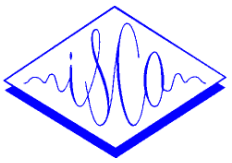


International Symposium on Chinese Spoken Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

AUTOMATIC STRESS PREDICTION OF CHINESE SPEECH SYNTHESIS*

[(1)]*Jian-Hua Tao* [(2)]*Sheng Zhao* [(3)]*Lian-Hong Cai*

Department of Computer Science and Technology, Tsinghua University, Beijing
{(1)jhtao,(3)clh-dcs}@tsinghua.edu.cn (2)sz00@mails.tsinghua.edu.cn

ABSTRACT

The stress 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 stress and acoustical parameters. The normalized compared acoustic parameters are induced to facilitate the stress detecting from the speech. Furthermore, a rule-learning approach is proposed to predict stress in unrestricted Chinese text. In order to improve the accuracy rate of prediction rules, the most effective linguistic features related to stress are selected according to several experiments. The method is proved to be very successful and has been integrated into our speech synthesis system. We get 86% accurate rate of stress prediction. Further listening tests also show that the expressive force of synthesized speech is improved a lot compared to the systems based on traditional method.

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 stress 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 stress, such as how to detect the stresses in the corpus with high consistency and how to predict the stresses from the linguistic information.

The first question exists in the phase of corpus design and labeling. Normally, stress is not a very well defined term in literature. A common definition of prominence is that it refers to those words or syllables that are perceived as standing out from their environment. Perceived syllable stress was interpreted as a gradual parameter by Fant & Kruckenberg. Subjects rated the perceived stress of syllables on a 30-point scale. The authors investigated a small corpus of Swedish and found linear relationships between perceived prominence and acoustic and articulatory parameters. They also investigated the consistency 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. It is difficult to determine the stresses within the influence of various syllabic tone patterns. In the paper, some acoustical measurements are carried out on F0, duration, and silence, in order to disclose the relationship between stress and acoustical parameters, and to make stress labeling in high precise and high consistency. Results show that stress is influenced not only by pitch range and duration of the syllables, but also by the neighboring silence and neighboring stresses.

To get the relationship between linguistic processing and acoustic processing, some data-driven methods have been introduced in English, such as Classification and Regression Tree (CART), Hidden Markov Model (HMM), neural network auto associators. Whereas, rule based stress prediction is still the popular method for most of the current Chinese speech synthesis systems. As a result, the naturalness and flexibility of the system are limited in a certain extent. In the paper, rule learning approach is proposed to predict stress in unrestricted Chinese text. In order to improve the accuracy rate of prediction rules, the most effective linguistic features related to stress are selected according to several experiments. The method is proved to be very successful and has been integrated into our speech synthesis system. We get 86% accurate rate of stress prediction. Compared to other methods, it can be though as a very high performance. Further listening tests also show that the expressive force of synthesized speech is improved a lot compared to the systems based on traditional rule based method.

The paper is organized as following. In Section2, the acoustic features of stresses are analyzed to make automatic stress tagging for the corpus. Section 3 analyzes the mapping the patterns between syntactic information and stresses. In Section 4, a rule-learning algorithm is described in detail, which is used to predict the stress automatically from the linguistic information. The results are analyzed in Section 4. Section 5 presents the conclusion and the view of future work.

2. STRESS DETECTING

As we know, tone is the most important and basic prosody feature in Chinese. There are four lexical tones exist, and each tone contains relatively constant pitch patterns. The pitch movement of stressed syllables in Chinese is complicated that it cannot be described as one line intonation model. Pitch range of syllables can be described as top-line and bottom-line correlates to the stressed components.

In the paper, we try to describe a method on how to detect stress from the speech in the large Chinese speech database. Firstly, to be able to compare with normal behaviors, the acoustic parameters, duration, top F0 and bottom F0 of stressed syllables, are divided by their statistic average values (normal behaviors)..

