system consequently [2]. To synthesize speech with more natural and expressive intonation, accent information should be adopted into above two prosody predictors. Therefore, corpora labeled with AI are needed.

In principle, accent annotation could be implemented by subjects. But to process a large size of corpus manually, human-consuming is a critical problem. Our experience in manual annotation shows that the time-consuming of BI (Break-Index, i.e., prosody structure) labeling is about 10~20 times of speech uttering, and contrastingly the time-consuming of AI labeling is about 40~60 times. For manual labeling, another problem is how to keep the consistence among those subjects, if more people have to be involved to label large corpus annotating. Substitutively, we have to find a method to automatically annotate the accent.

Variously acoustic manifestation of Chinese accent has been investigated [4] [5]. Our effort focuses on the leverage of those surface features lying in speech signal. It is based on the following assumptions: a), those prosodically acoustic features adequately reflect various prosodic events including prosody structure, intonational modality, and accent. b), if it is only constrained with prosody structure and intonational modality, and each unit is with the same AI, the utterance could only be a neutral type of speech, i.e., with neutral prosody. c), for a same utterance's script, those surface features' difference between the vivid speech and the neutral one should be mainly caused by accent varying, or partly at least. Based on those three assumptions, the main idea of our approach is, a), the neutral one could be predicted by a prosody parameter predictor which has been trained with a corpus annotated only with the information of prosody structure and intonational modality. b), AI could be retrieved from the difference between the predicted one and the real one reasonably.

IBM Mandarin TTS Corpus [3] was used to examine the proposed method. Based on the corpus, a neutral prosody parameter predictor was built at first. And then, an acoustic measure, Approximation-Ratio was calculated, based on the differentials between real prosody parameters and neutral (i.e. predicted) ones. The corpus was annotated with AI, i.e., the discrete value of Approximation-Ratio. Eventually, an AI supported prosody parameter predictor was built, and the reasonableness of such AI detector was examined by difference analysis and a perceptual evaluation.

The paper is organized as following: Section 2 introduces the definition of AI. Section 3 shows the procedure of automatic detection of AI. And Section 4 presents the examining tests. Conclusions and discussions are presented in Section 5.

2. Chinese AI, its Definition and Manifestation

2.1. Definition

The term AI refers to the varying degrees of those prosody features which are perceived as accented or unaccented. From the view of linguistic aspect, AI is relevant to the distribution of semantic emphasis or focus in a sentence. From the view of acoustic aspect, it is realized as varying prominences of acoustic prosody parameters. In the space of its realized acoustic features, accent should be varying continuously. But as a set of symbolic transcription, accent must be scaled into several discrete levels. According to the experience of BI definition [3], the set of AI was defined and corresponding to a set of explicit prosody meanings.

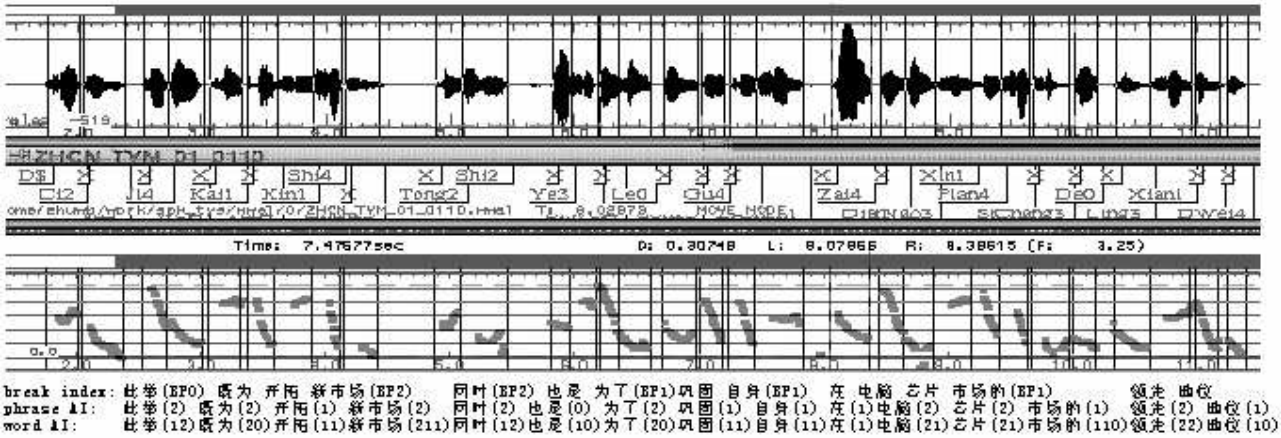

Figure 1. One sample of prosody description with AI

Considering the application purpose, we define that AI crosses two layers in prosody structure: intonational phrase and prosody word. The definition is as following,

At phrase layer, AI is conveyed by prosody word, and is divided into 3 levels in our definition.

- 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 we have to face to is that the total acoustic features in speech signals cover not only accent but also various pragmatic and emotive functions, and lexical tones especially in Chinese. To identify the accent distribution in a speech signal, we should understand how accent and other prosodic events, including prosody structure and intonational modality and lexical tone, affect the acoustic features in general.

