PERCEPTUAL DISTORTION ANALYSIS AND QUALITY ESTIMATION OF PROSODY-MODIFIED SPEECH FOR TD-PSOLA

Shi-Han Chen, Shun-Ju Chen, and Chih-Chung Kuo

Advanced Technology Center, ICL, ITRI, Hsinchu, Taiwan

ABSTRACT

TD-PSOLA is one of the most widely used prosodic modification techniques. However, perceptible distortions are introduced occasionally and how TD-PSOLA affects speech quality has not been fully understood and controlled. In this paper, we present a quality estimation method before performing modification. By exploiting relationship between prosodic modifications and subjective scores, 27 distance measures are proposed and respective performances are given and compared. Extensive search is used to find every possible combination among these measures, and the best correlation between the predicted and subjective scores is 87.6%, which can be obtained by linear regression of 4 proposed distance measures. The proposed method does not require synthesizing target and can be used both in online unit selection and off-line corpus design of TTS systems.

1. INTRODUCTION

Prosodic modification has been a subject of great interest in many different research areas. In speech synthesizers, prosodic modification is one of the important factors to create natural and expressive synthetic speech. On the other hand, synthesis units could be reduced if there was no artifact in synthetic speech after prosodic modification. TD-PSOLA [1] is one of the most widely used prosodic modification techniques. However, it introduces perceptible distortions after certain degree of modification.

In order to generate high quality and expressive synthetic speech, an objective quality measure of prosody-modified units is required to control the quality of synthetic speech. In other words, the objective measure can be used to better select synthesis units both in corpus design and online unit selection. However, few works were done in this area and most previous works were related to objective measures of discontinuities at concatenation points [2-5]. Some researchers avoided this issue by performing modification only in some limited range, however it was reported that distortion became perceptible at only 2% pitch modification [6]. In the area of speech coding and transmission, more objective quality measures exist such as PESQ [7]. However, it is not suitable for the problem since spectrums are always changed after prosodic modification, even when quality of synthetic speech is high. In [8] the authors considered speech modified by OGIresLPC, and they used distances between original and target pitch contours to predict synthetic speech quality. However, correlation between subjective and predicted scores was only 62.7%.

In this paper, we first analyze how prosodic modification affects synthetic speech quality. Based on these analysis results, we propose 27 distance measures that can be used to predict synthetic speech quality. Subjective test materials and scoring method are carefully designed to develop prediction models and evaluate the performance of each measure. Extensive search is also performed to evaluate every possible combination among these measures. Considering both prediction ability and robustness of the prediction model, 4 distance measures are combined and correlation between subjective and predicted scores equals to 87.6%. The proposed method uses only original and target prosodic information without synthesizing target speech, therefore it is suitable to be used both in off-line corpus design and real-time unit selection of TTS systems.

The paper is organized as follows. Section 2 first describes the design of subjective test, and in Section 3 we present the perceptual distortion analysis of TD-PSOLA. The proposed distance measures are described in Section 4, and evaluation results are given in Section 5. A discussion of the results and of future work concludes the paper.

2. DESIGN OF SUBJECTIVE TEST

In order to obtain the most reliable subjective scores that will be used in prediction model development, it is very important to carefully design the subjective test.

2.1. Content and quality balanced materials

It was mentioned in [6] that quality degradations introduced by TD-PSOLA have certain interactions between formants and fundamental frequencies. Therefore, in order to cover different speech types as many as possible while keeping the test materials small for efficient subjective test, five groups of different Mandarin vowels, /a/, /i/, /u/, /ε/, /o/, which are located in different vertices of Vowel Triangle, are chosen from corpus of the ITRI TTS system. This corpus contains a female speech whose pitch ranges from 50 to 480 Hz. Therefore although only female speech is

considered, this widely distributed pitch could compensate the unbalanced test materials in some sense. The test materials' structure is illustrated in Figure 1. Each group contains 40 isolated vowels with equally distributed Mandarin tones, and each vowel can generate 39 prosody-modified vowels by using other vowels' prosody as target. In order to create quality-balanced materials, 9 prosody-modified vowels with equally distributed subjective qualities are chosen for each vowel by 2 volunteers, and for each vowel they form a testing-set that contains 1 original vowel and 9 corresponding prosody-modified vowels. There are totally 1800 isolated prosody-modified vowels in the test materials.