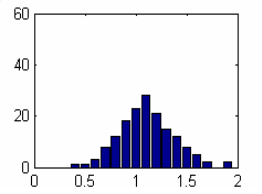
$$R_{dur,i} = \frac{Duration\ of\ \ Syllable\ i}{Average\ Duration\ of\ \ all\ \ Syllables} \quad (1)$$

$$P_{top,i} = \frac{Top\ \ F0\ \ of\ \ Syllable\ \ i}{Mean\ \ value\ \ of\ \ Top\ \ F0\ \ of\ \ all\ \ Syllables} \quad (2)$$

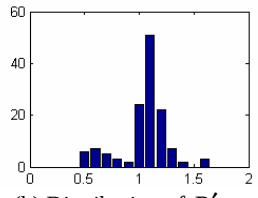
* Supported by 863 program (2001AA114072)

$$P_{bottom,i} = \frac{Bottom\ \ F0\ \ of\ \ Syllable\ \ i}{Mean\ \ value\ \ of\ \ Bottom\ \ F0\ \ of\ \ all\ \ Syllables} \quad (3)$$

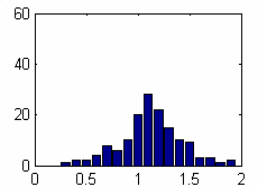
Here, i means syllable i. $R_{dur,i}$, $P_{top,i}$ and $P_{bottom,i}$ represent the compared acoustic parameters, their distributions are shown in Figure 1(a),1(c),1(e).



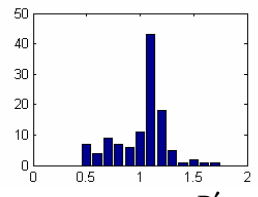
(a) Distribution of $R_{dur,i}$



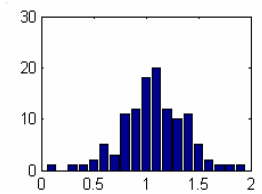
(b) Distribution of $R'_{dur,i}$



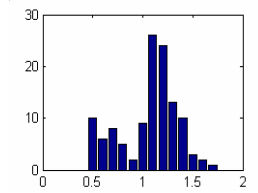
(c) Distribution of $P_{top,i}$



(d) Distribution of $P'_{top,i}$



(e) Distribution of $P_{bottom,i}$



(f) Distribution of $P'_{bottom,i}$

Figure 1, statistic distribution of compared acoustic parameters. x-axis denotes $R_{dur,i}$, $P_{top,i}$, $P_{bottom,i}$, $R'_{dur,i}$, $P'_{top,i}$ and $P'_{bottom,i}$ respectively, y-axis denotes the number of stressed syllables whose $R_{dur,i}$, $P_{top,i}$, $P_{bottom,i}$, $R'_{dur,i}$, $P'_{top,i}$ or $P'_{bottom,i}$ are equal to the value listed in x-axis.

It can be seen that the center of $R_{dur,i}$, $P_{top,i}$ and $P_{bottom,i}$ is more than 1 a little. It means the top F0, bottom F0 and duration are probably enlarged in stressed syllables. But, it can not be determined clearly, since there exists Gauss distribution with a large scale for each statistical result. There is still a large amount of stressed syllables, whose centers of $R_{dur,i}$, $P_{top,i}$ or $P_{bottom,i}$ are smaller than 1.

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 normally, 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 duration and F0 are normalized by the following steps,

$$S_i = \frac{Duration\ \ of\ \ Syllable\ \ i}{Average\ \ Duration\ \ of\ \ Syllable\ \ i\ \ with\ \ Tone} \quad (4)$$

$$H_i = \frac{Top\ \ F0\ \ of\ \ Syllable\ \ i}{Mean\ \ value\ \ of\ \ Top\ \ F0\ \ of\ \ Syllable\ \ i\ \ with\ \ Tone} \quad (5)$$

$$B_i = \frac{Top\ \ F0\ \ of\ \ Syllable\ \ i}{Mean\ \ value\ \ of\ \ Top\ \ F0\ \ of\ \ Syllable\ \ i\ \ with\ \ Tone} \quad (6)$$

