PROSODY IN SPEECH TECHNOLOGY

- National project and some of our related works -

Keikichi Hirose

Department of Frontier Informatics, School of Frontier Sciences, University of Tokyo
Bunkyo-ku, Tokyo, 113-0033, Japan
hirose@gavo.t.u-tokyo.ac.jp

ABSTRACT

Prosodic features of speech are known to play an important role in the transmission of linguistic information in human conversation. Their roles in the transmission of para- and non- linguistic information are even much more. In spite of their importance in human conversation, from engineering viewpoint, research focuses are mainly placed on segmental features, and not so much on prosodic features. With the aim of promoting research works on prosody, a research project "Prosody and Speech Processing" is now going on. A rough sketch of the project is first given in the paper. Then, the paper introduces several prosody-related research works, which are going on in our laboratory. They include, corpus-based fundamental frequency contour generation, speech rate control for dialogue-like speech synthesis, analysis of prosodic features of emotional speech, reply speech generation in spoken dialogue systems, and language modeling with prosodic boundaries.

1. INTRODUCTION

Acoustic features of speech consist of those segmental and supra-segmental. Each phone has its distinctive segmental features and, from this viewpoint, segmental features can be said to directly correspond to characters of written language. Supra-segmental features, on the other hand, are mostly related to vocal folds vibrations and may extend to a range wider than phones or syllables, viz., words, phrases, and so on. They are tightly related to prosody, and are usually called prosodic features. Different from the segmental features, prosodic features have no direct correspondence to written characters and, from this aspect, are peculiar to spoken language. They play important roles when humans transmit information through speech. Surely, segmental features have major roles in the transmission of linguistic information, such as word meanings, the roles of prosodic features come larger for higher-level information, such as syntactic and discourse levels. They come dominant for attitudes and emotions, known as "kansei" information.

Spoken language technologies, such as speech synthesis and recognition, showed a great advancement recently by concentrating mostly on segmental features. This is because the current technologies cover only text-reading (even if not reading text actually, sounding as text reading) speech. When we try to extend the technologies to more spontaneous speech, prosodic features should be more taken account of.

2. RESEARCH PROJECT ON PROSODY AND SPEECH PROCESSING

In view of the necessity of a unified study on prosody with its clear formulation, a new research project on prosody has started from October 2000 for the period of 3.5 years. It is one of Scientific Researches of Priority Areas, supported by Ministry Science, Culture, Sports, and Education, Japanese government.

The project "Realization of advanced spoken language information processing from prosodic features (abbreviated as Prosody and Speech Processing)" has 8 groups to cover from fundamentals to applications of prosody researches: one group for administration and the other 7 groups for individual researches. In order to avoid the research goal being diverse, a rather limited number of researchers (4 to 6) belong to each group. Totally, 40 members are involved in the project. Adding to these official members, a large number of researchers, mainly graduate students, are also involved. The project will last till March 2004 with annual budget around 140 million Japanese yen. The research results are reported at plenary meetings twice a year and other intra-group meetings. Besides the plenary meetings, "Symposium on Prosody and Speech Processing" is held annually with several invited speakers from foreign countries. The first workshop was held on January 31, 2002 in Tokyo with two invited speakers. Next meeting is scheduled on February 6, 2003 also in Tokyo. Computer programs and databases developed through the project will be open for academic use.

Administration Group, headed by Keikichi Hirose, handles various affairs for a smooth operation of the project, including regulation between research groups, organization of academic meetings, publication of research outputs, and so on.

In order to clarify how prosodic features are related to convey linguistic and para-/non- linguistic information, a good (quantitative) modeling of prosody is necessary. From this viewpoint, Modeling Group (Analysis of Prosody, its formulation and Modeling) headed by Hiroya Fujisaki, professor of emeritus, University of Tokyo, conducts research works with a focus on the generation process model of fundamental frequency (F_0) contours (known as Fujisaki's model). A program on the automatic estimation of model parameters will be developed and opened to the public use. The relationship of the model with other prosody models will also be investigated.

Prosodic features of speech are subject to change by various factors, such as, speaking styles, situation effects, individualities, regional effects, and so on. Multiplicity Group (Variability of Prosody and its Quantitative Expression) headed by Masuzo Yanagida, Doshisha University, will tackle this prosodic variability problem. Emotional speech is in the research scope. Collection of speech in real situation is also aimed at.

Currently, corpus-based methods are successfully applied in speech technology areas. Corpus Group (Design of Prosodic Corpora and Automation of the Developing Process) headed by Shigeyoshi Kitazawa, Shizuoka University, will provide several types of prosodic databases (speech database with prosodic labels) to meet various requirements from other groups. They include reading style speech, semi-spontaneous speech, and so on. Some databases will include expression pictures. Several existing prosodic labeling scheme will be checked to produce a new scheme. Labels will include the commands of the generation process model of F_0 contours. Automation of labeling is also a major concern of the group.

