PROSODIC MANIPULATION SYSTEM OF SPEECH MATERIAL FOR PERCEPTUAL EXPERIMENTS

Nobuaki MINEMATSU† *Seiichi NAKAGAWA*† *Keikichi HIROSE*‡
mine@tutics.tut.ac.jp *nakagawa@tutics.tut.ac.jp* *hirose@gavo.t.u-tokyo.ac.jp*

† Dept. of Information and Computer Sciences, Toyohashi Univ. of Tech.,
1-1 Hibarigaoka, Tempaku-chou, Toyohashi-shi, Aichi-ken, 441 JAPAN

‡ Dept. of Information and Communication Eng., Univ. of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113 JAPAN

ABSTRACT

In perceptual experiments, quantitative manipulation of acoustic features in speech material is often required. And obviously, it can be realized only with speech synthesis techniques. Some of the authors have conducted a series of perceptual experiments, through which they have felt necessity of a system to generate more natural speech. With these backgrounds, a speech stimuli generation system was developed using an analysis re-synthesis technique, where users can freely manipulate prosodic features of input speech and the manipulated material is obtained as synthetic speech.

Degree of resemblance to human speech (henceforth, RHS degree) of the synthesized material was investigated in evaluation experiments. As a result, no perceptual difference was found between synthesized sentences with wrong accents and spoken sentences with the same wrong accents. Furthermore, RHS degree of synthesized sentences with correct accents exceeded that of spoken sentences with flat *F0* contours. These results clearly indicate that this system is useful for the preparation of speech stimuli in perceptual experiments.

1. INTRODUCTION

Some of the authors have been conducting a series of perceptual experiments[1] in order to analyze and model the human process of spoken language perception. Currently they are working on the roles of prosodic features in word and sentence speech perception[2]. These perceptual experiments of course require speech stimuli, correctly and quantitatively controlled in their acoustic features. Consequently, before the experiments, researchers have to bother themselves with gathering an adequate set of speech stimuli. In some cases, they may have to make an original speech synthesizer to prepare the stimuli with required characteristics. In this situation, the researchers without engineering training will have quite difficulties in the preparation of stimuli. This may be a reason why lots of psychologists are dealing only with letters or images as stimuli. After these considerations, we thought that we could assist such researchers by making some of our tools freely available after a tune-up. In this paper, a system for prosodic manipulation of speech material is described, a tentative version of which was already used in [2] for the generation of stimuli.

In case of using synthesized speech as stimuli in perceptual experiments, the quality of the speech should be sufficiently high. However, it is difficult to synthesize speech material of such high quality with current TTS (Text-to-Speech) technology. Consequently, the system was developed on an analysis re-synthesis technique[3]. While this system always requires original speech material to be processed, it has a benefit that it can deal with the material in any language.

Although methods based on concatenation or overlapping of waveforms have been introduced to the analysis re-synthesis technique[4], they require quite accurate pitch extraction to segment input speech pitch by pitch. With the current technology, however, completely error-free extraction is still difficult. Furthermore, if this method is adopted for the experiment which requires a great number of speech stimuli, it should be necessary to check the quality of generated stimuli before the experiment. It must be very hard task for experimenters. These considerations led the authors to design the system based upon the conventional source-filter model of speech production, thus the pitch by pitch segmentation is not necessary. Instead, some schemes discussed in Section **2.** were incorporated into the system to improve the quality of the re-synthesized speech.

In most cases of evaluating analysis re-synthesis systems, the degree of distortion between original and re-synthesized speech has been focused upon. Considering the perceptual experiments where only re-synthesized material is used as stimuli, however, RHS degree of the material should be directly taken into account rather than the distortion from the original speech. In other words, the authors consider that a little distortion is allowable so long as the re-synthesized speech is judged as human speech by subjects. In the previous experiment[5], lots of words were uttered with wrong accents by a male speaker for the preparation of stimuli. If RHS degree of the re-synthesized sentences is comparable with that of human speech, the material for such an experiment as above can be automatically generated and prepared by using the system. Namely, one of the objectives in this study is to investigate to what extent RHS degree can be maintained in the stimuli generated by a system based on the conventional source-filter model[3].

