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

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

RHYMTH CORRELATION OF SPEECH SYNTHESIS SYSTEM

Jianhua Tao

National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
jhtao@nlpr.ia.ac.cn

ABSTRACT

There has been a rapid progress of speech synthesis, however it is still hard to make good objective evaluation of the speech intonation while training the speech synthesis system. Unlike the traditional method, Standard deviation of intonation, which normally makes the speech synthesis system sounds smooth and flat but with less expressiveness, the paper integrates the rhythm correlation in the evaluation based on the tangential intonation. Furthermore, the paper makes the comparing among three typical evaluation methods, Listening Test, Standard Deviation of Intonation, Standard Deviation of Intonation & Tangential Intonation. It proves the method introduced in the paper could generate better synthesis results than others with even less training corpus.

1. INTRODUCTION

Speech synthesis has been developed steadily over the last decades and it has been incorporated into several new applications. For most applications, the intelligibility and comprehensibility of synthetic speech have reached the acceptable level. However, in prosodic, text preprocessing, and pronunciation fields there is still much work and improvements to be done to achieve more expressive sounding speech.

As long as speech synthesis needs to be developed, the evaluation and assessment play one of the most important roles, especially if you want to get more expressive speech output. Several individual test methods for synthetic speech have been developed during last decades. But the traditional evaluation methods are usually designed to test speech quality in general, however most of them are subjective methods, such as MOS listening test. For trainable speech synthesis system, such as the system based on neural network or HMM, the objective intonation evaluation is much more important than subjective methods. It offers the methods to get better synthetic modules after training. Currently, most of the training systems use the Root Mean Squared Error (RMSE) as the standard deviation of intonation (SDI) between system's output and the training references. For this kind of evaluation, the system could easily be trained and reach the average speaking style, smoothing but a little flat (with less expressiveness).

On the other hand, intonation, stress, and duration are called prosodic or supra-segmental features and may be considered as the melody, rhythm, and emphasis of the speech at the perceptual level. The prosody of continuous speech depends on many separate aspects, such as the meaning of the sentence and the speaker characteristics and emotions, shown in Figure 1. Unfortunately, written text usually contains very little information of these features and some of them change dynamically during speech.

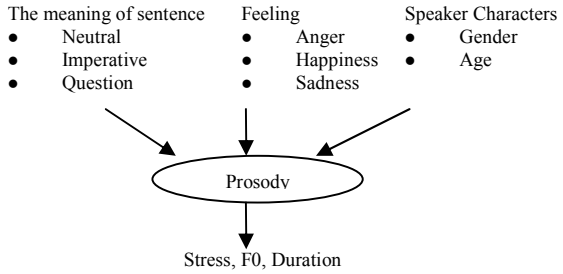

Figure 1, Prosodic dependencies

One of the drawbacks of the standard deviation of intonation is that it computes the error between two F0 values or duration at given time points, where intuitively a method which computes errors in pitch-event alignment both with respect to time and F0 would seem more ideal. Research has shown that listeners are particularly sensitive to stressed syllables and able to predict when they will occur [10] (Cutler & Foss 1977). Rhythm is a regularly recurring sequence of events or movements, which include a beat or stress, and imposes structure on sequences. To generate more expressive speech synthesis results, it is then very useful to integrate rhythm correlation into the intonation evaluation. The method proposed here is a step in that direction, but it will not remove the traditional standard deviation of intonation. It inserts tangential intonation information, which is related to rhythm detection, as more evaluation parameters. The analysis shows that the method could make the trainable system more expressive than other two methods, Listening Test (LT) and SDI.

The paper is organized as following, section II reviews two typical evaluation methods for speech synthesis system, LT and SDI method. In section III, the paper introduces the rhythm correlation in intonation evaluation. In section IV, the paper compares the three methods, and proves the efficiency of the rhythm correlation for trainable speech synthesis system. Final conclusion is described in section V.

