Conversational speech synthesis model for virtualized-egos

Yoshihiro Toyama *1 Li Qing *2 Toyoaki Nishida *2 Sadao Kurohashi*2

*1 Graduate school of Engineering, the University of Tokyo  *2 Graduate school of Information Science and Technology, the University of Tokyo

A virtualized-ego is a conversational agent that represents a member of a community and can talk on behalf of a real member. We propose a conversational speech synthesis model for virtualized egos that can handle natural spoken characteristics on the view of the structure of discourse. We show how speech sound with the naturalness of spoken speech is created using this model. We also show the details of speech sound analysis to illustrate the relation between discourse structure and prosodic features.

1. Introduction

As the computer science and media technology advanced, the way people interact with each other has been changing. Intelligent Agents mediate people in getting to know and communicate with each other. A virtualized agent is a conversational agent that represents a member of a community and can talk on the behalf of the real member [1]. But the voice output is far from the level. It seems a non-native speaker to the deficiency that inspires us to construct a conversational speech synthesis model for virtualized egos (VEs).

Within technical advances in the last two decades, a speech synthesis system is now capable of producing synthetic speech with intelligibility approaching that of natural speech. It is generally accepted that linguistic structures are essential in communication by means of synthetic speech. Therefore, researches on the relationship between prosodic characteristics and the structure of discourse are expected to distinct the role the discourse of structure plays on the prosody and improve the synthetic speech.

In this paper, we propose a conversational speech synthesis model which enables VEs to speak on behalf of a real member in a community. In order to construct such a model we analyze the relationship between prosodic characteristics and the structure of discourse in details.

2. A conversational speech synthesis model for virtualized egos

We propose a conversational speech synthesis model shown as Fig.1. There are two special requirements need to meet because it is a model for virtualized ego S.

- One is to represent a real member to give a presentation in a natural casual way like a discourse on a symposium.
- The other is to compile separate sentences collected from a real member’s data such as past mailing-lists and text chatting histories into a text with a relatively clear relationship between sentences automatically.

To meet these requirements we construct a model with an architecture shown in Fig.1. Firstly, separate sentences from a real member’s daily text are input to a speech synthesis engine to form voice waveform and to linguistic structure analyzer to analyze the structure of the input discourse based on the morphological information and a dictionary which is made in advance. Next, the output discourse structure is sent to the prosody analyzer to obtain the prosodic parameters with the help of the data on prosodic characteristics of discourse structure and speaking style, respectively. Finally, prosodic characteristics are added to the original waveform to achieve a natural speech waveform in the final waveform synthesizer.

As shown in Fig.1, the model is composed of four components according to their different functions at different levels. Though every one of them severes for different functions in the model, it is their cooperation that makes the whole model work. Every component is essential for a conversational speech.
synthesis model for VEs. In the following are the four components described.

1. Components of speech synthesis engine and waveform synthesizer

The two components serve to transform text to voice output in the whole model. Though the two components play different roles at different levels, their roles can be realized by utilizing a simple TTS system which can free us from the difficulty of constructing the model from the beginning. We introduce them into the model as a fixed component set. In our research we employ a speech synthesis software named SmartTalk3.0 [4] because it can allow us to tune the speech rate, pause and intonation to some extends and simplify the modeling of prosodic rules.

2. Linguistic structure analyzer

The component is to compile separate sentences into a discourse as the input of the following prosody analyzer. The component plays the role of analyzing the structure of discourse on the base of the morphological information supplied by JUMAN [5] and a dictionary on the structure of discourse made in advance. Building the component needs the help of natural language processing techniques. As proposed above, the input sentences are extracted from the daily texts of a real member. The component should be able to not compile these separate sentences into a text simply, but it is more important that it should present a relative clear relationship among these sentences. In fact, the above process is a challenging work because the relationship among individual sentences turns out to be ambiguous in most cases with fewer conjunctions which are concerned to be important clues to analyze the structure. The temporary solution is to encourage users to extend the type of information presented to the model, for example, ask them to note sentences with linguistic tags.

3. Prosody analyzer

The component is the central part of the model because of the role prosody plays in the communication by VEs. The component realizes adding prosodic information into a discourse from the component of linguistic structure analyzer and passing the processed discourse to waveform synthesizer. The prosodic information assisting the realization of prosody controlling should include both the prosodic characteristics of discourse structure and speaking style. A real intelligent agent to mediate people interact with each other should be able to speak in the same speaking style as a real member because the style of speaking addresses central issues in the domain of speech communication. In the communication by VEs, a component having the ability to control the style and emotion of synthetically generated spoken language is expected. Besides, it is well-accepted that the structure of discourse plays a crucial role on the naturalness of synthesis spoken speech because it includes the information not only from the structure but also from the meaning. To obtain the prosodic information related to the structure of discourse is the first step in our researches.

From the description of the four components it becomes clear that combining speech synthetic approaches, natural language processing, and linguistic structure analyzing, and prosody controlling into a conversational speech synthesis system for VEs is the major feature of the model. The model requires not only the sufficient supports from the separate components at different levels but also the cooperation among them. We begin our researches from analyzing the relationship between the structure of discourse and prosody. The details are shown in the third section of this paper.

3. Prosody analyzing based on the structure of discourse

The structure of discourse is closely allied to pragmatics and linguistics, perhaps parts of their combination, and is considered to be the study of the organization and the dynamics of conversation. It conveys the boundaries between topics, between paragraphs with a topic, and between sentences within a paragraph. These boundaries are generally realized by different durations of pausing, speech rate changes. Within a single sentence or utterance, small phrases may be realized shorter pauses or by pitch changes. What matters is the making of structure and salience: pauses and emphasis must be placed correctly, and the hierarchy of phrasing and prominence must be adequately conveyed.