2. SPEECH GENERATION BASED ON ANALYSIS RE-SYNTHESIS

2.1. Approximation of Vocal Tract Characteristics

As mentioned in Section **1.**, since the system was developed based upon the source-filter model, a digital filter must be constructed for the approximation of vocal tract characteristics. While LPC or PARCOR based filtering is often carried out for the approximation, it results in assuming that speech is generated through a mere AR model. As is well-known, however, an AR model does not have any zero in frequency characteristics and it sometimes derives inadequate modeling of speech. While approximation through some kinds of models such as AR is very useful in control theory, its modeling capability should be considered. Namely, in developing the system, non-parametric vocal tract approximation is desirable. This discussion brought us to adopt LMA (Log Magnitude Approximation) filter[6], where any form of frequency characteristics can be realized in logarithmic scale so long as they can be represented with an infinite dimension of cepstrum coefficients. By using MLSA (Mel Log Spectrum Approximation) filter[7], speech spectrum can be modeled in Mel or Bark scale. After preliminary experiments, however, these two filtering showed no difference in the quality of re-synthesized speech. So, we adopted LMA filter which has simpler structure than MLSA filter. As for the cepstrum coefficients, unbiased cepstrum[8] was adopted. The reason for this selection is described in the following section.

2.2. Generation of Source Signals

Source signals to the approximation filter were generated using residual signals, which are obtained by the inverse filtering to original speech. The inverse LMA filter can be realized by sign-reversed cepstrum $\hat{c}_i(t)(= -c_i(t))$, where i and t represent dimension and time respectively.

As will be discussed in the following section, modification of *F0* contour or speaking rate does not require any change in unvoiced segments of speech. Although, as for power modification, unvoiced parts should be changed, it can be realized as an adequate manipulation of $c_0(t)$ only. Therefore, the residual waveform can be directly used as the source signals for unvoiced segments.

For voiced segments of speech, it is necessary to change the shape of the residual waveform according to the prosodic modification. As told in Section **1.**, however, the system should be developed so that the accurate pitch extraction is not required. While the simplest way to generate glottal source signals is to use a pure pulse train, it will degrade the quality of re-synthesized speech. To avoid this degradation without pitch by pitch segmentation, zero-phase conversion is often carried out, after which every component of the signals will have zero in phase characteristics with their frequency characteristics unchanged. As shown in **Figure 1**,

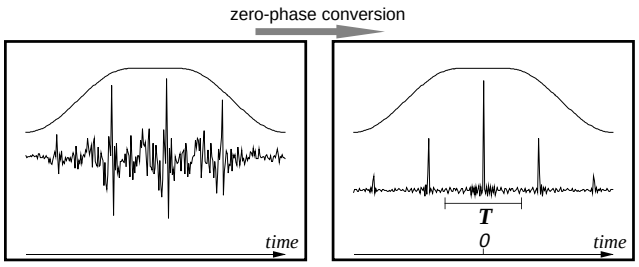

Figure 1: Zero-phase conversion and a pitch waveform used for re-synthesis (indicated by **T**).

by this conversion, the waveform comes to have the largest pulse at time $t = 0$. In this system, this conversion is also applied and a pitch waveform after *F0* modification is defined in this system as **T** in the figure, where the largest pulse exists in the middle. With this procedure, we can obtain a pitch waveform without any waveform edition such as zero-padding to the extent of double the pitch period of input speech. While zero-phase conversion does not change frequency characteristics of the signals, it may degrade the quality of re-synthesized speech. This degradation, however, is thought to be largely suppressed by using the unbiased cepstrum. This is because the unbiased cepstrum coefficients are calculated so that the energy of output signals from the inverse filter by $c_1(t)$ to $c_N(t)$ (N is the dimension of cepstrum and $c_0(t)$ is not used here) is minimized. This implies that use of the unbiased cepstrum makes it possible to minimize the segmental information left in the residual signals.

