Prosody Processing for Speech Synthesis and Speech Recognition

Keikichi Hirose

Department of Information and Communication Engineering
Graduate School of Information Science and Technology, University of Tokyo, Japan
hirose@gavo.t.u-tokyo.ac.jp

Abstract

In view of the importance of prosody in human speech communication, we are intensively working on how to generate prosodic features in speech synthesis and how to utilize them in speech recognition. As for the speech synthesis, we already have developed a corpus-based method for synthesizing prosodic features in the framework of the generation process model of fundamental frequency contours. This fiscal year, an improvement was realized in emotional speech synthesis by including emotion levels in the input parameters of the predictor of the model parameters. As for the speech recognition, a new scheme of detecting fillers during spontaneous speech recognition process was developed. When a filler hypothesis appears during the decoding process of the speech recognizer, a prosodic module checks morphemes, which are hypothesized as fillers, to really be fillers from their prosodic features and outputs the filler likelihood scores. When the score exceeds a threshold, a prosodic score is added to the language score of the hypothesis as a bonus. Experimental results indicated that the proposed scheme could improve the performance of spontaneous speech recognition.

1. Introduction

Communication through speech is the most basic way of human communication, and therefore, a matured speech communication technology is indispensable for the realization of smooth communication between machines (computers) and humans. Speech involves two aspects with rather different features: segmental and prosodic ones. While the segmental aspect is mostly represented by features on vocal tract shape and types of sound source and play an important role in realizing phone sounds, the prosodic aspect is mostly represented by features on vocal folds vibration and play an important role in the transmission of accent, intonation and rhythm. Although both aspects should be well processed in speech technologies, processing of prosodic features has rather been neglected, especially in the speech recognition. Major reasons of this situation will be that the prosodic features cover temporal spans longer than those the segmental features do, and are subject to change widely due to individuals, situations and so on. These factors complicate the prosodic features, and thus make the systematic works on prosody rather difficult. To solve this undesirable situation, research works have been conducted placing a major focus on prosody for the realization of advanced speech technologies. Surely, our research works cover those not related to prosody, such as adaptation of acoustic and language models for robust speech recognition and so on. Recently we have been conducting a series of works on sound separation in less conditions; number of mixed sounds (microphones) is smaller than the number of sound sources. We incorporated several sophisticated methods, such as Hilbert transformation, empirical mode decomposition, and so on, to tackle with the difficult problems [1]. Although rather preliminary, a scheme was developed for localizing moving sources and separating them [2]. This technology is quite important to realizing speech recognition in a realistic situation: several people moving around while they are making conversation.

As for the works related to prosody, the followings are the major results obtained this fiscal year:

1. Corpus-based generation of fundamental frequency (F_0) contours in the framework of the generation process model of F_0 contours (F_0 model) [3]. The method predicts the model commands using binary decision trees with inputs on the text to be synthesized. Because of constraints by the F_0 model, no serious degradation will happen in synthetic speech. Using the method, emotional speech synthesis was realized. Especially, better emotion was realized by taking emotion levels for each *bunsetsu* (basic unit of Japanese syntax consisting of content word(s) followed or not followed by particles) into account.
2. Two-step generation of F_0 contours of Standard Chinese with tone nucleus model [4, 5]. The method is based on the superposition of tone components on phrase components in logarithmic frequency. The tone components are generated by concatenating F_0 patterns of tone nuclei, which are predicted by a corpus-based scheme, while the phrase components are generated by a set of rules. Experiments of F_0 contour generation showed that synthetic speech with high naturalness was possible.
3. Detection of fillers in spontaneous speech recognition [6]. The method is to check the likelihood of a filler candidate in speech recognition process being really filler from prosodic features, and, if yes, adds a prosodic score to the language score of the recognition hypothesis. A comparative recognition experiment with and without the filler checking process was conducted for 100 utterances of spontaneous speech, which are included in the corpus of academic meeting presentations of the Corpus of Spontaneous Japanese. Seven fillers originally miss-recognized as non-fillers are correctly recognized as fillers when the prosodic features are counted, while no fillers originally recognized as fillers are wrongly recognized as non-fillers.
4. Tone recognition of Standard Chinese using tone nucleus model and neural network. The method recognizes tone types of syllables in continuous speech of Standard Chinese using five-layered perceptron. With inputs of prosodic features of current and preceding/following syllables, 86.5 % of correct recognition (including tone 0) was obtained. A slight improvement was further realized by discarding transition parts. The results are better than our method based on representing prosodic features by hidden Markov models [7].
5. Realization of concept-to-speech conversion in a spoken dialogue system on road guidance [8, 9]. The method generates reply speech form the content to be conveyed to the user from the system. To realize this concept-to-speech conversion, a new scheme of sentence generation was developed. It handles the concept in phrase units in a LISP form and concatenates them to generate a sentence. Syntactic structure is kept throughout the sentence generation so that it can be reflected to the prosodic control during speech synthesis. The prosodic control also takes account the phenomenon observable in human conversation: focusing words according to their novelty. The method was realized in

the road guidance system. Its trial use showed that a smooth conversation between the user and the system was possible.

