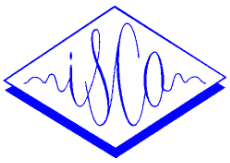

ISCA & IEEE Workshop on Spontaneous
Speech Processing and Recognition
Tokyo Institute of Technology, Tokyo, Japan
April 13-16, 2003

MODELING OF ACCENT PERCEPTION IN CHINESE SPONTANEOUS SPEECH

Jianhua Tao

National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, Beijing, China, 100080
jh_tao@yahoo.com

ABSTRACT

The accent was proved to be the essential links between linguistics and acoustics, and behaves as an important parameter for prosody processing and unit selection in speech synthesis system. In the paper, some acoustical measurements are carried out on F0, duration, silence in order to disclose the relationship between accent and corresponding acoustical parameters. The normalized acoustic parameters are induced to facilitate the accent detecting. Based on this, a model is proposed to predict accent in spontaneous speech. The method is proved to be very successful and has been used to label accent automatically for our large corpus. Further listening tests also show that the labelling results reaches 86% accurate rate, which is nearly the same as the agreement of hand labelling results. Furthermore, different native's evaluations are also compared in the paper.

1. INTRODUCTION

During the last several years, there has been a rapid progress in Chinese speech synthesis. Now, the method of unit selection and concatenation, accompanying with large corpus, is used widely in the systems design. Nevertheless, the accent was still proved to be the essential links between linguistics and acoustics, and behaves as an important parameter for prosody processing and unit selection. However, it is a real hard work for us to handle the accent, such as how to detect the accents in the corpus with high agreement and how to generate a sophisticated method to label the accent in spontaneous speech with high accurate rate.

Normally, accent is not a very well defined term in literature. A common definition of accent is that it refers to those words or syllables that are perceived as standing out from their environment. Perceived syllable accent was interpreted as a gradual parameter by Fant & Kruckenberg. Subjects rated the perceived accent of syllables on a 30-point scale. They investigated a small corpus of Swedish and found linear relationships between perceived prominence and acoustic and articulatory parameters. They also investigated the agreement of their labellers and obtained high correlations; this was confirmed by de Pijper & Sanderman for boundary prominence. Grover et al. showed that the reliability of word prominence ratings is higher for a 10-point scale than for a 4-point scale.

As we know, Chinese is a tonal language and syllable is normally assigned as the basic prosodic element in processing. Each syllable has a tone, and has a relatively steady F0 contour. The tone shapes often deviate from the expected canonical one in spontaneous speech, such as the tonal variations. It is difficult to determine the accents within the influence of various syllabic tone patterns. The paper carried out some acoustical measurements on F0, duration, and silence, in order to disclose the relationship between accent and acoustical parameters. To process the tone in intonation, top line and bottom line of syllables are used to describe the feature of accent in speech. The parameters have already been used to describe the characters of intonation earlier[1][7]. Experiments show that accent is influenced not only by top line and duration of the syllables, but also by the neighboring silence and neighboring accents.

Further more, the paper also describes a model based on linear discriminant function and neural network to detect accent from the spontaneous speech. The model facilitates accent labeling of corpus with very high accurate rate. The labeling results are proved to be very useful and essential for generating the accent predicting model from linguistic part in TTS system.

The paper is organized as following. In section2, the paper describes the preparation of the corpus and criteria of hand labelling which is essential to get the high agreement. In section 3, the acoustic features of accent syllables are analyzed. To facilitate the accent detecting, the normalized acoustic parameters are induced here. Based on above analysis results, section 4 generates two models based on linear discriminant function and neural network. In section 5, some experiments are designed to test the model. Listening tests show that the labelling results reaches 86% accurate rate, which nearly the same as the agreement of hand labelling results. Furthermore, different native evaluations are also compared in this section.

