HEARER MODEL BASED STRESS PREDICTION FOR CHINESE TTS SYSTEM

Guo-Ping HU, Qing-Feng LIU, Yu HU, Ren-Hua WANG
iFly Speech Lab, University of Science and Technology of China, Hefei, China
Email: applecore@ustc.edu

ABSTRACT

People often feel tired if he/she listens synthesized speech for a long time. This is mainly because synthesized speech is too flat and never stresses the focus. Different to traditional TTS research approach of simulating speaker, this paper does the stress prediction research from the point of the hearer. An ideal hearer model is first proposed to predict the stress distribution based on the hypothesis: people speak with limited stress effort and distribute the limited effort to ensure that the hearer can understand the speaker easily. Then according to the limited research resource, this paper modifies the ideal hearer model and presents a practical model. Experiments show that the stress prediction achieves an acceptable rate of 87.36%.

Keywords: Hearer model, Stress prediction, Speech synthesis

1. INTRODUCTION

Phoneme, rhythm and stress are the most needed three augments from text in the text-to-speech system. With the development of TTS technology, a lot of research work has been done on the phoneme and rhythm processing and promotes the intelligence and naturalness of synthesized speech quite a lot. But the stress in synthesized speech is almost invariant, which is quite unnatural. So stress information should be analyzed from speech, predicted from text and performed in the synthesizer to promote the naturalness.

Here we should first present our stress definition. In this paper, the stress is defined as the strength of syllable that people pronounced. The stress can be apperceived and represented in speech as amplitude, pitch and duration. Such as the stress of noun and verb are often stronger than auxiliary word.

Also lots of research has been done on stress with the development of TTS technology, such as Chilin Shih (2003) [1] and Steffen Werner (2004) [2]. For Chinese, researchers also tried to analyze the stress and predict the stress, such as Tao Jianhua (2002) [3]. Almost all the researchers analyze the stress from recorded natural speech and try to simulate the behavior of the speaker. It is right but not the only one right. Speech is created for communication, and communication has two participators: speaker and hearer. People always prefer less effort in communication [1] [2], so people always just put heavy stress on the key part of the sentence and leave the left part unstressed, which forms a stress distribution. Though the stress (energy) of TTS system is limitless, but for naturalness, TTS system also needs to distribute limited stress into different part of the sentence. It is quite difficult to predict the stress distribution by simulating the speaker, because there are lots of different distributions that are acceptable for same sentence. But standing on the position of the hearer, under the limitation of stress sum, we can choose one of stress distributions that can make the hearer easiest to understand the speaker. The approach is to put the limited stress on the key part of the sentence that can mostly restore the meaning of the sentence. This is the basic idea of hearer model that will be presented in this paper.

The remained parts are structured as following: In section 2, we explain and demonstrate the ideal hearer model. In section 3, we simplify hearer model and present a practical model. And in section 4, the experimental result is presented. Finally, we give our conclusions in section 5.

2. IDEAL HEARER MODEL

We must simulate the function of hearing first before we apply hearer model. We define how well the hearer understands speaker as how many transcripts the hearer can write down. Though for state-of-the-art TTS system, the understandable rate is almost 100%. But this score is achieved by the system synthesizing each syllable with extreme stress (which loose naturalness) and listener listening with 100% energy (which make the hearer easy to feel tired). The hearer model is just to simulate one relaxed hearer who just put 20%~80% energy on hearing. We wish to find good stress distribution and perform the stress into speech, so we can lead the

0-7803-8678-7/04/$20.00 ©2004 IEEE

hearer to put his limited energy on the key part of the sentence. This is why the model is not named as listener model but hearer model. Now we will describe the hearer model with its three components in detail.

Confusion set:

Confusion set is defined as *C(P,s)*. Here *P* is the phoneme of one unit. The unit of hearer model can be syllable, word or phrase. *s* is the stress of the unit, values from 0.0 to 1.0. *C* is a set of units, which is a function of *P* and *s*. *C* is defined as the possible transcript of *P* if *P* is pronounced with stress *s*. *s* = 1.0 means the pronunciation is loud, slow and accurate, which ensure the phoneme can not be misheard and can only be transcribed to the units whose phoneme is *P*; and s = 0.0 means the pronunciation can be misheard as any other phonemes and can be misunderstood as any unit. *C* will become larger when *s* decreases. Assuming the unit as the Chinese word, the *C(li4xing4, 1.0)* should be *{例行,厉行,力行}*, *C(li4xing4, 0.9)* may be {例行,厉行,力行,*理性*}, and *C(li4xing4, 0.0)* should be the whole Chinese word set.