Because of the limitation of space, in the following sections, works on 1 and 3 are introduced.

2. Corpus-based generation of F_0 contours for emotional speech synthesis

Emotional speech synthesis has been conducted as a part of work to realize various styles in synthetic speech, which is necessary to increase the usability of spoken dialogue systems in "real-world applications." We already have developed a corpus-based synthesis of F_0 contours in the framework of F_0 model [9]. By predicting the model commands instead of F_0 values, a good constraint will be automatically applied on the synthesized F_0 contours; still keeping acceptable speech quality even if the prediction is done somewhat incorrectly.

This fiscal year, we newly took the level of emotion into account. By labeling the degree for each *bunsetsu*, and by adding it as inputs to the F_0 model command predictor, a better emotional control was realized in synthetic speech.

2.1. Prosodic corpus

Speech corpus used for the experiment was utterances of a female narrator. She was asked to read the 503 sentences, which are the same with those of the ATR continuous speech corpus, in 3 types of emotion (anger, joy, sadness), and calmly. After recording she was asked to mark the parts of sentences, where she placed emotion specially. The current experiment was done on "anger." In the experiment of F_0 model parameter prediction, 503 sentences were divided into two groups: 453 sentences for training and 50 sentences for testing.

2.2. F_0 model parameter prediction

The parameters of F_0 model are predicted through the following processes:

1. Prediction of phrase command.
2. Prediction of prosodic word boundary location.
3. Decision of accent types.
4. Prediction of accent command.

Processes 1, 2 and 4 are conducted using binary decision trees (BDT's). The CART (Classification And Regression Tree) included in the Edinburgh Speech Tools Library [10] was utilized to construct BDT's. Stop threshold, represented by the minimum number of examples per leaf node, was set to 40. One BDT was constructed for each model parameter for the processes 1 and 4. So predictors for these processes consisted of plural BDT's. In the following subsections, phrase and accent command prediction processes (processes 1 and 4) are addressed, since the emotional level information is assumed to be effective for these processes.

Information on the current *bunsetsu* in question and that on directly preceding *bunsetsu* were included in the input parameters for the phrase command predictor as shown in Table 1. Punctuation marks of the text were not included, because of the large variation according to writing styles. Since the depth of syntactic boundary has a tight relation with the phrase command, boundary depth code (BDC) between the preceding and current *bunsetsu*'s was added to the input parameters. The last three parameters in the table were added to count for the influence of the preceding phrase command on the current phrase command. The category numbers in the parentheses are those for the preceding *bunsetsu* and are larger than those of the corresponding parameters of the current *bunsetsu* by one to represent "no preceding *bunsetsu*." BDC denotes the depth of the boundary between the current and preceding *bunsetsu*'s, and was obtained by a simple calculation from the corresponding KNP code [11]. The input parameters for accent command predictor were selected similarly.

Table 1. Input parameters for the F_0 model parameter prediction.

Input parameter	Category
Position in sentence	28
Number of morae	21 (22)
Accent type (location of accent nucleus)	18 (19)
Number of words	10 (11)
Part-of-speech of the first word	14 (15)
Conjugation form of the first word	19 (20)
Part-of-speech of the last word	14 (15)
Conjugation form of the last word	16 (17)
Boundary depth code (BDC)	20
Phrase command for preceding *bunsetsu*	2
Number of morae between the preceding phrase command and the head of the current *bunsetsu*	25
Magnitude of the preceding phrase command	Continuous

We add emotion levels of the current and preceding *bunsetsu*/prosodic-words into the input parameters of the phrase and accent command predictors: 1 when speaker included emotion specially, and 0 when not. To check the validity of the emotional level for F_0 model parameter prediction, experiments are conducted in the four conditions as shown in Table 2.

Table 2. Use of emotional levels in F_0 model parameter prediction. Symbols "o" and "x" respectively indicate when the levels are used and not used.