In Xu [4], those surface features are the indirect reflections of the prosody events, which could be 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're some adjunctive variances while the strength is changed, including the intensity and the duration varying, and pause being inserted sometime also.
- Prominence, the differentials 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 differentials 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, maybe it is a virtual one.

At the layer of phrase/sentence, the manifestation of different levels of prosody word's accent is specified as below:

- A2, Pitch range is enlarged, while pitch register is abnormal, and articulatory strength is supernormal. Sometime it is additively with inserted pause
- A1, Pitch register is normal, and articulatory strength is normal
- A0, Pitch range is compressed, and/or articulatory strength is weakened

Where, the range of pitch refers to the one of the prosody word; and the register of pitch refers to the one of the word also. Articulatory strength refers to the one of the most stressed syllable in the prosody word.

At the layer of prosody word, the manifestation of stress pattern is listed as below:

- S2, articulatory strength is supernormal.
- S1, articulatory strength is normal
- S0, syllable's pitch curve is implemented as a neutral tone or articulatory strength is weakened.

Figure 1 shows a sentence being annotated with prosody boundary, accent level of word, and stress pattern in word, which was implemented manually.

3. Automatic Detection of AI

3.1. Main idea

The tasks for AI automatic detection is to annotate each utterance in a corpus with AI defined in section 2. The inputs include: speech waveform, phonetic alignment, and prosody structure annotation. We started at quantizing two aspects of accent: articulatory strength and prominence.

According to the analysis on the manifestation of accent, both articulatory strength and prominence are of relative

properties. Only after being compared with the normal one, could one syllable be determined its articulatory strength degree, and could the prominent part be distinguished from the whole utterance. In IBM former TTS system, only information of prosody structure as prosody description features was used to build the acoustic prosody predictor using DT (Decision Tree) method [2]. Given the prosodic structure context of the utterance, a set of prosody parameters can be predicted. Because no AI information were involved when the prosody DT was trained, the integrated influences from varying AIs were to enlarge the statistic variances of those samples clustered in each leaves in the DT, while to keep the statistic means no changed.

It is natural that the intonation represented by the mean of predicted ones could be treated as the 'neutral' one. Because everything else, including phone context and prosody structure, are equal, and all processed sentences are in the same intonation modality: declarative one, the differences between the real parameters and the 'neutral' ones should be strongly related with AIs. So it is reasonable to retrieve the AIs from the differences.

3.2. Approximation-Ratio

There are 4 tones in Mandarin, each one 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.

Its prominence is mainly presented through pitch register and duration. Except tone3, if pitch is higher, being compared with 'neutral' one, the syllable is with larger prominence. The case for tone 3 is converse. For all 4 tones, if duration is enlarged, the syllable is also with larger prominence. Considering these facts synthetically, one parameter, Approximation-Ratio is defined to present one syllable's accent,

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

Where, one syllable S_i Approximation-Ratio $AppRatio(S_i)$ is the sum of the difference $Diff()$ between parameters of real speech $C_{real,j}(S_i)$ and predicted ones $C_{pred,j}(S_i)$, weighted with coefficients $W_{t,j}()$ special for tone t; $Tone(S_i)$ is the tone of syllable S_i, j refer to the j'th elements of prosody vector. And before $Diff()$ is calculated, pitch and duration are normalized at word or phrase level, to eliminate the mean deviation of the utterance. The 'neutral' one is as the reference that is for sure.

3.3. Procedures of AI detection

The implement to detect AI follows the below procedure.

Step 1 Pitch-Smoothing. F0 was processed in way of V. Strom [6], and was calculated for every 10 ms, then was filtered by a low pass filter and then reset in voiced regions and adapted to the linear interpolated F0 in unvoiced regions, totally 5 iterations.

Step 2 Parameterization. Pitch, pitch-slope and duration of syllable were as three elements forming the feature vector, and were transformed into z-scores after log operation. Due to the most appropriate F0 contour of a tone was best approximated in the final portion of a syllable [4], the pitch in feature vector was present by the value of F0 at the middle of the syllable, and the pitch-slope was the log slope of F0 at the final portion of the syllable, while the duration was the log duration of the syllable.

Step 3 Prosody Decision Tree Building. Here, the decision tree was a One-Gaussian target tree which generated the context equivalent clusters of target values. The context information includes: tone context, phone context, and prosody structure information. The feature vectors were the three dimension parameters.

Step 4 Approximation-Ratio Detecting. With the giving context information, The corresponding 'neutral' prosody features of each utterance in the corpus were predicted by the prosody target tree,. Then the Approximation-Ratio of each syllable was calculated, and normalized at word layer and phrase layer respectively. At phrase layer, word Approximation-Ratio was simply set as the max value among those syllables'.