The aim of Synthesis Group (Prosody Control for High-Quality Speech Synthesis), headed by Keikichi Hirose, is to establish an advanced synthesis technology of prosodic features, enabling to generate synthetic speech highly human-like in various utterance styles. Scope of the study is not only limited to linguistic information, but also covers para- and non- linguistic information. Realization of a spoken dialogue system, whose reply speech is highly acceptable for users, is also aimed at. Although both of heuristic and statistical frameworks are introduced for the research works, major focuses are place on two corpus-based methods for F_0 contour generation; HMM-based method and method based on the generation process model.

Recognition Group (Use of Prosodic Information in Speech Recognition and Understanding) headed by Kazuhiko Ozeki, University of Electro-Communication, pursuits research works with the final goal of developing sophisticated ways to incorporate prosodic features in speech recognition process. Currently a major focus is placed on the linguistic information contained in the prosodic features, though para- and non-linguistic information is in the research scope. Several novel schemes to exploit prosodic information for various problems in speech recognition and understanding have been developed already, including use of dependency analysis of Japanese sentences, combined use of phonemic and accent information, summarization of spoken materials, and so on. There are several contributions from our laboratory, such as prosodic boundary detection using mora transition modeling, dynamic beam search using prosodic boundaries, separate modeling of word transitions across and not across prosodic boundaries, and so on.

Dialogue System Group (Discourse System with Enhanced Prosody Control) headed by Tetsunori Kobayashi, Waseda University, investigates the relation between prosodic features and various dialogue phenomena. A modeling will be constructed through analyzing human dialogues for prosodic control during conversation. Several dialogue acts, including turn taking, will be clarified through the modeling. The results together with those obtained other groups will be tied up to realize an advanced dialogue system.

Medical Application Group (Applications of Prosody processing technology to the Field of Medical and Welfare) headed by Hideki Kasuya, Utsunomiya University, concentrates on the prosody disorders. In parallel with basic understanding of the physiological and physical mechanisms of prosody control, acoustic properties of speech utterances of the patients are examined. Its final scope is to explore a useful means for the examination, evaluation and enhancement of the prosody disorders as well as to develop aiding devices for the patients.

3. CORPUS-BASED GENARATION OF F_0 CONTOURS UNDER GENERATION PROCESS MODEL CONSTRAINTS

In corpus-based methods for F_0 contour generation, F_0 movements can be directly related to linguistic information of the input texts. An HMM-based method successfully generated synthetic speech with highly natural prosodic features by counting F_0 delta features [1]. These methods without F_0 model constraints theoretically can generate any type of F_0 contours, but have possibility of causing un-naturalness especially when the training data are limited. Several methods are reported under the ToBI labeling strategy. Constraints by the ToBI system are beneficial in avoiding unlikely F_0 contours being generated. The major problem of ToBI system is that it is not a full quantitative description of F_0 contours, which causes some limitations to the quality of synthesized F_0 contours.

From these considerations, a corpus-based synthesis of F_0 contours in the framework of the generation process model was developed [2]. By predicting the model commands instead of F_0 values, a good constraint will automatically applied on the synthesized F_0 contours; still keeping acceptable speech quality even if the prediction is done incorrectly. Although current constraints are limited to the model's command response features, further constrains are possible based on various knowledge on model commands, such as on command timing as compared to the segmental boundary locations.

In our method, prediction of F_0 model parameters is done for each accent phrase, and a sentence F_0 contour is generated using the F_0 model after the prediction process is completed for all the constituting accent phrases. Input parameters for the statistical methods are therefore, linguistic features of the accent phrase in question, such as position in sentence, mora numbers, accent type, and so on. Besides, the linguistic information of the preceding phrase and the syntactic boundary depth between the two phrases are added. As for the latter parameter, information on the direct modifier obtainable through text analysis using the Japanese text parser

KNP [3] is used with no manual correction. Output parameters are the model command amplitudes and timings, represented as offset values from corresponding mora boundaries. Binary decision tree (BDT) and multiple linear regression analysis (MLRA) are used as statistical methods.

Experiments on F_0 contour generation by the method were conducted using the ATR continuous speech corpus of 503 sentences [4]. Utterances by male speaker MHT were selected, since J-ToBI labels were attached. Among them, 388 sentences (2803 accent phrases) and 48 sentences (262 accent phrases) were used as the training data and test data, respectively. Small mean square error between the sentence F_0 contour generated using parameters predicted by the method and that of the "best" approximation by the model indicated the validity of the method. This was fuurher confirmed by the subjective evaluation: listening test of the synthetic speech.