2. LISTENING TEST AND STANDARD DEVIATION OF INTONATION

2.1. Listening test and system training

The traditional evaluation procedure is usually done by subjective listening tests with response set of syllables, words, sentences, or with other questions. These methods may be used when the synthesized speech is used through some transmission channel, but they are not suitable for evaluating speech synthesis in general. This is because there is no unique or best reference in a TTS system, not only in the acoustic characteristics, but also in the high-level part which determines the final quality (Pols et al. 1992). On the other hand, the apperceive can not be fully represented by acoustic simulation, subjective listening could still be used for speech synthesis training.

Then, the question is we revise the results while we meet synthesis error in application? There are two methods, making more rules based on experience, or adjusting the weight based on statistic method, while the system meets an unnatural error in testing. Figure 2 shows a model for weight adjusting method of the prosody model which is based on the context information and prosody cost function. The method is briefly described as following [7],

- Find unnatural sentence in the listening test.
- Locate the syllable in which the speech unit or prosody template is bad for synthesis result.
- Change the prosody unit of the syllable to find a group of the suitable units to make better F0 output.
- Calculate the center of all suitable units.
- Change the weights related to context information in prosody cost function.

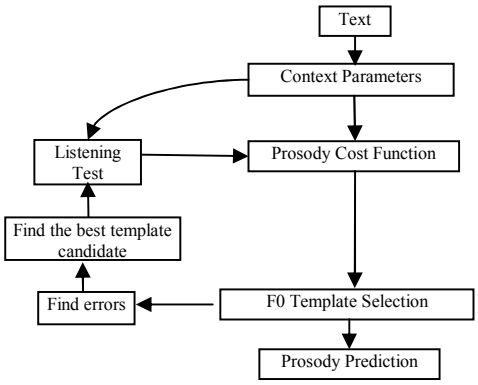

Figure 2, Diagram of Listening Test Based System Training

2.2. Standard deviation of intonation (SDI)

Perhaps the most popular objective evaluation method used for trainable speech synthesis is standard deviation of intonation. That is based on Root Mean Squared Error (RMSE) between the output F0 contours, duration and training references. The training methods will be the procedure which makes the smallest deviation.

For F0 contour, the standard F0 deviation could be got by,

$$d_C^2 = \frac{1}{T}\sum_{t=0}^{T}[f^O(t) - f^R(t)]^2 \quad (1)$$

$f^O(t)$ means the output F0 of the speech synthesis in time t, and $f^R(t)$ is the target F0 in the training set. T is the whole duration of the sentence. The method treats one contour as a reference and compares the other to it. The distance between the contours is measured in a direction normal to the reference contour. This is to incorporate a time component based on the assumption that the contours rates of change are similar, as well as the contours themselves being similar. The warping methods are also necessary for adapting the error computing.

For syllable based duration prediction, the standard duration deviation could be got by,

$$d_D^2 = \frac{1}{N}\sum_{n=1}^{N}[D^O(n) - D^R(n)]^2 \quad (2)$$

$D^O(n)$ means the output duration of the speech synthesis in syllable n, and $D^R(n)$ is the corresponding target duration. N is the whole amount of syllables of the sentence.

An inherent problem with such deviation is that pitch and time are measured on independent with different units which cannot necessarily be combined in a straight forward manner. Speech rhythm organizes speech into regularly occurring temporal units or events (Fox 2000). In this way the predictability of speech events is increased and thus the intelligibility of utterances (Lehiste 1970).

3. TANGENTIAL INTONATION

There has been a longstanding debate in studies of rhythm whether the beat or the regular recurrence is more important (time vs. accent controversy; for a discussion see Adams (1979, 9 ff.)). However, it seems likely that both are equally important. A third characteristic of rhythm is that it creates the "expectation that the regularity of succession will continue" (Abercrombie 1967, 96).