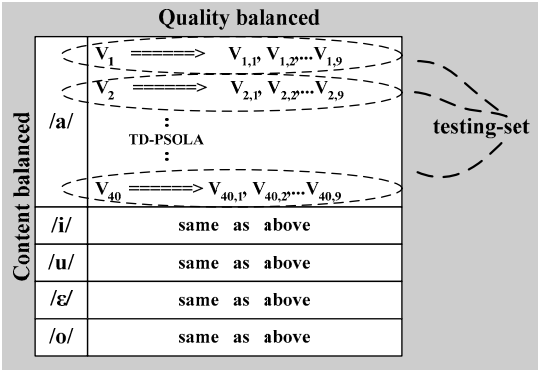

Figure 1 Structure of the test materials

2.2. Reliable scoring method and scores calculation

The scoring method we used in this work is very similar to CCR (Comparison Category Rating), which was defined in ITU P.800 [9]. Listeners were presented with a pair of isolated vowels on each trial and the scale from –3 to 3 is used to judge the quality of the second sample relative to that of the first. Before each test, listeners were given demos about what range of quality and types of distortions will be. However, in order to get the most reliable scores, all combinations in a testing-set, which are C(10,2)=45, are judged by listeners, and this makes it different from CCR. Simple averaging could be used to derive the final score for each testing-set. There are totally 9 male and 7 female college students participated in this test, who have never experienced in such test before.

3. DISTORTION ANALYSIS OF TD-PSOLA

In this paper, we focus on the prosodic modification using TD-PSOLA. Other modification algorithms may lead to different results, even though TD-PSOLA is one of the most widely used prosodic modification methods.

Since TD-PSOLA performs pitch-scale and time-scale modification on speech, it is reasonable to assume that perceptual quality could have some correlation with the amount of modification in pitch and duration. In order to analyze influences of these two factors separately, the original vowels of test group 1 (vowels /a/) are chosen from the test materials. For each original vowel, we generate another 10 pitch-scale and 10 time-scale modified vowels in the range of 0.5-2. Therefore a total number of 400 pitch-scale and 400 time-scale modified vowels are listened. Results are shown in Figure 2 and 3, and we describe several observations here:

1. From the figures it is clear that pitch-scale modification is the main factor of speech quality degradation. Time-scale modification only has some perceptible quality degradation when modification scale is large. It also shows that perceptible distortion still exists in the commonly used range 0.5-2.
2. Decreasing F_0 has greater influence on speech quality than increasing F_0. This result is the same as those described in [6] and [8]. On the other hand, quality seems to be a little bit more easily influenced when lengthening duration of speech.
3. Mandarin tone-4 vowels, which have rapid decrease in pitch contours, are more sensitive to pitch modification. On the other hand, tone-1 vowels that have flat pitch contours are less influenced. It brings to a conclusion that speech with faster changing pitch contour could be more easily distorted by pitch-scale modification. However, this is not clear in decreasing F_0, possibly because decreasing F_0 has such a great influence on speech quality so that even speech with flat pitch contour has no advantage.