3. PROSODIC MODIFICATION

3.1. Manipulation of F_0 Contour

For *F0* contour modification, because of the following two reasons, this system was designed so that all the manipulations were conducted through a functional model of *F0* contour generation (henceforth, *F0* model)[9]. One is that it is easier to represent the degree of modification quantitatively by a model-based manipulation of *F0* contours rather than by other non model-based schemes like hand-free edition. The other is that the *F0* model is based upon the human mechanism of speech production. Therefore, *F0* manipulation through the model is expected to prevent users to some extent from generating an impossible *F0* contour, which cannot be produced by humans. In this model, an *F0* contour is well represented only by phrase and accent components and the minimum value of *F0*. The formulation is as follows:

$$\log(F0) = \log(F0_{\min}) + \left[\begin{array}{c}\text{phrase}\\ \text{components}\end{array}\right] + \left[\begin{array}{c}\text{accent}\\ \text{components}\end{array}\right].$$

Figure 2 shows the two types of components and the *F0* contour after their summation. The system can extract *F0* and estimate parameters characterizing the components and $F0_{\min}$ before the prosodic manipulation. Although, for this estimation, it is necessary to extract *F0* from input speech, it is required only to calculate the above parameters. And

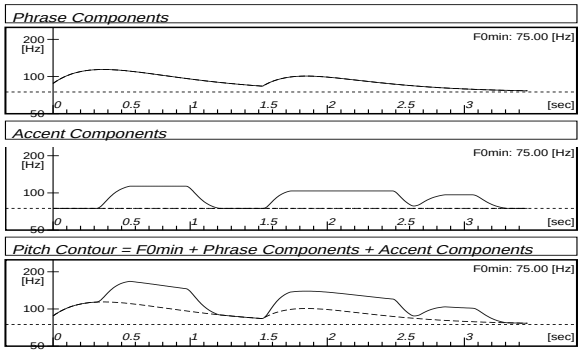

Figure 2: Phrase / accent components and their summation

they will be manipulated to generate stimuli for perceptual experiments. This means that the requirement for the fine pitch extraction to the system is looser than that to pitch synchronous methods. Users' manipulations are of course carried out through a GUI. And based upon the resulting *F0* value and the residual signals after the zero-phase conversion, a pitch waveform is obtained, shown as **T** in **Figure 1**.

3.2. Speaking Rate Modification

As for speaking rate modification, we basically followed the method used in [10][11], where the modification is realized by processing only voiced segments with unvoiced segments left unchanged. In [10], a pitch waveform is obtained through pitch by pitch segmentation and then, by concatenating the duplication of segmented waveforms, lengthened speech is generated. In our study, the segmentation is not conducted. Instead, by repeating the same frame and by producing glottal source signals accordingly, lengthened speech is realized.

Lengthening of the entire voiced segments, however, sometimes produced unexpected additional voiced sounds. In order to cope with this problem, the following two voiced segments were excluded from the segments for lengthening[3]. **(A)** voiced consonant segments detected by Δpower and **(B)** spectral transition segments detected by norm of Δcepstrum. **Figure 3** shows the method of determining the segments for lengthening. While **(A)** is obtained as a segment corresponding to a bottom-to-top jump of Δpower, **(B)** is calculated as a segment which has a larger norm of Δcepstrum than the threshold. It is clearly shown in the figure that the total duration of the segments for lengthening is reduced from that of voiced segments indicated by **(V)**.

Re-synthesis is conducted after the above prosodic modifications. And it is expected that the following post-processing will improve further the quality of the re-synthesized speech. In this procedure, the unbiased cepstrum coefficients are calculated again from the re-synthesized speech, which are referred to as $\tilde{c}_i(t)$ here. Then, LMA filter corresponding to "$c_i(t) - \tilde{c}_i(t)$" is constructed. By performing this LMA filtering to the re-synthesized speech, the obtained signals will take on more exactly the same segmental features as those of the original speech.