This allows recipients to concentrate attention to these events which highlight the semantically significant parts of the utterance and removes the need of constant attention to any speech input. "Rhythmic structure thus produces useful perceptual redundancy in speech by constraining the time when (important) articulatory events may occur" [9] (Allen & Hawkins 1980: 229). Ramus, Nespor & Mehler (1999) developed an acoustic correlate of rhythm class based on the segmentation of speech into vowels and consonants. It is derived from the behavior of newborns who appear to be able to distinguish between different rhythm types without any knowledge of language-specific phonological properties. The main criticism leveled at this method of measuring speech rhythm is that it does not take account of changes in speaking rate (Grabe & Low to appear).

For Chinese prosody, lots of research results has pointed out both F0 range and duration play an important role in stress determination, and also for rhythm [8]. Of cause, rhythm also contains the prosodic boundary information, such as prosody word, prosody phrase, etc. To simplify the rhythm correlation, we only make the discussion on prosody phrase level. So, the training corpus must be labeled with prosodic phrase information first.

The method described in the paper attempts to utilise the notion that differences between the contours can be brought about by either pitch displacement errors, time displacement errors or a combination of both. The *tangental estimation method* is then used to estimate a distance perpendicular to the direction of change of the contour.

3.1. Deviation of Tangential F0 Range

To calculate these distances, the contours are estimated locally by straight lines drawn from the F0 range, see figure 3. Some one calls these, F0 top lines and F0 bottom lines. The lines are constructed by taking the data from two consecutive frames and using these time and F0 range as end points for the lines. Then a point is chosen on each line and the distance between the points is taken as the distance between the contours of the outputs and the references.

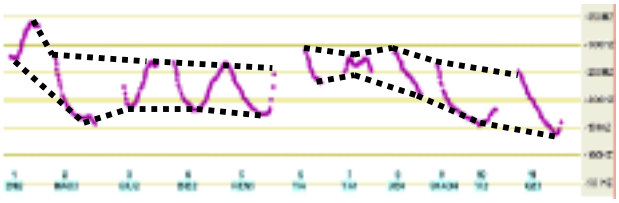

Figure 3, Top line and Bottom line in phrase level

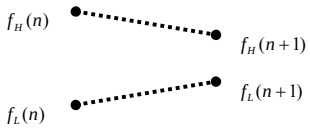

Figure 4, The tangent of top line and bottom line

To get the tangent of the top line and bottom line, we consider a small period of the line from syllable n to syllable n+1. Then, we could get tangential results related to syllable n (see figure 4),

$$\alpha_H(n) = \frac{f_H(n+1) - f_H(n)}{t_H(n+1) - t_H(n)} \quad (3)$$

$$\alpha_L(n) = \frac{f_L(n+1) - f_L(n)}{t_L(n+1) - t_L(n)} \quad (4)$$

Here, $\alpha_H(n)$ is the tangential top line in syllable n, and $\alpha_L(n)$ is the tangential bottom line. $f_H(n)$ means the maximal F0 of syllable n and $f_L(n)$ is the minimal F0 of the same syllable. $t_H(n)$ denotes the time related to maximal F0 of syllable n, $t_L(n)$ is the time related to minimal F0 of syllable n.

The measured distances are then combined in same way that the RMSE calculation is computed.

$$d_\alpha^2 = \frac{1}{N-1}\sum_{n=1}^{N-1}\{[\alpha_H^O(n) - \alpha_H^R(n)]^2 + [\alpha_L^O(n) - \alpha_L^R(n)]^2\} \quad (5)$$

Here, $\alpha_H^O(n), \alpha_L^O(n)$ mean the result of $\alpha_H(n)$ and $\alpha_L(n)$ from the output of speech synthesis system. $\alpha_H^R(n)$ and $\alpha_L^R(n)$ are the target references.