Prediction	Con. 0 (Original)	Con. 1	Con. 2	Con. 3
Phrase command	x	o	x	o
Accent command	x	x	o	o

As an objective measure to evaluate the F_0 contour generated using the predicted F_0 model parameters, the mean square error between the generated contour and the target contour is defined as:

$$F_0MSE = \frac{\sum_t (\Delta \ln F_0(t))^2}{T}, \qquad (1)$$

where $\Delta \ln F_0(t)$ is the F_0 distance in logarithmic scale at frame t between the two F_0 contours. The summation is done only for voiced frames and T denotes their total number in the sentence. The results are summarized in Table 3, where average F_0MSE values are shown for 4 conditions listed in Table 2. A better prediction was realized by taking the emotional levels into account. The effect is larger for accent components as compared to phrase components.

Table 3. Average F_0MSE's of F_0 contours generated using the model parameters predicted in four different conditions.

	Original	Con. 1	Con. 2	Con. 3
Close	0.0696	0.0713	0.0692	0.0714
Open	0.0755	0.0750	0.0745	0.0742

2.3. Speech synthesis and evaluation

Two versions of synthetic speech were compared: one with F_0 contours, which were generated by the original method

(without emotional levels) and the other by the new method (with emotional levels). Segmental features were generated using the HMM-based speech synthesis toolkit [12].

Synthetic speeches for 20 sentences by the new method (condition 2) and 10 sentences by the original method were randomized and presented to 12 Japanese, who were asked to check *bunsetsu*'s where they feel higher emotional levels than other parts. The option of no *bunsetsu* with higher emotional level was allowed. When the checked parts coincide with those checked by the speaker (see section 2.1), even if they are partly, they are counted as their emotional levels being correctly realized in synthetic speech. When F_0 contours are generated by the new method, 92.7 % of *bunsetsu*'s with higher emotional level in the original utterances are correctly perceived so in the synthetic speech. The rate decreases to 78.2 % when F_0 contours are generated by the original method.

In order to evaluate how the designated emotion can be conveyed by the new method, another listening test was conducted. Each of 30 sentences was synthesized by both new (condition 2) and original methods and the two versions of synthesized speech were presented to 9 Japanese speakers. They were asked to select the version, to which they felt the designated emotion (anger, for the current experiment) clearer. The version by the new method was selected in 79.3 % probability. These results on the listening tests indicate the validity of adding *bunsetsu*-based emotional levels in realizing designated emotion in synthetic speech.

3. Detection of fillers in spontaneous speech recognition

We have developed a new method of using filler information for continuous speech recognition: to calculate the likelihood of fillers appearing in the decoding process of speech recognition using prosodic features (prosodic module), and, if the likelihood is high, increase the score of the hypothesis with the fillers. As for the prosodic module, a neural network was adopted, though other options were also possible.

3.1. Configuration of the method

Baseline speech recognition engine is Julius, developed as an open-software for continuous speech recognition. The engine conducts quick coarse search (1st pass search) first and then conducts detailed search backwoods (2nd pass search) [13]. The 1st pass is the frame synchronous beam search with (morpheme) bi-gram language model and the 2nd one is N-best stack decoding search with (backward) tri-gram language model. When calculating the likelihood of hypotheses, the weight of the language score to the acoustic score was set to 8.0 throughout the current experiment. The prosodic module calculates probability of a morpheme being a filler (henceforth, filler likelihood score). Although the module can calculate the filler likelihood scores for all the morphemes included in the input utterance, in the current method, it needs to calculate only for those hypothesized to be fillers in the 2nd pass search process. The language score is changed depending on the result of the prosodic module. Our preliminary experiment showed that reducing the language score when the likelihood score being low degraded the final recognition rates. Taking this into account, a certain value (bonus) is added to the language score only when the filler likelihood score exceeds a threshold. Henceforth we call this value as the prosodic score. Since there is no clear difference in the recognition performance, whether the prosodic score is changed according to the filler likelihood score or is kept constant, we set it to a constant value. The threshold and the prosodic score are respectively set to 0.5 and 5 in the experiments shown in section 5. Surely, if we reduce the prosodic score, the number of false filler detection may decrease, but the number of filler recovery by the prosodic module may also decrease.

3.2. Speech material

The speech material used for the experiments is 100 utterances (including one or more fillers) by 7 males and 6 females, which are selected from the corpus of academic meeting presentations included in the Corpus of Spontaneous Japanese (CSJ) prepared under a national project [14]:

http://www2.kokken.go.jp/~csj/public/index.htm/

In the original corpus, all the utterances of each speaker are recorded in a file. So, we first segmented it into utterances and then selected 100 utterances so that each of them includes one or more fillers, and does not include any restatements or coughs. The numbers of fillers in the 100 utterances sorted in the order of frequency are, 185 /eH/, 82 /e/, 16 /sonoH/, 14 /ma/, 13 /maH/, 12 /eQto/, 11 /ano/, etc. (Symbols "H" and "Q" mean elongation of previous vowel and gemination, respectively.)