Normalized compared acoustic parameters are defined as,

$$R'_{dur,i} = S_i \Big/ \frac{1}{N}\sum_{i=1}^{N} S_i \quad (7)$$

$$P'_{top,i} = H_i \Big/ \frac{1}{N}\sum_{i=1}^{N} H_i \quad (8)$$

$$P'_{bottom,i} = B_i \Big/ \frac{1}{N}\sum_{i=1}^{N} B_i \quad (9)$$

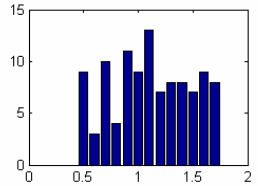
Then, we can get the new distribution shown in figure 1(b),1(d),1(f). It can be seen that $R'_{dur,i}$, $P'_{top,i}$ and $P'_{bottom,i}$ of most stressed syllables are more than 1. The distribution of them shows that normalized compared acoustic parameters represent the stressed syllables much better. To get the more relationship between stressed syllable with other information, silence and adjacent weakened syllables are also taken into account. The compared duration of silence, and statistical number of weakened syllables closed to stressed syllables are got by the following formulas. The results are shown in figure 2.

$$S_{Prev} = \frac{Length\ \ of\ \ Silence\ \ before\ \ Stressed\ \ Syllables}{Average\ \ Length\ \ of\ \ Silence} \quad (10)$$

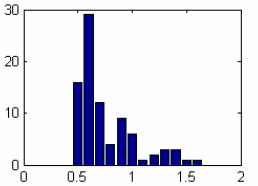
$$S_{Next} = \frac{Length\ \ of\ \ Silence\ \ after\ \ Stressed\ \ Syllables}{Average\ \ Length\ \ of\ \ Silence} \quad (11)$$

$$N_{Prev} = \frac{Number\ \ of\ \ Adjacent\ \ Previous\ \ Weakened\ \ Syllables}{Total\ \ Number\ \ of\ \ Stressed\ \ Syllables} \quad (12)$$

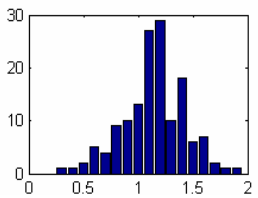
$$N_{Next} = \frac{Number\ \ of\ \ Adjacent\ \ Succeeding\ \ Weakened\ \ Syllables}{Total\ \ Number\ \ of\ \ Stressed\ \ Syllables} \quad (13)$$



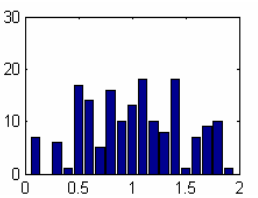
(a) Distribution of S_{Prev}



(b) Distribution of S_{Next}



(c) Distribution of N_{prev}



(d) Distribution of N_{next}

Figure 2, statistic distribution of S_{Prev}, S_{Next}, N_{prev} and N_{next}. x-axis denotes S_{Prev}, S_{Next}, N_{prev} and N_{next} respectively, y-axis denotes the number of corresponding syllables.

From figure 1 and figure 2, some results can be got,

A. The pitch movement of syllable stress is realized by shifting up of the pitch with relatively constant pitch contours and enlarging the duration, which confirm the Wu's view [2].

B. Stressed syllables are also influenced by the neighboring silence. Normally, succeeding silence of stressed syllable is shortened. But there is no enough facts support that stressed syllables are influenced by the preceding silence.

C. Experiments also show that the stress may also be influenced by the previous syllables. Normally, the syllables are weakened if they appear before a stressed syllable. There is no enough phenomenon support that the stress syllables are also influenced by the following syllables.

Based on above knowledge, a criteria of stress determining can be defined by,

$$A_n = a \cdot R'_{dur,n} + b \cdot P'_{top,n} + g \cdot P'_{bottom,n} + h \cdot S_{Next} + d \cdot A_{n-1} + C \quad (14)$$

Where, A_n means stress degree of the syllable n in the sentence. a, b, g, h, d are the coefficients (between 0 and 1). C is the constant. In our database, we use $a = 0.7, b = 0.65, g = 0.45, h = 0.3, d = 0.5$, however the coefficients may be various to different database. Furthermore, more accurate and efficient coefficients can also be generated by a training method.

3. SYNTAX TO STRESS MAPPING

A linguistic theory of syntax-stress mapping must consider the underlying syntactic structure in terms of its hierarchical organization, especially if the syntax of a given language allows different directions of branching as it is the case in an Object-Verb-language (OV) such as Chinese. In Chinese, the *syntactic* OV-parameter means that in structures with verb-final word order, i.e., in most subordinate clauses, the verb takes its argument from its left.