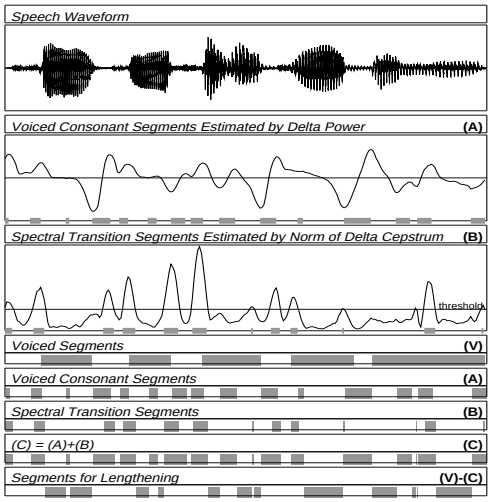

Figure 3: Determination of segments for lengthening

4. EVALUATION EXPERIMENTS

Evaluation experiments were carried out on the re-synthesized speech, which were generated with 10 kHz and 16 bit sampling, 25.6 msec window length and 5.0 msec frame rate. *F0* extraction was conducted every 5 msec and the unbiased cepstrum of 33 dimensions, including $c_0(t)$, were used.

4.1. Speech Material

Dozens of Japanese sentences comprising five familiar words were prepared. After dividing these sentences into the following eight subgroups, groups **A** to **C** were uttered by four male speakers (**SP1**$\sim$**SP4**) in a speaking manner of each subgroup. The rest were synthesized by a rule-based synthesizer. Groups **B** and **C** were re-synthesized after prosodic modification under a certain condition of each subgroup. In the list below, **correct/wrong/flat** indicate the accent conditions used in the utterance or the re-synthesis.

A-1 Human speech(correct)
Sentences uttered with **correct** accents (by **SP1,2,3,4**)

A-2 Human speech(flat)
Sentences uttered with **flat** *F0* contour (by **SP1,2,3**)

A-3 Human speech(wrong)
Sentences uttered with **wrong** accents (by **SP1**)

B-1 Re-synthesized speech(flat→correct)
Sentences re-synthesized with **correct** accents from spoken sentences uttered with **flat** *F0* contour (by **SP1,2,3**)

B-2 Re-synthesized speech(correct→wrong)
Sentences re-synthesized with **wrong** accents from spoken sentences uttered with **correct** accents (by **SP1**)

C Re-synthesized speech(modified speaking rate)
Sentences re-synthesized with **correct** accents after speaking rate modification. Modification rates of their durations were 0.8, 1.0, 1.2, 1.5, and 1.8 (by **SP2**)

D-1 Synthesized speech by a rule-based synthesizer(correct)
Sentences synthesized with **correct** accents

D-2 Synthesized speech by a rule-based synthesizer(wrong)
Sentences synthesized with **wrong** accents

4.2. Procedures

After dividing all the speech material into three groups and randomizing them in each group, they were presented to six subjects through a speaker with an interval of 25 sec. Namely, this experiment consisted of three sessions, each of which had about thirty speech stimuli. As mentioned in Section **1.**, the subjects were asked to evaluate RHS of each stimulus using a scale of seven degrees (0∼6). At the same time, they were also requested to write down what they heard in the stimuli. Before the sessions, two human speech stimuli with correct/wrong accents and two synthetic stimuli with correct/wrong accents were presented to the subjects to let them know in advance what kind of speech would be given to them in the following experiments.

4.3. Results and Discussions

Figure 4 shows the averages and standard deviations of the RHS scores for each subgroup. Both values of **A-1** were, however, calculated except the results of **SP-4** because the prosodically modified material by **SP-4** was not prepared in the experiments. Modification rates of duration in **C** are also indicated in the figure (0.8∼1.8).

Comparison between **A-1** and **C** with the rate 1.0 shows no difference (significant level: $p \approx 15.5\%$). It indicates that the glottal source generation method described in Section **2.2.** is valid at least without prosodic modification. Between **A-1** and **A-2**, and between **A-1** and **A-3**, all of them being human speech, significant differences were found. This implies that it is difficult to evaluate RHS with no correlation to naturalness of speech. Although the score of **B-1** is significantly larger than that of **A-2** ($p < 0.1\%$), it does not reach the score of **A-1**. It is considered to be due to some of the following reasons. The *F0* contour manipulation scheme was probably so immature that the re-synthesized speech might include some unnaturalness. Since power and duration was unchanged through *F0* contour manipulation, they might degrade the quality. Or prosodic modification may require the adequate alternation of segmental features. In any case, this problem should be solved in the future work. However, no difference was found between **A-3** and **B-2** ($p \approx 91.8\%$), which definitely indicates that the system is useful at least to prepare prosodically modified speech stimuli.