2. CORPUS PREPARING AND HAND LABELLING

The whole speech database used for analysis and training in the paper contains 3167 sentences coming from ten year's PEOPLE'S DAILY, and was read by a professional female speaker with broadcasting style, neutral speed and neutral mood. The average length of each sentence is 20 syllables. Three sophisticated phonetists were asked to label an accent syllable which is perceived as standing out from their environment. The labeling results were classified into three different levels, strong accent, accent and primary accent. Previous study showed that the classification of these levels was not easy[9]. The three kinds of accent are surmised to have slightly different characteristics. In this work, the difference is defined much clearer. The syllable was labeled as strong accent if it is perceived as strong prominence in whole sentence, labeled as accent if it is perceived as standing out from a prosodic phrase, and labeled as

primary accent if it is only can be perceived as standing out among two or three syllables. Table 1 shows the amount and distribution of the different accent syllables in corpus.

Type	Strong Accent	Accent	Primary accent
Amount	*2111*	*3365*	*11876*
Percent	*3.2%*	*5.1%*	*18%*

Table 1, The amount and distribution of the different accents in corpus

There are 17352 accent syllables in total. It takes up 26.3% of the whole corpus. Table 2 shows the agreement of hand labeling among different persons.

Type	Strong Accent	Accent	Primary accent
Agreement	*95%*	*83%*	*72%*

Table 2, The agreement of each accent in labeling

Where, agreement is defined as,

$$AG_n = \partial_n / (\frac{1}{J}\sum_{i=1}^{J} A_{n,i}) \quad (1)$$

∂_n denotes the amount of syllable with same labeling, n means different type of accent, J is the amount of persons who labeled the corpus and i means different person. $A_{n,i}$ denotes the amount of accent n and was labeled by person i.

Here, we found that the agreement of labeling results in strong accent is much higher than others.

3. ACOUSTIC FEATURES ANALYSIS OF ACCENT

3.1. Normalization of acoustic parameters

As mentioned above, tone is the most important prosody feature in Chinese. Normally, the F0 and duration are various to different tones and syllables[2]. Such as the top F0 of tone 1 is relatively higher than other tones, while they are in the same linguistic surroundings. But it doesn't mean the syllables with tone 1 are always stronger than others, even they may have higher pitch. To reduce the influence of tone and inherent syllabic feature, the F0 and duration are normalized by pitch range and average duration according to different tones.

$$F_{t,nor} = (F0_t - F_{t,\min}) / (F_{t,\max} - F_{t,\min}) \quad (2)$$

$$Dur_{t,nor} = (Dur_{t,real} - Dur_{t,mean}) / Dur_{t,mean} \quad (3)$$

Here, $F0_t$ means the F0 value of current syllable with tone t. $F_{t,\min}$ and $F_{t,\max}$ are the minimum and maximum F0s of all syllables with tone t. $Dur_{t,real}$ denotes the real duration of current syllable, and $Dur_{t,mean}$ is the average duration of all syllables with tone t. Table 3 shows their distribution.

	TONE 1	TONE 2	TONE 3	TONE 4	TONE 5
$F_{t,\max}$	*423HZ*	*403HZ*	*398HZ*	*412HZ*	*401HZ*
$F_{t,\min}$	*287HZ*	*167HZ*	*98HZ*	*143HZ*	*76HZ*
$Dur_{t,mean}$	*298ms*	*201ms*	*256ms*	*198ms*	*128ms*

Table 3, The values of $F_{t,\min}$, $F_{t,\max}$ and $Dur_{t,mean}$ in the corpus

The F0 movement of accent syllable in Chinese cannot be described as one line intonation model. F0 range of syllables can be described as top-line and bottom-line correlates to the accent components[7]. The modification of the range is somewhat as a graph drawn on an elastic band would be magnified when stretched (Chao, 1933)[1]. The F0 movement of accent syllable is traditionally realized by shifting up of the F0 with relatively constant F0 contours[1]. Then the top and bottom lines are very useful for accent determining. To reduce the influence of tone patterns, they are also normalized by pitch range with different tones. Those are,

$$F_{t,nor,top} = \max(F_{t,nor}) \quad (4)$$

$$F_{t,nor,bottom} = \min(F_{t,nor}) \quad (5)$$

3.2. F0 movement

Analysis in the paper tries to disclose the how the normalized top and bottom lines behave within different accents in spontaneous speech. Figure 1 and Figure 2 show their relationship. The X-coordinate means the different type of accent, from normal to strong accent. Y-coordinate denotes the average value of normalized top line or bottom line of each accent.

