# A Method of Speech Synthesis for Spoken Dialogue Systems and Psychological Assessment of the Synthetic Speech

**Keikichi HIROSE, Noboru TAKAHASHI, Toru SENOO, Mayumi SAKATA and Hiromichi KAWANAMI**
Department of Information and Communication Engineering,
Faculty of Engineering, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, 113 Tokyo, JAPAN
e-mail: hirose@gavo.t.u-tokyo.ac.jp

## 1. Introduction

In order to realize the smooth communication, the response speech should not only be with high-quality but also be easily understood by the users. Therefore, linguistic information, such as syntactic structure and discourse structure, should well be conveyed by the prosody of synthesized speech. Since, in the case of text-to-speech conversion, it is hard to extract this kind of higher-level information from input texts, prosodic rules developed for the conventional text-to-speech conversion system are far from enough for the above requirement. In the case of output generation in dialogue systems, higher-level information is available during the process of sentence generation. Another important issue in realizing smooth communication is the generation of good answer. Since we have a limited size of short term memory, much information should not be conveyed simultaneously by an utterance. However, on the contrary, too small information per an utterance may increase the number of interactions between system and user to convey necessary information. There should be a compromise between these two extremes. Another compromise should also be considered on how to generate a surface sentence from a semantic representation. If known information is included in the surface sentence, it is tedious to the user, but, in some situations, is useful to avoid the misunderstanding between the user and the system. When the input to the system is conducted only by speech, as is often the case of spoken dialogue systems, the possibility for confusion may increase in the dialogue flow because of misrecognition of the input. Therefore, in such the cases, duplication of known but key information comes more useful to keep the coherence of the dialogue. If the response sentence includes such a duplication, the user can easily extract the necessary information from the speech reply when the words conveying such information being prosodically emphasized, and, thus, can maintain an efficient dialogue.

The final goal of our project is to construct a system for generating response speech (Fig. 1), which should be included in the total communication system with spoken dialogue. Based on the considerations above, the following three items were selected as major topics of the research:

1) Generate sentences from response contents of deep-level semantic representation, which may include information on focal position, ellipsis and anaphora.

2) Synthesize response speech with prosodic features, naturally sounding as dialogue speech and well representing high-level linguistic information.

3) Synthesize high-quality speech also from the viewpoint of segmental features.

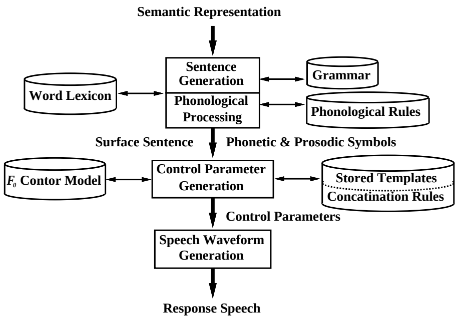


Figure 1: Total configuration of a system for generating response speech.

As for the first item, we have developed a method of dialogue processing in a spoken dialogue system with questions and answers. We also have developed a method of controlling ellipses and focal positions based on the degree of novelty of the information. As for the second item, we have conducted a comparative study on the prosody of dialogue-style and reading-style and have constructed prosodic rules for the synthesis of speech output in spoken dialogue systems. As for the third item, we have adopted a formant synthesizer with multiple paths previously developed. Recently, a rather good quality is obtained for synthetic speech with a scheme of waveform concatenation. This scheme, however, was not adopted, because it lacks flexibility in prosodic control and is not appropriate for the current research. In the coming sections of this report, the above results are partly described.

**2. Response sentence generation [1]**

A method has been developed for the generation of speech reply with proper elliptical expressions and prosodic emphases in a dialogue system on weather conditions, which was designed to facilitate the maintenance of electric power systems by supplying necessary weather information to a patrol man. Although the system has several modes of conversation with the user, the developed method is mainly related to the major mode of questions and answers.

In order to make an access to the database and to find out necessary information for the user, the system needs information on "date/time, location and content of questions," which are the case elements of the semantic representation of user input and also of system response. A stack was prepared for each of these three items to manage the dialogue flow. By restricting the dialogue to that of questions and answers, and by forbidding the sudden jump in the dialogue topics, the dialogue flow can be represented by pairs of answers and questions called "fundamental routines of dialogue (FRD)." Case element fillers in user inputs and system responses are stored sequentially in the corresponding stacks with FRD numbers, which increase as the dialogue proceeds. If the user question does not include fillers for all of the above three items, compensation is conducted by referring to the content of the stacks without asking back to the user. If compensation is failed, the system asks back to the user to give the lacking information. When all of the necessary information is obtained, the system makes an access to the database and generates the answer in semantic representation of the list form.

For each case element fillers of the semantic representation, a flag was selected out of four flags +1, 0, -1, -2. Flag "+1" was assigned to the case element filler of the target information, viz., the key words of the answer. The assignment of other flags was conducted by comparing the number of current FRD with that of FRD attached to the same filler in the stack. If the difference between the numbers exceeded 2 (or if the same filler could not be found), flag "0" was assigned, and, otherwise, flag "-2" was assigned. Flag "-1" will not appear in the current method, but is reserved for the control of prosodic emphasis as explained later. These flags are utilized for the control of elliptical expressions and prosodic emphases, such as to include case element filler with flag "+1" in the surface sentence with prosodic emphasis.