For reasons of explanatory adequacy, we make use of theories which consider the information structure. In Jacobs (1993), for both the so called 'normally intonated' sentences, e.g., widely focused sentences, and sentences containing narrowly focused constituents, stress positions are predictable by terms of *integration*. Stress positions in terms of their relative prominence (e.g., the weight of accents distributed over a syntactic structure) is thus calculable if the position is fixed.

3.1. Influenced by syntactic structure

In a first step of syntax-stress mapping, the syntactic structure and the position have to be determined. There are 191 patterns concerned for Chinese totally in the paper. Let us assume a basic syntactic structure reflecting the superficially linear VO-order (A1):

(A1) **我们认为他错了**(*We think he is wrong.*)

Sentences (A1) are structurally locally different as to whether NP2 '*He*' is the object of verb_1 (as in A1). If A1 is focused widely, i.e., and if syllable or word is located at a very high branching node in the syntactic structure, the 'normal' default accentuation has to be applied: in A1 the second verb 'think' is accented.

Other syntactic information such as word boundaries, phrase boundaries, syllable positions (in word or phrase), also affect the performance of the stresses.

3.2. Influenced by keywords

The stressed syllable can also be influenced by the keywords appeared in the sentences. Such as, **今天的报告主题是"战争与和平"** (*The title of today's talk is " War and Peace"*). In this sentence, the word "*war and peace*" was extremely emphasized to be the key topic of the sentence. Nevertheless, some typical verbs may also deduced emphatic components, such as, **是**(*is*), **说**(*say*), **做**(*do*), etc. Same phenomenon also appears in some adverbs, **很**(*very*), **非常**(*much*), **太**(*too*), **肯定**(*surely*), etc.

3.3. Influenced by tonal sequences

According to recent tonal sequence models (Reyelt, Grice, Benzmü ller, Mayer, and Batliner 1996 for German), the main stress positions derived from the syntactic and information structure as described above serve as anchor points for the association of tonal sequences. In Chinese, stresses can be assumed to be realized preferrably by tonal/pitch variations.

Despite the syntax to stress mapping, stress can also be influenced by the different kind emotions. Within different speaking mood, speaking speed and emphatic component are variously changed. However, it's very hard to us to handle them, being lack of semantic and concept parsing ability.

4. RULE LEARNING METHOD

4.1. Rule Templates

Rule Templates is the basic and most important features for rule learning based stress prediction. In our work, linguistic features are classified into different levels, syllable, prosody word, prosody phrase, and sentence level. As a common problem, it's still hard work to perform the syntactic parsing in real time. Features used for stress prediction should be retrieved much reliably and efficiently. Here we separate the parameters related to stress prediction into two types, basic features (*BFEATS*) and advanced features (*AFEATS*).

Part-of-speech (POS) sequences are the most popular features used in the previous research. And it's much easier to automatically get POS tags from unrestricted Chinese text than other deep syntactic structures such as syntactic phrase or components. Two feature sets based on POS features were induced. One is a base feature set (*BFEATS*), using a stress label history of previous five words and a POS window of five-word width, three to the left and two to the right. POS features are derived from three POS sets simultaneously. The first one is the POS set of the tagger having 30 POS tags. The second one is much larger, in which the top 100 frequent words themselves are treated as independent POS tags in addition to those in the first set. The last one has only two tags: content words or functional words. The content words are those belonging to POS tags that are open word set. The functional words are on the contrary. The adoption of multiple POS sets results in POS features of different granularity.

The other feature set (*AFEATS*) is based on *BFEATS*, which includes some additional features: (1) the length of each word in the POS window, in syllable; (2) the length of the sentence, in words and syllables; (3) the position from the start and end of the sentence, words; (4) the distances from the current syllable to the last preceding stressed syllable. These features are all numeric and related to length or distance.

POS_0 POS_1 => STRESS_1
POS_-1 POS_0 POS_1 => STRESS_1
STRESS _0 POS_-1 POS_0 POS_1 => STRESS_1
POS_0 POS_1 WLEN_0 WLEN_1=> STRESS_1

Table 1, Examples of rule templates

4.2. Rule learning method

A classifier is a function that maps the input feature vector $\overrightarrow{F} = (x_1, x_2, ..., x_n)$ to a confidence that the input belongs to a class. Research on machine learning has concentrated in inducing rules from unordered set of examples, especially feature–based induction, an inductive formalism where examples are described in terms of a feature (attribute) vector. And knowledge represented in a collection of rules is understandable and effective way to realize some kind of intelligence. The rules induced by machine sometimes can be combined with handcrafted ones to improve overall performance.