Step 5 Discretization. At word level, each syllable's Approximation-Ratio was discretized into 3 levels. The threshold was determined based on the distribution of Approximation-Ratio, and initially in proportion reported in Wang [5], which were the perceptual counts as listed in Table 1. At phrase level, each prosody word Approximation-Ratio was discretized into 3 levels also, and splitting in proportion listed in Table 2 initially.

Table 1. Proportion of syllable pattern, derived from Wang [5]

Stress	S0	S1	S2
Proportion (%)	10	65	25

Table 2. Proportion of word AI, derived from Wang [5]

Word AI	A0	A1	A2
Proportion (%)	20	45	35

The corpus IBM Mandarin TTS database [3] had been processed, in which 5k cells of sentence were recorded, and were pronounced by a professional female speaker. It was required the speaker to utter each given scripts in a kind of neutral intonation, without intended emotion or semantically specific focus. So, what accents collected in the corpus are mainly the 'default' ones, i.e. they are not semantic accents but syntactic ones.

4. Preliminary Experiments

4.1. Difference analysis

To evaluate the reasonableness and efficiency of the AI detector, a new prosody parameter predictor, a decision tree, was trained with AI information besides other features. 5k

cells of sentence were involved here also. And then, several objective measures were implemented.

1. The distribution of prosody parameters at each leaf in the DT became more compact. Other things being equal, the leaf number of DT with AI was increased than the one of former DT without AI. The leaf number was increased from 418 to 542.
2. The real prosody parameters were compared with the predicted ones. With AI, the predicted prosody parameter was more closed to the real one than without AI, while the variance of the difference between real one and predicted one was reduced about 10~16% relativistically. Table 3 shows the variances of the difference between real ones and predicted ones.

Table 3. the variances of the difference between real prosody parameters and predicted prosody parameters

Parameters	pitch	pitch-slope	duration
AI	0.49	0.43	0.54
Non-AI	0.54	0.51	0.60

4.2. Perceptual evaluation

The former module of prosody parameter predictor in TTS system was replaced with the AI one. We manually adjusted the AI annotations of the test sentences as the input of prosody parameter predictor. Then we could get the synthesized speech with expectant accents which were set manually. One perceptual evaluation showed that the accent manifestation was distinguishable and acceptable.

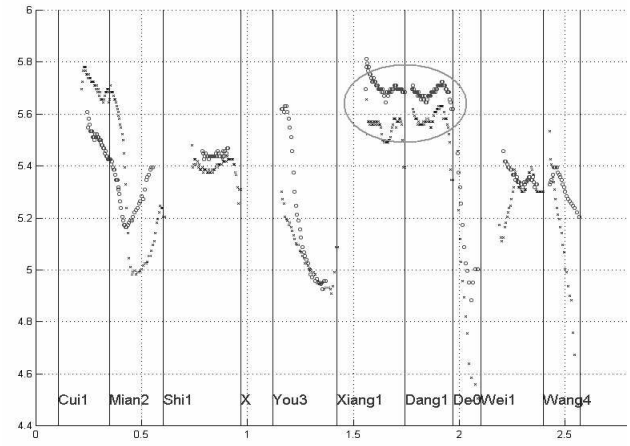

Figure 2. Pitch contours for real speech (in x) and synthesized one (in o). F0 of the accented word '***xiang1dang1de0***' becomes higher.

Figure 2 shows two pitch contours of synthesized speech and real speech. The script of the sentence is 催眠师有相当的威望 (the hypnotist is with considerable prestige), in PinYin, ***cui1 mian2 shi1 you3 xiang1 dang1 de0 wei1 wang4***. In this case, 相当的 (***xiang1 dang1 de0***) was enforced into A2, while AIs of the rest words were set into A1. The realized F0 of two synthesized tone-1 syllables (***xiang1 dang1***) were higher, and the word was perceived as accented one significantly.

5. Conclusions and Discussions

One acoustic measure, Approximation-Ratio was proposed to represent the AI of Chinese. Prosody parameters predicted by non-AI DT were treated as the 'neutral' ones. And because everything else was equal, the differentials between real speech and 'neutral' speech were reasonably assumed to be caused by accent varying. AI was the discretization of Approximation-Ratio, which was retrieved from the differentials.

Both difference analysis and subjective evaluation showed that the AI annotated by such an automatic detector was reasonable and efficient.

All the weight coefficients used in the detector were adjusted based on expert's intuition, which was of less strictness. Optimization of those weight coefficients is another research topic to be investigated in the future.

Actually, the corpus annotated with AI is for training the two components: prosody event predictor, and prosody parameter predictor. For the later one, the DT method could be a possible solution as we had done in the paper. For the former one, how to train the prosody event predictor, with the information generated from AI detector, is worthy to be further investigated.

6. References

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

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

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

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

[5] Wang, B., *The Research on Perception of Prosody in Mandarin,* Ph.D. dissertation, Institute of Psychology, Chinese Academy of Sciences, Beijing, China, 2002

[6] Strom, V., "Detection of accents, phrase boundaries and sentence modality in German with prosody features", *EUROSPEECH 1995*, Madrid, Spain, 1995