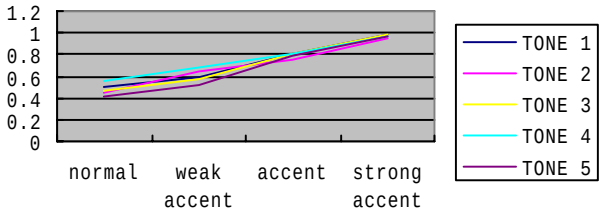

Figure 1, relationship between accent and normalized top line

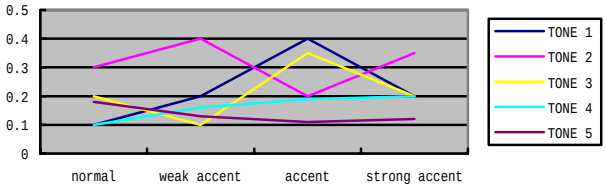

Figure 2, relationship between accent and normalized bottom line

From the figures, the F0 movement of accent syllable is realized by linearly shifting up the top line of the syllables. But there is no clear evidence to support that bottom line will also changed with accents. This confirms the Sheng's view[7]. In his work, he thought that a high top line leads an accent in spontaneous speech and deep bottom line always means the boundary of a rhythm[7].

To get the model of accent prediction, we have to solve another interesting question. Are the top lines of accent syllables always higher than those of normal syllables? Table 4 shows the correlation among them.

	Normal	Primary accent	Accent	Strong accent
Normal	---	*36.8%*	*19%*	*3%*
Primary accent	---	---	*12%*	*9%*
Accent	---	---	---	*11%*
Strong accent	---	---	---	---

Table 4, Correlation among accent types

The values in the table are got by,

$$P_n = | B_{n,n-1} | / | A_n | \quad (6)$$

Where, $| A_n |$ means the amount of syllables with accent level n. $| B_{n,n-1} |$ is the amount of syllables with accent level n whose normalized top line is lower than that of accent n-1.From table 4, it shows there are still some parts of the syllables whose F0s are lower than those with lower accent levels. It means that linear relationship between average top line and accent level cannot be used as criteria to determine the accent in real speech.

3.3. Duration

Duration is another important prosody parameter for accent perception. The relationship between average normalized duration of the syllable and the accent type is shown in figure 3.

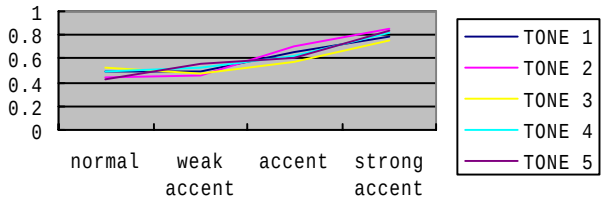

Figure 3, relationship between accent and normalized duration of syllables (X-coordinate means the different type of accent, from normal to strong accent. Y-coordinate denotes the average normalized duration in each accent.)

It shows that duration of accent syllables is enlarged compared to normal syllables. Much similar to the results shown in table 4, this relationship cannot be used as criteria for accent annotation either.

3.4. Combination of F0 and Duration

Since accent is related to both F0 and duration. Further analysis tries to find the joint relationship among them. According to the knowledge in section 3.2, normalized top line of the syllable is the perfect parameter to be the substitution of F0.

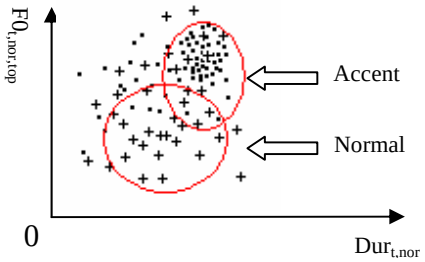

Figure 4, Relationship among accent, normalized top line and normalized duration of syllable (X-coordinate means the normalized duration of syllables, from 0 to 1. Y-coordinate denotes the normalized top line of syllables, from 0 to 1)

In figure 4, the cross dots denote the distribution of normal syllables, the small circle dots represent the distribution of accent syllables. The two big circles line out the main distribution areas of normal and accent syllables. it can be found that the distribution area of accent syllables are up to that of normal syllables, though some parts of them are overlapped, and detailed and clear classification of accent distribution area was very hard to get.