5. CONCLUSIONS

In this paper, a prosodic manipulation system of speech material for perceptual experiments was described. Considering the preparation of speech stimuli in the experiments, the analysis re-synthesis method based upon the conventional source-filter model was adopted with LMA filter and some schemes for the source signal generation. In the system, *F0* contour manipulation was realized through *F0* model, which enables the quantitative description of the modification. Evaluation experiments showed that the re-synthesized sentences with wrong accents had a similar RHS score with human speech with wrong accents and that the re-synthesized sentences with correct accents had an even better score than human speech with flat *F0* contours. These results indicate the validity of the system to generate prosodically modified stimuli. The system is open for the academic use and a small demonstration can be seen/heard at
`http://www.gavo.t.u-tokyo.ac.jp/~mine/prosody-e.html`.

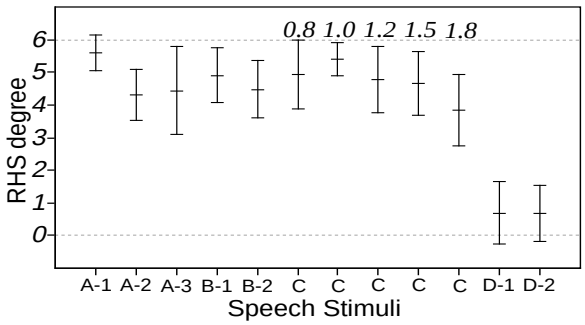

Figure 4: RHS degree for each group

ACKNOWLEDGMENT

This work was supported by Charitable Trusts Ono Research Fund for Acoustics, The Hori Information Science Promotion Foundation, and a Grant-in-Aid for Scientific Research from the Ministry of Education, Science and Culture, Japan.

REFERENCES

1. N.Minematsu *et al.*, "The Influence of Semantic and Syntactic Information on Spoken Sentence Recognition," *Proc. ICSLP'92*, pp.153-156 (1992).
2. N.Minematsu *et al.*, "Role of Prosodic Features in the Human Process of Speech Perception," *Proc. ICSLP'94*, pp.1151-1154 (1994).
3. N.Minematsu *et al.*, "Development of a Speech Stimuli Generating System for Perceptual Experiments," *Technical Report of IEICE*, SP95-134, pp.41-48 (1996, **J**).
4. T.Takagi *et al.*, "A Speech Prosody Conversion System With a High Quality Speech Analysis-synthesis Method," *Proc. of EUROSPEECH'93*, pp.995-998 (1993).
5. K.Hirose *et al.*, "Experimental Study on the Role of Prosodic Features in the Human Processes of Spoken Word Perception," *Proc. ESCA Workshop*, Working Papers 41, pp.200-203 (1993).
6. S.Imai, "Log Magnitude Approximation (LMA) Filter," *Trans. IEICE*, J63-A, 12, pp.886-893 (1980, **J**).
7. S.Imai *et al.*, "Mel Log Spectrum Approximation (MLSA) Filter for Speech Synthesis," *Trans. IEICE*, J66-A, 2, pp.122-129 (1983, **J**).
8. S.Imai *et al.*, "Unbiased Estimation of Log Spectrum," *Trans. IEICE*, J70-A, 3, pp.471-480 (1986, **J**).
9. H.Fujisaki *et al.*, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J.Acoust. Soc. Jpn.*, 4, pp.233-242, (1984, **J**).
10. A.Nakamura *et al.*, "Real Time Voice Speed Converting System with Small Impairments," *J.Acoust. Soc. Jpn.*, 7, pp.509-520, (1994, **J**).
11. N.Shimizu *et al.*, "Improvement of Naturalness on Speech Rate Conversion Algorithm" *Report of Autumn Meet. Acoust. Soc. Jpn.*, 3-P-9, pp.299-300, (1993, **J**).

J = "in Japanese"