3.2. Duration Deviation and Tangential Duration Deviation

The tangential duration deviation could be estimated in the similar way as tangential F0 range.

$$\beta(n) = D(n+1) - D(n) \quad (6)$$

$$d_\beta^2 = \frac{1}{N-1}\sum_{n=1}^{N-1}[\beta^O(n) - \beta^R(n)]^2 \quad (7)$$

Here, $\beta(n)$ is the tangential duration of syllable n, $\beta^O(n)$ is the result of $\beta(n)$ from the output of speech synthesis system and $\beta^R(n)$ is the target references.

3.3. Standard Deviation of Intonation & Tangential Intonation (SDI&TI)

To get the both intonation continuity and rhythm information, it is very necessary to combine the rhythm evaluation with normal F0 and duration deviation. Then, tangential intonation is integrated with normal one. That is,

$$d_f^2 = a \cdot d_F^2 + b \cdot d_\alpha^2 \quad (8)$$

$$d_d^2 = p \cdot d_D^2 + q \cdot d_\beta^2 \quad (9)$$

Here, a, b, p, q are the weights which could be assigned as experiences or statistic results.

4. ANALYSIS AND DISCUSSION

4.1. Comparing among the methods

To know how efficient the method described in the paper behaves, we made the comparing among three methods, LT, SDI and SDI&TI. Here, we made some tests, among ten students who were studying the linguistic in the university. Two kinds of systems were selected, neural network based system and corpus based unit selection system, both of them are based on syllable level. The training corpus contains 10,000 sentences, about 150,000 syllables.

To compare the results, Pair Comparison method is used, the method is usually used to test system overall acceptance [11] (Kraft et al. 1995). An average listener of a speech synthesizer will listen to artificial speech for hours per day so the small and negligible errors may become very annoying because of their frequent occurrences. Some of this effect may be apparent if few sentences are frequently repeated in the test procedure [11](Kraft et al. 1995). Stimuli from each synthesizer are compared in pairs with all n(n-1) combinations, and if more than one test sentence (m) is used each version of a sentence is compared to all the other version of the same sentence. This leads total number of n(n-1)m comparison pairs. Final MOS is got from the comparing results [11], shown in figure 5.

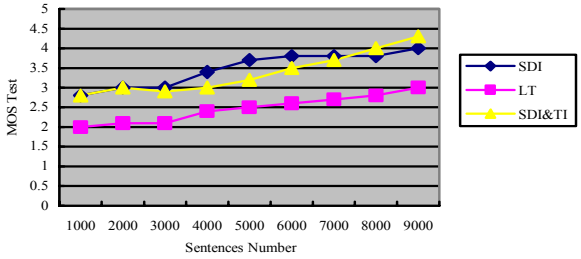

Figure 5, Comparing among three evaluation methods

In figure 5, the x-axis denotes the MOS of the speech synthesis outputs, which is based on 50 testing sentences from newspaper and reader's digest. The y-axis means the amount of the sentences used for training.

To reduce the influence from text analysis part, all of the testing sentences have been labeled with prosodic phrase boundaries. Final results are shown in figure 5. From figure 5, it

shows that the SDI&TI method is comparable with the SDI and LT method. At the beginning, SDI method could reach high score quicker than the other two methods, however the other two methods behaves better with more training set. LT method might be the slowest trained system to reach stable outputs, though the final results might be not bad. SDI&TI method behaves between them, it is trained faster than LT method, and slower than SDI, but it reaches stable output and generate more expressive speech with much less corpus than the others.

This is possibly due to the nature of the synthesized data. The pairs of utterances that are presented to the subjects were designed to be on a smooth continuum with respect to their RMSE scores. However, their perceptual score do not form such a smooth continuum. They were judged either as very similar, or distinctly different which seems to be random variation across listeners. These correlation results show the amount of variance within the variable score which can be accounted for by individual independent variable.

4.2. Influenced by context information ?

The most optimal way to test the suitably for individual application is to perform the test in a real environment. As we have mentioned, the above comparing is based on the prosodic phrase level. But, is there any influence from the context information? Table 1 shows the comparing results among the methods in the different accuracy of the prosody phrases. Here, all of the phrase boundaries are still manually labeled. The accuracy means the agreement rate among the three labelers.