4. MODEL OF ACCENT PREDICTION

4.1. Other related acoustic features

Since the accent is defined as standing out from their environment, the detecting of accent syllables will also be influenced by some other neighboring acoustic parameters, such as silence, adjacent F0 and duration, etc. Normally, silence shrinks after the accent syllables[10]. Some other experiments also show that there are also some rules between accent syllables and adjacent ones. Normally, the syllables are primaryened if they appear before an accent syllable. [10]

To get the similar processing as F0 and duration, silence is also normalized by its range,

$$Sil_{nor} = (Sil - Sil_{\min}) / (Sil_{\max} - Sil_{\min}) \quad (7)$$

Then, the normalized previous and next silence are defined as $Sil_{prev,nor}$ and $Sil_{next,nor}$.

4.2. Basic model

Combination of all acoustic parameters, a criteria of accent determining can be defined by,

$$Y^m = f(F0^m_{nor,top}, Dur^m_{nor}, Sil^m_{prev,nor}, Sil^m_{next,nor}) \quad (8)$$

Where, Y^m means the accent level of syllable m in the sentence. $F0^m_{nor,top}$ is the normalized top line to syllable m, Dur^m_{nor} is the normalized duration of syllable m, $Sil^m_{prev,nor}$ and $Sil^m_{next,nor}$ are the previous and succeeding normalized silence of the current syllable. As mentioned in 3.1, all of the normalized parameters were got in different tone patterns.

With linear discriminant function, (9) can be changed as,

$$Y^m = \alpha \cdot F0^m_{nor,top} + \beta \cdot Dur^m_{nor} + \gamma \cdot Sil^m_{prev,nor} + \eta \cdot Sil^m_{next,nor} + \delta \cdot Y^{m-1} + C \quad (9)$$

$\alpha, \beta, \gamma, \eta, \delta$ are the coefficients (from 0 to 1). C is the constant.

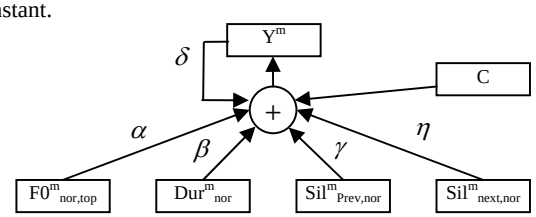

Figure 5, Accent prediction model based on linear discriminant function

According to the corpus, the coefficients are assigned by experience, $\alpha = 0.65, \beta = 0.65, \gamma = 0.3, \eta = 0.5, \delta = 0.5$, however the coefficients may be various to different database. Furthermore, more accurate and efficient coefficients can also be generated by a statistic method.

4.3. Neural Network based model

A neural network is a kind of function that maps the input feature vector to a confidence that the input belongs to a class. In the case of accent determining, the input feature is an acoustic vector $\vec{V}^m = (F0^m_{nor,top}, Dur^m_{nor}, Sil^m_{prev,nor}, Sil^m_{next,nor})$ whose sub-parameters have already been normalized from 0 to 1. The architecture of the neural network is shown in figure 5. The hidden layer consists of 32 nodes with an activation function tanh(x). To integrate the neighboring information, more acoustic features of previous two syllables, current syllables and next syllables are involved into the input parameters.

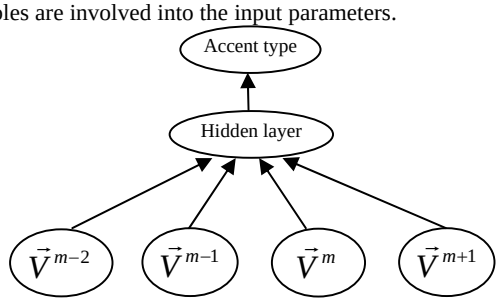

Figure 6, The Architecture of NN based accent prediction model

To get the trained model, 2000 sentences were used for training, 500 sentences used for validation and others used for testing. Both full batch procedure and LineSearch procedure are chosen to train the neural network.