C4.5 rule learning is a decision-tree based induction method (Quinlan, 1986), which can handle both discrete and continuous features and is robust to noise and sparse training. Transformation-based learning (henceforth TBL) is introduced to fulfil part-of-speech tagging in (Brill, 1995). The central idea behind TBL is to start with some simple solution to a problem, and apply transformation rules to correct errors. TRBL is usually applied to classification tasks in NLP such as tagging and base-NP chunking. A general TRBL toolkit (Grace and Radu, 2001) is used to accelerate our work.

The two algorithms are both supervised learning and can be used to induce rules from examples. But they also have difference from each other. The constrained TRBL experiment uses rule templates that were equivalent to the types of questions that could be asked in the decision tree, with the exception that the decision tree includes questions about groups of categorical features and the TRBL templates considers only one or two values. For this constrained case, where TRBL is at some what of a disadvantage, but TRBL gives slightly better performance than the decision tree. Surprisingly, the difference is not significantly affected by reducing the training data (up to 2/3 reduction). The best case TRBL includes a two-feature combination (and") rule template and performance improves. In both cases of TRBL, several different initialization rules were tried, with only insignificant differences in performance. Results reported here are based on initializing all syllables with a stress, in which case the first rules learned effectively lead to the simple content-word system.

The first experiment was the prediction of pitch accent locations with phrase boundaries known. Features used for prediction has been described in 4.1. Results for decision tree prediction and TRBL are presented in table 3.

5. RESULTS AND ANALYSIS

5.1. The corpus preparing

Speech database used in this evaluation is the continuous male speech database of phoneme balanced 3000 Chinese sentences chosen from the four years' news paper of PEOPLE DAILY. There are 19806 Chinese characters in the corpus, which constitute 13375 words. All of the sentences were labeled with stress in three-level with automatic labeling method based on formula (14). And, two experienced annotators, guided by a senior phonetist, checked all of the results by listening. The labeling results of them achieve a high consistency rate of 93.1%.

5.2. Evaluation Parameters

Stress prediction was evaluated with subjective or objective measure. The subjective measure is generally performed by perceptive tests, which are undoubtedly convincing but time-consuming to conduct on large corpus. In this paper, only the objective measure is adopted. As a classification task, stress prediction should be evaluated with consideration on all the stresses. The trained classifiers are applied on a test corpus to predict the label of each stress.

Method	***Precision***
Rule based	0.72
CART	0.83
TRBL	0.86

Table3: Prediction precision

From table 3, we can find the rule based method is superior to the pure rule based method. In contrast, decision trees are aimed at classifying independent vectors, though questions about local context can be incorporated by making Markov assumptions and using dynamic programming and the most likely sequence. For this reason, TRBL tends to be less sensitive to data scarcity, and is better able to learn parameters associated with independent factors.

6. CONCLUSIONS

The paper compared the acoustic parameters of stressed syllables with the corresponding normal status, and draw some efficient views for automatic stress labeling of large speech database. Facilitated by the manual checking, a high labeling consistency can be acquired. The paper also introduced a new approach to symbolic prosodic label prediction based on transformational rule-based learning. Experiments on stress prediction with a news corpus show that TRBL gives a small improvement over simple decision tree predictors, despite a more simplified approach to set membership rule design. In addition, the experiments showed that stress prediction benefits from phrase structure, but not vice versa. The use of average absolute distance is proposed as a new metric for design and evaluation of stress prediction, which is motivated by the graded acoustic cues observed for different stresses. On the other hand our results are based on cross validation tests, which give a better estimation of the performance when the classifiers are running on noisy inputs.

Thanks Prof. Zhou Tongchun for his effort of labeling most the training corpus.

7. REFERENCES

[1] Paul Taylor, Black and Alan.W.Black. Assigning phrase breaks from part-of-speech sequences, Computer Speech and Language v12, 1998.

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

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

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

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