Not only the phoneme is the media of information, but also the rhythm. Such as "yu3yin1he2cheng2le0", the corresponded text could be "语音 合成了" or "语音盒 成了" depend on the position of pause. So we add the pause into the hearer model too. TTS system always distinguishes four level of pause in speech [4]. But considering the prosody word pause shares the most information that contained by the pause, we just define pau0 as no pause and pau1 as prosody word pause and define *C(pau0,1.0) = {pau0}* and *C(pau0, 0.5) = {pau0, pau1}*, etc.

Language model:

Given the phoneme and stress of each unit in one sentence, noted as $C(P_i,s_i)$ is fixed. And now language model is employed to choose one unit from each $C(P_i,s_i)$ to compose one sentence. The choosing should ensure the composed sentence having maximum probability according to the language model.

Predicting algorithm:

For each given sentence, the phoneme of each unit is fixed, noted as P_i, i = 1,2,…,n, where n is the unit count. The target of the hearer model is to find the optimal s_i which obeys following two limitations:

1） s_i =argmin |S'-S|

2） s_i =argmin $\sum s_i$

Where *S* denotes the original sentence, and *S'* is the new composed sentence according to the P_i and s_i, and |*S'*-*S*| means the count of the units which are different in *S'* and *S*.

The algorithm is a NPC problem if we try to find the global optimal s_i. But the |*S'*-*S*| is a monotonic decreasing function of each s_i. So the min |S'-S| must be achieved at position s_i =1, ∀ i. Here we can use hill climbing algorithm, initialize from the position where each s_i =1, and try to find the minimum of $\sum s_{i..}$

3. PRACTICAL HEARER MODEL

Section 2 demonstrates the ideal hearer model. We call it ideal model because there are still some problems in the model that are difficult to solve, such as confusion set construction method, base unit selection and the computing complexity.

Confusion set construction:

The biggest problem is the construction of confusion set, and experiments show that predicted stress result is very sensitive to the confusion set construction. As the definition of confusion set, confusion set should be constructed as a set of unit with similar phoneme. So for Chinese, we should define the similarity of each syllable, which needs lots of research. We try two approaches:

1) The similarity of two syllables is defined as multiplication of the similarity of their initial, final, and tone, and the similarity of the initial, final and tone is partly shown as examples in table 1. Noted as Construction method 1.
2) The similarity of two syllables is defined as the confusion probability in speech. We let persons listen to the segmented speech of each syllable individually and write down the phoneme they heard, and from this experiment the confusion set of syllable is built and partly shown in table 2. Noted as Construction method 2.

Table 1: Confusion set for initial, final (partly)

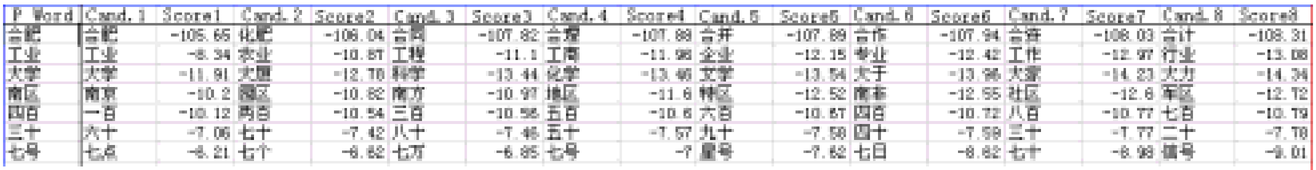
Phoneme\Similarity	1.0	0.9	0.8	0.5
ang	ang	ong,an	eng,ing	en,in,ian
ian	ian	an	en	in
h	h	g, k	d, t, p	n, l, r,

Table 2: Confusion set for syllables (partly)

Phoneme\Similarity	1.0	0.9	0.8	0.5
gui1	gui1	gu1	gong1	gou1
sheng4	sheng4	sheng3, seng4	shuang3, shen4	shen1, cheng4

Base Unit Selection:

Unit selection is the first and basic problem in the hearer model. There are several candidates for Chinese base unit, including the