In previous researches on texts, monologues and dialogues the following has been pointed out according to the different position in the discourse:
- At the beginning: the difference of maximum and minimum of F0 (pitch range) is large; the pause of utterance is long; Speech is slow relatively [6] [7].
• In the middle: speech rate become faster [7].
• In the end: the pitch range become small and speech become slow [6] [7].

3.1 Data and standards base on discourse structure

In our researches we analyze the prosodic information of five common people’s live records addressed on different symposiums. On symposium lecturers propose their ideas and the problems which they want to be discussed so that the speech is closer to the dialogues in our daily lives. The feature that both of them need the active attendance of the audience and their behaviors are interactive makes the data we analyze different from the speech announced on TV or recorded in labs. Due to the naturalness of the data of speech we can distinct the behaviors are interactive makes the data we analyze different from the speech announced on TV or recorded in labs. Therefore we make some necessary manual modifications on the original voice data according to the corresponding draft provided by the authors.

Based on Japanese grammar and the structure of sentence, we define that a subject in one sentence is a noun or pronoun in a phrase which contains class particles for

In addition of pauses after punctuations we study the distribution of pauses without punctuations. After the conjunctive particles except for

If there are parts of speech for conversion like

Besides, pauses are inserted after the phrases containing the two kinds of words.

(3) The prosodic characteristics of independent words

If these words locate in the beginning of one sentence, they are spoken in lower speech rates than in the middle of one sentence. Analyzing the differences among words, we find that conjunctions, adverbs and pronouns are spoken relative slowly. And adjective verbs, adjectives and verbs present higher values in the three aspects of F0 range, F0max and F0min than other kinds of words.

We analyze the prosodic characteristics from the relationship between sentences to the relationship inside phrases and obtain the above results. We consider the similarity of prosodic characteristics between sentences to the relationship inside phrases and the structure of discourse. It is the above results and the prosodic data we obtain from the analysis that we base on to construct a conversational speech synthesis model for VEs. Another goal of our analysis is to create some knowledge of the prosodic characteristics on the point of discourse structure to share with other researchers on speech synthesis.

3.3 Evaluation

We model the prosodic rules into the conversational speech synthesis system proposed in the second section of this paper. When we deal with pauses, we distribute pauses as major pause
or minor pause according to the prosodic data we obtain from the above analysis. In our researches, we take use of the software of variations on the speech rate and pauses to complete prosody controlling.

Two kinds of speeches are synthesized, one processed in prosody based on what we obtain in the stage of analyzing prosodic characteristics and the other produced by the original speech synthesis software. After comparing the perceptual effects of the two synthetic speeches, it turns out that the former shows in a more natural way than the original one.

3.4 Conclusion

We analyze the relationship between the structure of discourse and prosody, and apply the rules to a conversational speech synthesis system. To clarify the effect prosody plays on discourses, we classify discourses into three categories from sentences boundary to independent words boundary to facilitate the prosodic analysis. Basing on the categories, we analyze the prosodic characteristics (pauses, speech rates and the characteristics in the domain of F0) of five common people’s presentation records and extract three outstanding rules: the prosodic characteristics about the meaning of conversion is outstanding not only in the aspect of pauses and speech rate but also in the aspect of F0; due to the unique features of Japanese language, pauses are related to the conjunctive particles closely; the location of words in one sentence plays an important role on the characteristics of F0 domain. In the stage of modeling these rules, we synthesize two kinds of synthesis speeches to compare. From the point of perceptual effects the one processed on the prosody in pauses and speech rates speaks in a more natural way.

It is worthy to be mentioned that when we analyze the relationship between the structure of discourse and prosody we pay less attention to the differences of the prosodic characteristics among different five people. For example, they speak in different rates, as we know, at different speech rates the relations between pausing and boundaries is likely to change. The problem can be solved by taking comparison among the records with near speech rates. In addition, when we find that the trend of the changes in prosodic characteristics vary with different people, it is regarded that there is no general rules roughly instead of analyzing the deep meaning due to the differences. We believe there must be something important in the differences, perhaps the key of personal traits. It is likely to analyze the relations between personal prosodic characteristics and the structure of discourse more deeply provided more data from one person. Another one is that the data in our researches are live records of lectures with different contents, which enables us to find some relatively subjective relationship between the structure of discourse and prosodic information in discourse but also exposes us in the difficulty of clarifying the relationship between the prosodic characteristics and the meaning of discourses itself. It is possible to be solved by analyzing the voice data recorded in labs but at the cost of losing naturalness.

4. Future work

By now we have analyzed the voice data from five people and obtained the outstanding features of prosodic characteristics across them. However, though the synthesis speech processed by the rules is closer to spoken speech, what we have done is only the beginning of constructing a conversational speech synthesis system for VEs because the personal traits have not been the emphasis of our researches. In order to synthesize speech containing personal traits on the view of the structure of discourse, it is necessary to analyze a great amount of data from one person. We will do this research in the future.

We recommend applying the prosody controlling to a conversational speech synthesis for VEs, modeling the prosodic information on the structure of discourse into the system is only one of aspects to improve the support of prosody controlling. Since our final goal is to enable VEs not only to speak naturally but also to talk on the behalf of a member, the synthesis speech should hold more variations. To fulfill the variations, introducing speaking style controlling into our present speech synthesis model is another emphasis of our future work.

References