Table 1, Comparing results related to prosody phrases

Accuracy of Prosody Phrase	Comparing Results of the methods
50%	SDI&TI < SD I< LT
60%	SDI&TI < SDI < LT
70%	SDI&T I <LT < SDI
80%	LT < SDI&TI < SDI
90%	LT < SDI < SDI&TI

It is obviously found that the context information has strong influence on the evaluation methods. LT method behaves best with lowest accuracy of phrase boundaries, since LT method could make manual training and weight revising for the system. It's also could find SDI&TI method is the best one while we get the highest accurate prosodic phrase boundaries. The system generates more expressive speech than others.

5. CONCLUSION

Although the development of an acoustic correlate for speech rhythm is still in its beginning stages, it seems to be a promising step towards understanding the rhythmic structure of languages. Measures such as the tangential methods are a useful means of quantifying and thus supplementing auditory impressions of speech rhythm. They facilitate the direct comparison of rhythmic patterns in speech data. However, further research is urgently required. Amongst other things normative data has to be collected and a perceptual basis for the tangential method has to be established and labeled, i.e. determining the effect of a specific change on the listener's perception of speech rhythm.

It is quite clear that there is still long way to go before text-to-speech synthesis, especially high-level expressive synthesis, is fully acceptable. However, the development is going forward steadily and in the long run the technology seems to make progress faster than we can imagine. Thus, when developing a speech synthesis system, we may use almost all resources available, because in few years today's high resources are available in every personal computer. Regardless how fast the development process will be, speech synthesis, whenever used in low-cost calculators or state-of-the-art multimedia solutions, has probably the most promising future.

6. REFERENCES

[1] Jones, C.D., A.B. Smith, and E.F. Roberts, "High-Quality Text-to-Speech Synthesis : an Overview", Journal of Electrical & Electronics Engineering, Australia, vol. 17 n°1, pp. 25-37.

[2] Adams, C. (1979). "English Speech Rhythm and the Foreign Learner". The Hague: Mouton.

[3] K. Dusterhoff and A. Black. Generating F, Contours for speech synthesis using the Tilt intonation theory. ESCA Workshop on Intonation, Athens, Greece, 1997.

[4] D. J. Hermes. Measuring the perceptual similarity of pitch contours. Journal of Speech, Language, and Hearing Research, 41:73–82, February 1998.

[5] K. Ross. Modeling of intonation for speech synthesis. PhD thesis, Boston University, College of Engineering, 1994.

[6] J. 't Hart, R. Collier, and A. Cohen. A perceptual study of intonation: An experimental phonetic approach to speech melody. Cambridge University Press, 1990.

[7] Jianhua Tao, Xin Ni. Auditive learning based chinese f0 prediction. ICASSP2003, Hongkong

[8] Jianhua Tao, Acoustic and Linguistic information Based Chinese Prosodic Boundary Labelling, Tal2004, Beijing

[9] Allen, G. D., & Hawkins, S. (1980). Phonological rhythm: Definition and Development. Child phonology. Volume I: Production (pp. 227-255). New York: Academic Press.

[10] Cutler, A. and Foss, D. (1977). On the role of sentence stress in sentence processing. G, Regional Meeting, Chicago. Chicago Linguistic Society

[11] Kraft V., Portele T. (1995). Quality Evaluation of Five German Speech Synthesis Systems. *Acta Acustica* 3 (1995): 351-365.

[12] Buxton, H. (1983). "Temporal predictability in the perception of English speech". Prosody: Models & Measurements. Berlin et al.: Springer-Verlag.

[13] Classe, A. (1939). "The Rhythm of English Prose". Oxford, Basil Blackwell. Couper-Kuhlen, E. (1993). Cambridge: Cambridge University Press.