P Word	Cand.1	Score1	Cand.2	Score2	Cand.3	Score3	Cand.4	Score4	Cand.5	Score5	Cand.6	Score6	Cand.7	Score7	Cand.8	Score8
合肥	合肥	-105.65	化肥	-106.04	合同	-107.82	合理	-107.88	合并	-107.89	合作	-107.94	合资	-108.03	合计	-108.31
工业	工业	-8.34	农业	-10.87	工程	-11.1	工商	-11.98	企业	-12.15	专业	-12.42	工作	-12.97	行业	-13.08
大学	大学	-11.91	大厦	-12.78	科学	-13.44	化学	-13.46	文学	-13.54	大于	-13.96	大家	-14.23	大力	-14.34
南区	南京	-10.2	园区	-10.82	南方	-10.97	地区	-11.6	特区	-12.52	南非	-12.55	社区	-12.6	军区	-12.72
四百	一百	-10.12	两百	-10.54	三百	-10.56	五百	-10.6	六百	-10.67	四百	-10.72	八百	-10.77	七百	-10.79
三十	六十	-7.06	七十	-7.42	八十	-7.46	五十	-7.57	九十	-7.58	四十	-7.58	三十	-7.77	二十	-7.78
七号	七点	-6.21	七个	-6.62	七万	-6.85	七号	-7	星号	-7.62	七日	-8.62	七十	-8.98	信号	-9.01

Fig. 1: The score of each candidate of the prosody word in one sentence

following two:

1) Syllable: Treat each syllable as unit, and try to predict stress for each syllable. Using syllable as base unit needs to add the *pau0* and *pau1* described in section 2 to reserve the information contain in the rhythm.
2) Prosody word: Treat the prosody word as unit. The selection is more natural for speech procession. And because its boundary is just the pause in the speech, so *pau0* and *pau1* can be omitted. For practical hearer model, we choose prosody word as base unit.

We analyze people's misunderstanding in listening and find that mistaken syllable count in a word is always no more than one syllable, so we propose a new confusion set construction method here. We define the similarity of the two prosody words as the percents of same part in them, noted as Construction method 3.

Computing complexity:

It is easy to see that the computing complexity is quite high because we suppose stress is influenced with each other. It is rational to suppose the influence, but it leads to high computing complexity and makes the model to be an ideal model. For a practical stress prediction model, we need to cut down the computing complexity and cut off the influence between each stress. So we propose following algorithm:

1) Fix the context of the unit just to be their original text
2) Try each candidate unit in the confusion set to replace the unit and apply language model to calculate the probability of the string constructed with original context and each candidate.
3) Order the probabilities and predict the stress of the unit by the probabilities first 20 units using linear regression algorithm, the regression parameter is trained in corpus with stress annotation.

Figure 1 shows the probability and order of confusing prosody words for one sentence. From the figure we can induce that "南区", "四百", "三十", "七号" should have heavy stress because they have lots confusion units with higher probability. But "合肥", "工业", "大学" can have lighter stress because they are just the units with highest probability and can easily understand by hearer.

4. EXPERIMENTS

Corpus:

We annotated one corpus of 1165 sentences, 8336 prosody words. All the sentences have corresponding speech recorded from CCTV news report program. One assistant annotated the stress for each prosody word according the speech. We define and distinguish three levels of stress: 0 for light stress, 1 for medial stress and 2 for heavy stress. She annotated twice spanned by a long time, and her annotation consistency is 85.1%.

Language Model Training:

We employ a text processing system [4] to analyze each sentence in 1.3G byte text and generate the prosody word corpus as training corpus. Then we build the language model based on prosody word unit and the language model based on syllable including *pau0* and *pau1* using the language model toolkit from CMU [5].

Ideal hearer model result:

We implement the ideal hearer model first, which uses syllable as base unit and employs the confusion set construction method 1. Table 3 shows one result of this model:

Table 3: Result of ideal hearer model

Text	合	肥	工	业	大	学	南	区	四	百	三	十	七	号
Stress	0.2	1.0	0.7	0.7	0.1	1.0	1.0	1.0	1.0	0.1	1.0	0.9	1.0	0.7
Word stress	0.6		0.7		0.55		1.0		0.55		0.95		0.85	

From the result we can see the predicted stress of "南,区,四,三,七,百" is quite rational, but the result of "合,肥,工,业,大,学" seems randomly distributed within the two syllables in each word, but the average stress of each word is still rational.