3.3. Prosodic module

The prosodic module is constructed as a 5-layered perceptron with 3 middle layers, each of which has 20 units. These numbers were decided through some preliminary experiments. The input and output layers have 10 and 1 units, respectively. One unit of input layer accepts each of 10 input parameters. The output layer unit outputs the filler likelihood in the range between 0 and 1.

Figure 1 shows an example on how fillers appear in the F_0 contour of utterance. It is clear that they have low and level contours. Taking this feature into account, four F_0-related parameters such as F_0 range, F_0 gradient, and so on are included into the input parameters. Lengths of immediately preceding and following silences are also included in the input parameters, because they frequently co-occur with fillers as shown again in Fig. 1. In the current method, silences are detected simply searching periods whose waveform amplitudes do not exceeds a threshold.

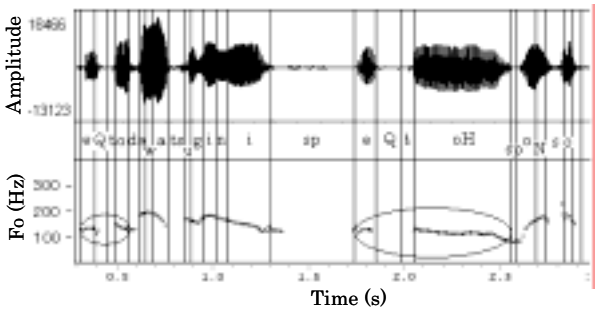

Figure 1. Waveform (upper panel) and F_0 contour (lower panel) for the utterance "eQto dewa tsugi ni eQtoH oNso ([Filler] Then, next [Filler] a phoneme…)" by a male speaker. The underlined morphemes are fillers. The circled parts of F_0 contour are those corresponding to the fillers. "sp" means a short pause.

An experiment of filler detection was conducted for the 100 utterances. First, all the utterances are segmented into phonemes by the forced alignment, and then their F_0's were extracted in order to calculate the input parameters. Twelve utterances were discarded where the input parameters were not properly extracted because of errors in segmentation and/or pitch extraction. Then, the rest 88 utterances (of 6 male and 6 female speakers) were divided into 76 utterances for training and 12 utterances (one utterance from each of 6 male and 6

female speakers) for testing. They include 306 fillers (in total of 2846 morphemes) and 39 fillers (in total of 420 morphemes), respectively. When morphemes with filler likelihood scores larger than 0.5 are assumed to be fillers, 29 fillers are correctly detected out of 39 fillers, while 13 fillers are incorrectly detected out of 381 non-filler morphemes.

3.3. Experiment

Speech recognition experiments were carried out for the 100 utterances using two versions of recognizer: one with prosodic module (proposed recognizer/method) and the other not (baseline recognizer/method). As explained already, the baseline recognizer is Julius for the spontaneous speech provided by the CSJ project. The acoustical (phone hidden Markov) models were trained using 486 hours of academic meeting presentations by 2496 people included in the CSJ corpus. The 100 utterances are included in these training speech samples. The language models were trained using transcriptions of 2592 lectures, which include 6.6 x 10^6 morphemes.

The utterance "kasetsu ga e shiji sa re mashi ta (The hypothesis was accepted.)," was recognized as "kasetsu ga ninshiki (recognize) sa re mashi ta." by the baseline recognizer, while it was recognized as "kasetsu ga e shi (do) sa re mashi ta" by the proposed recognizer. It is clearly shown filler /e/ (underlined in the example) is correctly recognized in the version with the prosodic module. Improvements at non-filler morphemes are also observable in the utterance "e kochira ga eH hana no aru (This one is with a nose…)," which was miss-recognized as "e kochiragawa (this side) eH hana no aru" by the baseline recognizer. It was correctly recognized when the prosodic module was introduced.

Table 4 summarizes changes in the recognition results caused by the introduction of the prosodic module. Seven fillers, miss-recognized by the baseline method as non-filler morphemes, are correctly recognized by the proposed method, while no fillers correctly recognized by the baseline method are miss-recognized by the proposed method. In the 100 utterances, a total of 389 fillers are included and 349 of them are detected by the baseline method. Therefore, 356 fillers are detected by the proposed method. Three non-filler morphemes correctly recognized by the baseline recognizer are miss-recognized by the introduction of the prosodic module. These errors can be avoided by decreasing the prosodic score, but improvement in filler detection also degraded. This type of miss-recognition is tightly related to the (sophisticated) search algorithms of the 2nd pass, such as: when a hypothesis survives beyond a threshold, hypotheses with shorter lengths are terminated. Because of these algorithms, the best hypothesis selected by the 2nd pass is not guaranteed to be really the best one. It is confirmed that all the three morphemes miss-recognized by the introduction of the prosodic module are correctly recognized in the "really" best hypotheses.