4. SPEECH RATE OF DIALOGUE SPEECH

In most current spoken dialogue systems, speech output is generated using text-to-speech (TTS) conversion devices (or software packages) commercially available. Although speech quality from these devices has recently improved considerably, there are still several problems to be overcome. One of such problems is that these devices are designed to synthesize read speech and, therefore, intonation and rhythm of the synthesized speech is often too monotonous for users. Dialogue speech generally shows wider dynamic ranges in its prosodic features than read speech. From this respect, we have been conducting a comparative study on the prosody of dialogue-style and that of reading-style, to construct prosodic rules for dialogue speech synthesis. As for the F_0 contours, differences between dialogue speech and read speech were analyzed based on the method similar to the previous section, and developed the prosodic control rules for dialogue-like speech synthesis [5]. Here, our research on another aspect of prosody, speech rate, is introduced [6].

Although precise analyses on segmental duration have been conducted for read speech, there have been relatively few studies for dialogue speech. In the case of dialogue speech, increased number of factors, including speaker-to-speaker variation, will affect the duration to a large extent, thus making precise analysis difficult. Therefore, instead of directly looking at segmental duration of dialogue speech, we analyzed its relative value as compared to read speech. By doing so, influences of various factors on segmental duration are suppressed, allowing us to find out clearly how speech rate of dialogue speech differs from that of read speech. Since prosodic rules for reading-style speech synthesis is already available, speech rate control for dialogue-style speech synthesis is possible by modifying segmental duration of read-style depending on the result.

Based on the model of F_0 contour generation, four prosodic units, prosodic sentence, clause, phrase and word, are defined. Speech rate was analyzed with respect to these units, especially to prosodic phrases. Reduction rate (of mora length) of dialogue speech as compared to its read speech counterpart is defined and is used for the analysis. Through the analysis of speech samples recorded during simulated dialogues, it was found that, in a prosodic phrase, dialogue speech rate starts with a value slightly larger than that of read speech. Then, it gradually increases and after passing through the middle of the phrase, decreases. This result was also supported through a linear regression analysis and a hearing test of synthetic speech.

5. PROSODIC FEATURES OF EMOTIONAL SPEECH

There are increasing interests in realizing and recognizing emotion conveyed by speech, and, therefore, a rather large number of analyses were conducted on prosodic features of emotional speech. Although general tendencies of prosodic features of emotional speech as compared to those of neutral utterances were clarified, such as F_0 increase in joy, the prosodic control done for the realization of emotional speech was rather ad hoc. In this paper, discussions are made through the analyses of F_0 contours and segmental duration on how humans control the prosodic features to express degrees in emotional speech [7].

In order to collect speech samples with several emotional degrees, scenarios were arranged where conversations were conducted between speakers A and B. Responding to the speaker A's questions/requests, speaker B repeats utterances with the same content but with increased emotional levels as the dialogue proceeds. The emotional levels are 5 including a neutral one. Two semi-professional actors of the Tokyo dialect were asked to simulate dialogues by referring to the scenarios. A scenario was arranged for each of anger, joy and sad.

Analysis of F_0 contours and that of segmental durations were conducted for speaker B's utterances. F_0 contours were analyzed by the method of analysis-by-synthesis using the generation process model. As for the segmental duration, the mora reduction rate (defined in section 4) was calculated as the index representing speech rate increase in emotional speech from neutral speech.

Although the general tendencies of emotional speech were observable in all the levels, it seemed all the tendencies did not become apparent evenly as the level increased. Humans may have several ways to express emotion and they may select one other than use them all.

It is said that segmental features are also important in realizing emotional speech. In order to reveal this, perceptual experiments were conducted using synthetic speech, where only one from F_0 contour, mora duration, power and spectrum (cepstral coefficients) of neutral speech was substituted to that of target speech. The results indicated the importance of segmental features was different depending on the type of emotion: large in the case of joy. It was also revealed that F_0 contribution was rather large for joy and sad.

6. SPEECH REPLY GENERATION IN SPOKEN DIALOGUE SYSTEM ON ACADEMIC DOCUMENT RETRIEVAL

As pointed out in section 4, speech output from most spoken dialogue systems is generated simply using TTS conversion devices. During reply sentence generation process, the system may have rich information, such as important words, syntactic structure of the sentence and so on, which should be reflected on prosody of reply speech. However, this process is rather difficult when we utilize commercially available TTS devices. Moreover, misreading may occur because of wrong linguistic processing in the TTS devices.