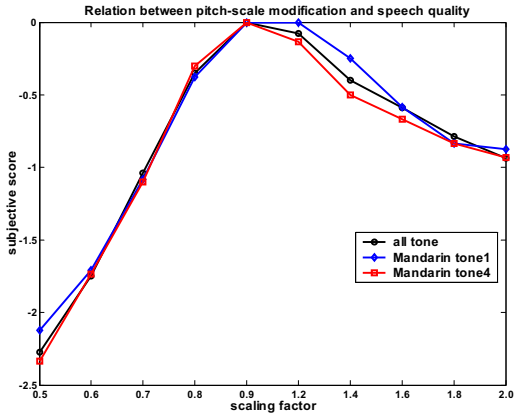

Figure 2 Pitch-scale modification and speech quality

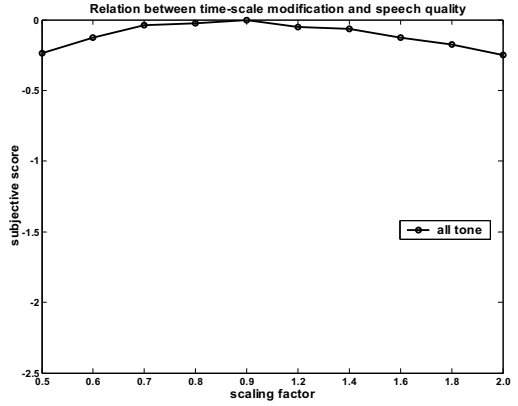

Figure 3 Time-scale modification and speech quality

4. SPEECH QUALITY ESTIMATION

Based on the observations of previous section, we proposed 27 distance measures that are summarized in Table 1. Most of them are pitch-related since pitch modification is the main factor of quality degradation.

DIST. TYPE	DIST. NUM.	DISTANCE DESCRIPTION
Pitch	D_1, D_{13}	{L_p norm, Max} of (pitch distance)
	D_2, D_{14}	{L_p norm, Max} of (pitch-ratio)
	D_3, D_{15}	{L_p norm, Max} of (Δ pitch distance)
	D_4, D_{16}	{L_p norm, Max} of (W_PSCALE pitch dist.)
	D_5, D_{17}	{L_p norm, Max} of (W_PSCALE pitch ratio)
	D_6, D_{18}	{L_p norm, Max} of (W_PSCALE Δ pitch dist.)
	D_7, D_{19}	{L_p norm, Max} of (W_PSLOPE pitch dist.)
	D_8, D_{20}	{L_p norm, Max} of (W_PSLOPE pitch ratio)
	D_9, D_{21}	{L_p norm, Max} of (W_PSLOPE Δ pitch dist.)
	D_{10}, D_{22}	{L_p norm, Max} of (W_EN pitch distance)
	D_{11}, D_{23}	{L_p norm, Max} of (W_EN pitch ratio)
	D_{12}, D_{24}	{L_p norm, Max} of (W_EN Δ pitch distance)
Time	D_{25}	Duration ratio
	D_{26}, D_{27}	{L_p norm, Max} (Pitch-mark discontinuity)

Table 1. Summary of the proposed distance measures

4.1. Pitch-related distance measures

We define *dist*_1 as the L_p norm of differences between original and target log-pitch contours:

$$dist\_1 = \left\{ \frac{1}{N} \sum_{i=1}^{N} [w(i) \cdot abs(F_{0s}(ms_i) - F_{0t}(i))]^p \right\}^{1/p} \quad (1)$$

where N and i are target pitch-mark number and index respectively. Since pitch-mark numbers of original and target may not be the same, they must be time-warped before distance calculation, and ms_i is the time-warped pitch-mark index of original. $w(i)$ is a weighting vector, which can lead to different distance measures in Table 1. L_p norm was used in PESQ [7] to give greater emphasis to localized distortions. Because *dist*_1 tends to give larger distances to pitch contours having higher pitch values, we define *dist*_2 as the pitch-normalized version of *dist*_1:

$$dist\_2 = \left\{ \frac{1}{N} \sum_{i=1}^{N} [w(i) \cdot abs(1 - F_{0t}(i) / F_{0s}(ms_i))]^p \right\}^{1/p} \quad (2)$$

In our previous works, it was shown that if original and target had different tones, i.e., different pitch contour shapes, distortion became larger. Therefore we define *dist*_3 as the L_p norm of differences between original and target Δ log-pitch contour,

$$dist\_3 = \left\{ \frac{1}{N} \sum_{i=1}^{N} [w(i) \cdot abs(\Delta F_{0s}(ms_i) - \Delta F_{0t}(i))]^p \right\}^{1/p} \quad (3)$$

and pitch contours with different shapes will give higher distortions. {D_1, D_2, D_3} can be obtained by setting $w(i) = 1$ in Eq. (1) to (3) for all *i*. D_1 and D_3 are similar with those defined in [8], except L_p norm is used instead of sum-of-squares, and pitch-synchronous distances are used here.