In the developed system, weather information is assumed to be supplied from the Meteorological Agency of Japan at every three hours. Since each issue covers a limited period, there will be many cases where the system cannot find information required by the user. Several schemes were included in the method to avoid the answer only of "We have no information."

## 3. Analysis and synthesis of prosodic features [2, 3]

With the aim of constructing prosodic rules for dialogue speech by modifying those already developed for read speech, a comparative study was conducted on the prosody of dialogue-style and that of reading-style. Several actors and actresses of common Japanese were asked to simulate a dialogue by referring to written texts on ski tour reservations. They were also asked to read individual sentences of the texts in a normal reading-style. These utterances were recorded and digitized to serve as the material for the analyses of prosodic features. Analy ses were conducted on fundamental frequency contour (henceforth, F0 contour), speech rate, and waveform power. As for F0 contours, analysis was conducted based on the F0 contour model, which enables the decomposition of a contour into phrase and accent components, and, thus realizing a quantitative representation of the contour with discrete phrase and accent commands.

The following general features were observed in F0 contours of dialogue speech as compared with those of read speech: 1) Larger mean value and standard deviation of F0, which can be explained as due to a larger baseline and larger command values of F0 contour. 2) Larger dynamic range in the command values. 3) Larger command values for words conveying key information. 4) Larger dispersion in the degree of F0 rise and delay in the onset timing of F0 rise for the sentence-final particle /ka/ for interrogation.

In order to make clear to what kind of words prosodic emphasis should be given, analyses were further conducted on how the amplitude of accent command increases in dialogue speech as compared with read speech, after classifying each word carrying an accent component according to its role in the dialogue: 1) contains the main target information of the question or stands for answer to a previous question, and is not a repetition (T), 2) essential for the understanding of the sentence and is not a repetition (E), 3) essential for the understanding of the sentence but is a repetition (RE), 4) not essential for the understanding of the sentence and is a repetition (R), 5) stresses, emphasizes or modifies the understanding of an essential or a target word (S), 6) interrogative pronoun (Q), and 7) others (O).

Since an accent component of F0 contour shows rather different feature depending on its location in a sentence and on its accent type, results were summarized separately for two types of positions and for two types of accent, viz., locating at the prosodic sentence initial (position 1) or not (position 2), and having a sharp F0 downfall (type D) or not (type F). In the case of position 1, no significant difference was found among distributions. For position 2, some significant differences were observed for the case of type D words: larger increase in group S and almost no increase in groups R and Q.

Based on the above results, several factors were selected which may affects the amplitude of accent commands. Concretely, 4 items (1. carrying a new/important information, 2. type of accent, 3. position in the phrase component, and 4. part of speech) were selected, with each item divided into several categories. A quantification method was then applied to represent the effects of factors on the command amplitude numerically. In general, larger effects were observed for dialogue speech. The results also showed larger effects of items 1 and 4. For items 2 and 3, command amplitudes were larger for type D accent and position 1, as shown already.

The quantification analysis was also conducted for the phrase command magnitudes, by taking 8 items (opening/closing an FRD, containing word(s) carrying important information, changing topics, number of morae of the phrase, and so on) into account. The results include: 1) When opening FRD, command magnitudes are larger for the sentence initial command and consequently are smaller for the following command. 2) When important words included in a phrase component, the corresponding command magnitude comes larger.

As for the speech rate, reduction rate of syllable duration from reading-style to dialogue-style was calculated. Generally, larger values and wider dynamic ranges were observed for dialogue speech. Results also showed that words with important information are not necessarily elongated. After normalizing the sentence length, the reduction rate was averaged over utterances without complicated syntactic structure. It was shown that, in a sentence, the speech rate of the dialogue-style started with a value close to the case of reading-style, gradually increased, and then slowed down toward the end of the sentence. Various factors, such as word length and syntactic structure, modify this fundamental tendency.

Based on the above results, prosodic rules were constructed for dialogue speech synthesis. As for the F0 contours, magnitude/ amplitude of phrase/accent command is represented by a linear equation, which is defined by the results obtained by the analyses with quantification method. As for the speech rates, an average reduction rate was defined as a function of position in a sentence and superposed on each original syllable duration. Synthesis was conducted using both former rules developed for read speech and these newly developed rules. A hearing test confirmed that the new rules can produce better prosody as dialogue speech.

## 4. Conclusion

Results were shown for the research work on the generation of response speech in spoken dialogue systems. Although the final goal of the research work was realized to a certain extent, further efforts are necessary for the improvements on speech quality and sentence generation for more complicated task.

## References

[1] K. Hirose and T. Senoo, "A method of generating speech reply with elliptical expressions and prosodic emphases," Proc. ESCA Tutorial and Research Workshop on Spoken Dialogue Systems, pp.233-236 (1995-6).

[2] K. Hirose, M. Sakata, M. Osame and H. Fujisaki, "Analysis and synthesis of fundamental frequency contours for the spoken dialogue in Japanese," Proc. 2nd ESCA/IEEE

Workshop on Speech Synthesis, pp.167-170 (1994-9).

[3] M. Sakata and K. Hirose, "Analysis and synthesis of prosodic features in spoken dialogue of Japanese," Proc. 4th EUROSPEECH, Vol.2, pp.1007-1010 (1995-9).

Key words: sentence generation, dialogue speech, speech synthesis, F0 contours, speech rate, prosodic emphasis