Practical hearer model:

According to the practical approach of hearer model, we implement the new model. We compare all kinds of confusion set construction approach by calculating the correlation rate between annotated prosody word stress and the probabilities of the first 20 confusing units. The result is shown in table 4.

Table 4: Comparison between 5 confusion set construction methods

Construct method	1	2	3	4	5
Coefficient	0.396	0.494	0.512	0.319	0.471

Note: method 4 means confusion set is constructed with the unique limitation that two units should have equal syllable count, and method 5 means the set is constructed with the limitation of two units should have identical phoneme transcription.

From the comparison we can draw conclusion that method 3 is the best method for confusion set construction.

Evaluation and scores:

We use both objective evaluation method and subjective evaluation method. The objective method is to measure the accuracy of the predicted stress level and annotated stress level. There are always several different acceptable stress level distributions for one sentence, so the subjective method is employed too. This method is to evaluate the stress result by human and give each sentence a score (0~5). 0 means totally unacceptable, 3 means just acceptable and 5 means that the predicted stress result is perfect.

We use 7390 prosody words as training set, and the left 946 prosody words as test set. Table 5 shows the objective evaluation.

Table 5: Objective evaluation result of practical hearer model

Training set	Annotated \ Predicted	Level 0	Level 1	Level 2
	Level 0	1487	686	56
	Level 1	659	2311	235
	Level 2	445	806	705
	Accuracy	60.9%		
	Serious Error Rate	6.78%		
Testing Set	Annotated \ Predicted	Level 0	Level 1	Level 2
	Level 0	167	137	8
	Level 1	118	245	24
	Level 2	73	104	70
	Accuracy	50.1%		
	Serious Error Rate	8.56%		

Note: Serious Error Rate means the rate of prosody word which has 2 level error between annotated stress level and predicted stress level.

If we always set the predicting result to be level 1, then the accuracy will be 43.4%(train set) and 40.1%(test set). From this we can see that the hearer model has its contribution to the stress prediction. And of cause the accuracy is quite low comparing with other predict problem. We think the low value result from the difficulty of the stress prediction.

We also carry subjective evaluation both in training set and testing set and the result is shown in table 6.

The average score is quite good and unacceptable rate is quite low also should be explained as the essentiality of the stress prediction.

Table 6: Subject evaluation result

Set	Average score	Unacceptable rate
Training Set	3.68	3.23%
Testing Set	3.64	12.64%

Note: unacceptable rate means the percent of the sentences whose score are less than 3.

Following is some examples of predict result:

这里(0) 有(1) **黑色的(2)** 玫瑰(1)

这里(2) 有(1) **红色的(0)** 玫瑰(1)

这里(1) 是(1) **科大(2)** **讯飞(2)** 研究(1) 中心(1)

中国队(2) *正式(1)* 踏上了(1) **第八次(2)** 冲击(1) 世界杯的(1) 征程(2)

Note: The bold prosody words' stresses are the rationally predicted, and the italic prosody word's stress is badly predicted.

5. CONCLUSION

In this paper we deduce one ideal hearer model based on a few hypotheses, and make it more practical by proper modification. Experiments show that the proposed hearer model is rational and achieves remarkable stress prediction result with 87.36% acceptable rate.

Also there are still lots of problems leaving for future research, such as the construction of confusion set and the prediction algorithm. Stress prediction from speech, detection from speech, and performing in the speech are still new fields that wait TTS researchers to explore.

REFERENCES

[1] Greg Kochanski and Chilin Shih, Prosody Modeling with Soft Templates. Speech Communication, V.39, Issue 3-4, pp. 311-352, 2003.

[2] Steffen Werner, etc, Toward Spontaneous Speech Synthesis——Utilizing Language Model Information in TTS, IEEE Transactions on speech And Audio Processing, VOL.12, NO. 4, July 2004

[3] Tao Jianhua, Jiang Dannin, Cai Lianhong, Rule learning based automatic stress prediction of Chinese speech synthesis, ISCSLP2002, Aug 2002, Taipei

[4] Bo Yin, Ren-Hua Wang, A Hierarchic Processing Model In Chinese TTS, ISCSLP2000, Beijing

[5] P.R. Clarkson and R. Rosenfeld, Statistical Language Modeling Using the CMU-Cambridge Toolkit, Proceedings ESCA Eurospeech 1997