In Section 3 we have shown that decreasing F_0 has greater influence on speech quality. Therefore we can define new weighted distance measures {D_4, D_5, D_6} by setting $w(i)$ in Eq. (1) to (3) as:

$$\text{W\_PSCALE}: w(i) = f(F_{0s}(ms_i) - F_{0t}(i)) \quad (4)$$

where $f(\bullet)$ could be any function, and in this work we use

$$f(x) = \begin{cases} c_1 & , \text{ if } x < 0 \\ c_2 & , \text{ if } x \geq 0 \end{cases} \quad (5)$$

where $c_2 > c_1$. And also, since speech with faster changing pitch contour may be more easily distorted, we can define {D_7, D_8, D_9} by setting $w(i)$ as:

$$\text{W\_PSLOPE}: w(i) = \exp(\alpha \times \Delta F_{0s}(ms_i)) \quad (6)$$

It is assumed that distortions in regions with lower energies such as beginnings and ends of syllables should be harder to perceive. Therefore we can define {D_{10}, D_{11}, D_{12}} by setting $w(i)$ as the normalized energy of *i*-th analysis window:

$$\text{W\_EN}: w(i) = \frac{E(i)^\beta}{\sum E(i)^\beta} \quad (7)$$

Although L_p norm can give greater emphasis to localized distortions, we can still give more emphasis to frames with largest distortions by defining {D_{13} to D_{24}} to be the maximum versions of {D_1 to D_{12}}, which can be derived by setting $p = \infty$ in Eq. (1) to (3).

4.2. Time-related distance measures

We define D_{25} as the duration ratio between original and target:

$$D_{25} = abs(1 - DUR_t / DUR_s) \quad (8)$$

Since in time-scale modification there may be pitch-marks repeats and deletions, we assume that pitch-mark continuity might also correlate with speech quality and we define D_{26} as the discontinuity of time-warped original pitch-marks:

$$D_{26} = \left\{ \frac{1}{N} \sum_{i=1}^{N} [pm\_discont(i)]^p \right\}^{1/p} \quad (9)$$

where

$$pm\_discont(i) = \begin{cases} 0, & \text{if } \Delta ms_i = 1 \\ c_3, & \text{if } \Delta ms_i = 0 \\ \gamma \bullet \Delta ms_i, & \text{otherwise} \end{cases} \quad (10)$$

and D_{27} can be defined as the maximum discontinuity of the time-warped original pitch-marks.

5. EVALUATION OF DISTANCE MEASURES

First, we perform linear regression analysis for each distance measure. 10-fold cross validation is used, and we calculate Pearson's correlation (R) between predicted scores and the 1800 subjective scores. To eliminate subjective voting variations, predicted scores are further mapped to

subjective scores using monotonic 3rd-order polynomial for each listener as described in P.862 [7]. The best p is searched for each distance and results are given in Figure 4, including those obtained using OGI's distances [8]. The best distance measure is D_4 (W_PSCALE + *dist*_1) whose R=0.800. It shows that W_PSCALE can greatly improve performance. W_EN can also help in most distances, while W_PSLOPE only improve D_1. {D_{13} to D_{24}} that use maximum distortion instead L_p norm are also better in some cases. Performance of D_{25} (duration ratio) is the worst, and D_{26}, D_{27} (Pitch-marks discontinuity) are much better than D_{25}. In the figure, some performances are already better than OGI's method, indicates that L_p norm of pitch-synchronous distances may be better than sum-of-squares.

Next, we try to combine these distance measures to further improve prediction performance. Since the possible combinations of 27 measures are so large, greedy search is first used to find sub-optimal combinations of distance measures by maximizing R/(*absolute prediction error*). We plot the performances for every different number of involved distance measures from 1 to 27 in Figure 5. The best performance whose R=0.891 and *error*=0.264, can be obtained by combining 13 distance measures. Combining more distance measures does not lead to better performance. Duration-related distances were not chosen until 10 distances are combined, indicates that pitch-scale modification is the main factor of quality degradation.

In order to make the prediction model as robust as possible, the number of involved distance measures should be kept low [7]. Here we combine only 4 distance measures, whose performance is located near the knee point of Figure 5. The best combination and p for each distance are found by grid search, and the result is {D_2, D_4, D_7, D_{24}} whose R=0.876 and *error*=0.274. It means combining all the proposed weightings can still improve performance, although D_4 only already give an 80% correlation.