Table 4. Numbers of morphemes where the recognition results are changed by the introduction of the prosodic module. "Baseline" and "Proposed" indicate speech recognizers without and with prosodic module, respectively.

(Baseline → Proposed)	Filler	Non-filler
Incorrect → Correct	7	4
Correct → Incorrect	0	3

4. Conclusions

An improvement was realized in the ability of expressing designated emotions in our corpus-based method of generating F_0 contours of emotional speech. Currently, the method is only trained for a speech corpus, and used for realizing the same emotion in the same voice quality. Further research is planned to realize emotional speech for a speaker without speech corpus of that emotion: applying deviations in acoustic features between emotional speech and calm speech of an actor/actress to other speaker's calm speech to generate his/her emotional speech.

A new method of detecting fillers in spontaneous speech during the speech recognition process was developed. Although some errors arose for non-filler morphemes, they were due to the search algorithm of the 2nd pass of the baseline recognizer Julian, and could be recovered by changing the algorithm. Further experiments are planned for increased number of utterances. It is known that speakers use fillers rather differently in their spontaneous utterances. Adaptation methods to cope with this variation are also in the scope of our future work.

5. References

[1] K. Molla, K. Hirose, and N. Minematsu, "Separation of mixed audio signals by decomposing Hilbert spectrum with modified EMD," *IEICE Transaction on Fundamentals of Electronics, Communication and Computer Sciences*, to appear (2006).

[2] K. Molla, K. Hirose, and N. Minematsu, "Localization based separation of mixed audio signals with binary masking of Hilbert Spectrum," *Proc. IEEE ICASSP*, Toulouse, to appear (2006-5).

[3] K. Hirose, K. Sato, Y. Asano and N. Minematsu, "Synthesis of F_0 contours using generation process model parameters predicted from unlabeled corpora: Application to emotional speech synthesis," *Speech Communication*, Vol.46, Nos.3-4, pp.385-404 (2005-7).

[4] Q. Sun, K. Hirose, W. Gu, and N. Minematsu, "Generation of fundamental frequency contours for Mandarin speech synthesis based on tone nucleus model," *Proc. EUROSPEECH*, Lisbon, pp.3625-3628 (2005-9).

[5] Q. Sun, K. Hirose, W. Gu, and N. Minematsu, "Rule-based generation of phrase components in two-step synthesis of fundamental frequency contours of Mandarin," *Proc. International Conference on Speech Prosody*, Dresden, to appear (2005-5).

[6] K. Hirose, Y. Abe, and N. Minematsu, "Detection of fillers using prosodic features in spontaneous speech recognition of Japanese," *Proc. International Conference on Speech Prosody*, Dresden, to appear (2005-5).

[7] J. Zhang and K. Hirose, "Tone nucleus modeling for Chinese lexical tone recognition," *Speech Communication*, Vol.42, Nos.3-4, pp.447-466 (2004-4).

[8] Y. Yagi, S. Takada, K. Hirose and N. Minematsu, "Improved concept-to-speech generation in a dialogue system on road guidance," *Proc. International Conference on CYBERWORLDS*, Singapore, pp.429-436 (2005-11).

[9] Fujisaki, H. and Hirose, K., "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *J. Acoust. Soc. Japan (E)*, Vol.5, No.4, pp.233-242 (1984-10).

[10] Edinburgh University, The Edinburgh Speech Tools Library, http://www.cstr.ed.ac.uk/projects/speeech_tools/.

[11] Kyoto University, Japanese Syntactic Analysis System KNP http://www-nagao.kuee.kyoto-u.ac.jp/projects/nl-resource/.Galatea Project, http://hil.t.u-tokyo.ac.jp/~galatea/regist-jp.html

[12] Galatea Project, http://hil.t.u-tokyo.ac.jp/~galatea/regist-jp.html

[13] A. Lee, T. Kawahara, and K. Shikano, K., "Julius – an open source real-time large vocabulary recognition engine," *Proc. EUROSPEECH*, Aalborg, pp. 1691-1694 (2001).

[14] K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation." *Proc. ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition*, *Tokyo*, pp. 7-12 (2003).