From this point of view, a scheme was developed in our system for academic document retrieval to directly generating speech reply from reply contents [8]. By doing so, we can realize speech reply with its prosodic features properly controlled to express syntactic structures and dialogue focuses. Sets of simple rules were developed for focus positioning and focus expression. The focus positioning rules include: to place a focus on the words conveying answering information. The focus expression rules are those slightly modified from the rules of dialogue prosody generation [5]. The validity of the rules was verified through evaluation experiments using the system. It was indicated that there existed users' preferences on the intonation of the reply speech.

7. N-GRAM LANGUAGE MODELING USING PROSODIC BOUNDARIES

Although control of prosodic features is one of major concerns in current speech synthesis, research works are rather rare on the use of prosodic features for speech recognition. We can even say that prosodic features have been expelled from speech recognition process. However, further advancements of speech recognition require a good use of prosodic features. Several methods developed in our laboratory were already reported in ICSP2001 [9]. In this paper, a new scheme is introduced, where prosodic boundary information is incorporated into N-gram language modeling [10].

The current statistical language modeling, known as N-gram, is only for written texts. As outputs of human process of sound production, spoken sentences cannot be fully represented only by written language grammars. Prosodic features are considered to represent structure of speaking, and should also be counted in the language model level. This consideration led us to an idea of separately modeling the word transitions for the two cases: one crossing and the other not crossing accent phrase boundaries. Since counting such transitions requires a large speech corpus, which hardly can be prepared, part-of-speech (POS) N-gram was first counted for a small-sized speech corpus for the two cases instead, and then the result is applied to word N-gram counts of a large text (newspaper) corpus to divide them accordingly. Thus, two types of word N-gram model can be obtained. Using ATR continuous speech corpus by two speakers, perplexity reduction from the baseline model to the proposed model was calculated for the word bi-gram. When accent phrase boundary information of the speech corpus was used, the reduction reached 11%, and when boundaries were extracted using our formerly developed method based on mora-F_0 transition modeling [11], it still exceeded 8%. The reduction around 5% was still observed for sentences not included for the calculation of POS bi-gram and using boundaries automatically extracted from another speaker's speech. The obtained bi-gram was applied to continuous speech recognition, resulted in a two-percentage improvement of word accuracy from when the baseline model was used.

8. CONCLUSION

After briefly introducing the Japanese national project on prosody and speech processing, some of our prosody research works related to synthesis were explained. A new scheme for speech recognition was also introduced. Corresponding to the increasing interest on prosody in speech processing, the first Prosody International Conference was held in this April at Aix-en-Provence with participants of more than 400. A special session will be held at ICSLP2002, Colorado on the topic of Prosody and Speech Recognition. Importance of prosody research may increase in the future research works, where speech other than read one should be dealt with.

9. REFERENCES

[1] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multispace probability distribution for pitch pattern modeling," Proc. ICASSP, 229-232 (1999).

[2] K. Hirose, M. Eto, and N. Minematsu, "Improved corpus-based synthesis of fundamental frequency contours using generation process model," Proc. ICSLP, to be published (2002-9).

[3] Kyoto University, Japanese Syntactic Analysis System KNP http://www-nagao.kuee.kyoto-u.ac.jp/projects/ nl-resource/.

[4] Speech Corpus Set B. http://www.red.atr.co.jp/ database_page/digdb.html

[5] K. Hirose, M. Sakata and H. Kawanami, "Synthesizing dialogue speech of Japanese based on the quantitative analysis of prosodic features," Proc. ICSLP, Vol.1, pp.378-381 (1996-10).

[6] K. Hirose and H. Kawanami, "Temporal rate change of dialogue speech in prosodic units as compared to read speech," Speech Communication, Vol.36, pp.97-111, Nos.1-2 (2002-1).

[7] K. Hirose, N. Minematsu and H. Kawanami, "Analytical and perceptual study on the role of acoustic features in realizing emotional speech," Proc. ICSLP, Vol.2, pp.369-372 (2000-10).

[8] S. Kiriyama, K. Hirose, and N. Minematsu, "Control of prosodic focuses for reply speech generation in a spoken dialogue system of information retrieval on academic documents," Proc. Speech Prosody, Aix-en-Provence, pp.431-434 (2002-4).

[9] K. Hirose, "Prosody, an important feature for advanced spoken language processing technology," Proc. ICSP, Taejon, V1, Vol.1, pp.35-40 (2001-8).

[10] K. Hirose, N. Minematsu, and M. Terao, "Statistical language modeling with prosodic boundaries and its use for continuous speech recognition," Proc. ICSLP, to be published (2002-9).

[11] K. Hirose and K. Iwano, "Detection of prosodic word boundaries by statistical modeling of mora transitions of fundamental frequency contours and its use for continuous speech recognition," Proc. ICASSP, Vol.3, pp.1763-1766 (2000-6).