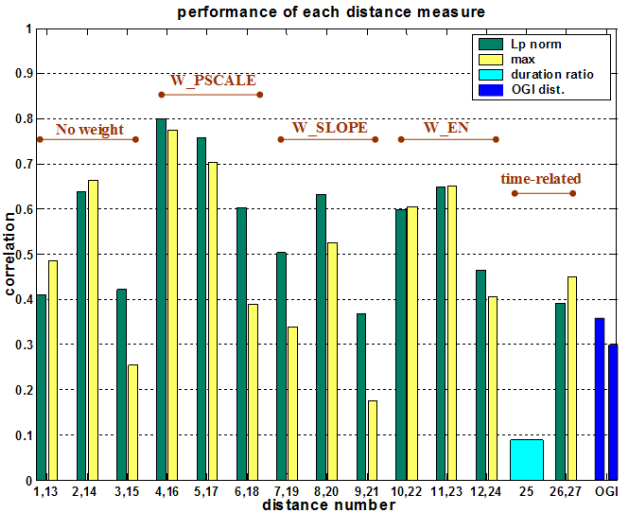

Figure 4 Performances of respective distance measures

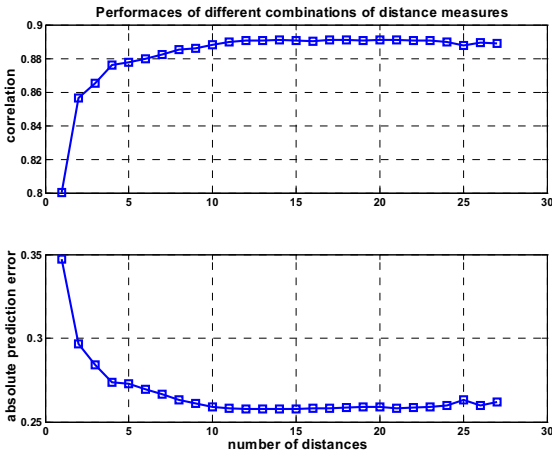

Figure 5 Performances of combined distance measures

6. CONCLUSIONS

We have analyzed how prosodic modification affects synthetic speech quality, and prediction performances of 27 distance measures are given and compared. Good result can be obtained by combining different weighted distances. However, more information can be extracted from synthetic speech to further improve performance, and complexity is not an issue since it can still be used in TTS corpus design.

7. ACKNOWLEDGEMENT

This paper is a partial result of FY95 project conducted by ITRI under sponsorship of the Ministry of Economic Affairs, Taiwan.

8. REFERENCES

[1] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text to speech synthesis using diphones," *Speech Communications*, Vol. 9, pp. 453-476, December 1990.

[2] E. Klabbers and R. Veldhuis, "Reducing audible spectral discontinuities," *IEEE Trans. on S.A.P.*, 9(1): 39-51, 2001.

[3] Y. Stylianou and Ann K. Syrdal, "Perceptual and Objective Detection of Discontinuities in Concatenative Speech Synthesis," *Proc. ICASSP'01*, Salt Lake City, 2001.

[4] Chu and Peng, "An objective measure for estimating MOS of Synthesized speech," *Proc. Eurospeech'01*, Aalborg, Denmark.

[5] J.H.L. Hansen, D. Chappell, "An Auditory-Based Distortion Measure for Segment based Speech Synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 5, pp. 489-495, 1998.

[6] Kortekaas R., Kohlrausch A. "Psychoacoustical Evaluation of the Pitch-Synchronous Overlap-and-Add Speech-Waveform Manipulation Technique Using Single-Formant Stimuli," *J.A.S.A.*, Vol. 101 (4): 2202-2213, 1997

[7] ITU-T Recommendation P.862, "PESQ: An objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs"

[8] E. Klabbers and J.P.H. van Santen, "Control and prediction of the impact of pitch modification on synthetic speech quality," *Proc. Eurospeech'03*, pp. 317-320, Geneva, Switzerland, 2003

[9] ITU-T Recommendation P.800, "Methods for subjective determination of transmission quality"