5. EVALUATION AND DISCUSSION

5.1. Evaluation

In order to verify the effectiveness of our method to detect accent, we compute the degree of agreement between natives' perception and our method. As a result of experiments, the degree of agreement between natives' perception and our method is greater than 86%. Compared to the results of the experiments using only one acoustic feature F0 (accurate rate 73% for linear discriminant method, 78% for NN based method), duration (accurate rate 70% for linear discriminant method and 68% for NN based method), the rate is improved by as much as 9%. These results show that the combination of multi acoustic features is a valid model of accent perception for Chinese spontaneous speech. Table 5 shows the results.

Method	Acoustic parameters	Average rate	Strong accent	Accent	Primary accent
Linear Discriminant	Top Line	*73%*	*89%*	*76%*	*58%*
	Duration	*70%*	*82%*	*61%*	*67%*
	All	*82%*	*94%*	*85%*	*71%*
NN based	Top Line	*78%*	*88%*	*81%*	*52%*
	Duration	*68%*	*76%*	*70%*	*65%*
	All	*86%*	*93%*	*88%*	*78%*

Table 5, The result of accent perception

In table 5, it also shows that there are not too much difference between NN based method and linear discriminant method, though NN based model can be adapted into new corpus much easily.

5.2. Evaluation among different natives

The different agreement of accent syllables in Table 2 indicates that each native speaker may perceive accent by using different criteria. In order to see the similarities and differences among different perception, and to realize a universal evaluation criterion, we conducted further experiments. We labeled speech samples by different native evaluators and estimated the parameter weights individually. Each weight trained by different native perception reflects each speaker's tendency of accent perception. From the trained weights, we found general agreement among all native evaluators. The weights of F0 and duration are similar. The values are about 0.45 and 0.25, respectively. The only difference is in the threshold value. This shows that all native evaluators perceive accent by the similar combination of the acoustic features and our model realizes the universal evaluation criterion. The reason for the difference in threshold is that it is difficult to judge whether there is accent or not due to neutral declarative speech for some syllables.

6. CONCLUSION

The paper carried out lots of acoustical experiments on F0, duration to disclose the relationship between accent of Chinese syllables and acoustical parameters in spontaneous speech. Results show that accent is influenced not only by top line and duration of the syllables, but also by the neighboring acoustic information. Based on this, We have proposed a method of automatic detecting accent. We utilize four acoustic features, top line, syllable duration, previous silence and next silence, and combine them by both linear discriminant function and neural network. The models were proved to be much successful in accent annotation of large corpus. It saves us lots of time and money. The labeling results were used for generating the accent predicting model from linguistic part in TTS system.

7. REFERENCES

[1] Wu Zongji, From Traditional Chinese Phonology to Modern Speech Processing, ICSLP2000

[2] W.Bei, Z.Bo, etc, The Pitch Movement of Word Accent in Chinese, ICSLP2000.

[3] Quinlan,J.R. Induction of decision trees. Machine Learning, 1(1):81-106, 1986

[4] Thomas Portele, Perceived prominence and acoustic parameters in american english, ICSLP96

[5] Chilin Shih and Greg P. Kochanski, "Chinese Tone Modeling with Stem-ML", ICSLP2000

[6] Achim F. Muller, Hans Georg Zimmermann and Ralph Neuneier, Robust generation of symbolic prosody by a neural classifier based on autoassociators, ICASSP2001

[7] Sheng Jiong, "Chinese intonation", Research of Chinese, 1992, Vol4, p16-24.

[8] Walter Daelemans, Jakub Zavrel. TiMBL: Tilburg Memory Based Learner, version 4.0, Reference Guide. ILK Technical Report 01-04, 2001.

[9] K. Jenkin, M. Scordilis: Development and Comparison of Three Syllable Stress Classifiers In *Proc. ICSLP*, 1996.

[10] Tao Jianhua etc, "Automatic stress prediction of Chinese speech synthesis", International symposium on Chinese spoken language processing, 2002, 8. Taipei