Long-term Analysis of Prosodic Features of Spoken Guidance System User Speech

Hiromichi KAWANAMI, Takahiro KITAMURA and Kiyohiro SHIKANO

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara zip-code 630-0101, Japan
kawanami@is.naist.jp

Abstract

As a practical information guidance system, we have been developing a speech-oriented system named "Takemaru-kun". The system has been operated on a public space since Nov. 2002. The system answers to user's question about the hall facilities, sightseeing, transportation, weather information around the city, etc. All triggered inputs to the system have been recorded since the operation started. And all system inputs during 22 months are manually transcribed and labelled for speaker's gender and age category. In this paper, we conduct a long-term prosody analysis of user speech to find a clue to obtain user's attitude from a user's speech. In this preliminary analysis, it is observed that F0 decreases regardless of age and gender category when the stability of the dialogue system is not established

1. Introduction

To realize a practical information service, we have been developing a speech-oriented guidance system named "Takemaru-kun"[Nisimura 2005]. This system has been practically operated at an entrance hall of a public community center in Ikoma-city, Nara, Japan since Nov. 2002. as shown in Fig.1. The system answers to a user's question, e.g., information about the community center, sightseeing, transportation or today's weather and about the agent "Takemaru-kun" himself. Since this system starts, all of the triggered inputs to the system are recorded. By analyzing and statistically model learning of the input data, the system robustness has been improved, i.e. (1). Rejection of unintended speech based on likelihood measurements using Gaussian Mixture Models (GMMs). (2) Removal of unnecessary input that have short input length for impulsive noise. (3) Discrimination between adult and child users on the basis of speech recognition scores using two parallel decoders.

On the other hand, prosodic features are not used for this system. It is well known that prosody plays an important role for carrying non-verbal information, such as speaker's attitude, emotion or individuality. If we can obtain such information from a user input, the response to the question can be more intelligent.

In this paper, the authors conduct a prosody analysis to find a clue to obtain a speaker's state from his/her speech.

This paper consists of five major sections. We will first describe an overview of our "Takemaru-kun" system in section 2. Section 3 explains the user speech database. The results of prosodic analysis are described in section 4. Here, we observe the global tendency from the viewpoint of a system revision, e.g. introducing speech-noise, adult-child discrimination and system stabilization. We conclude this paper in Section 5.

2. Information guidance spoken system "Takemaru-kun"

Fig. 1.Takemaru-kun speech guidance system (right) and the Ikoma North Community Center (left).

2.1. System structure

Figure 2 shows the structure of "Takemaru-kun". Synthetic speech, CG agent animation and a web browser are used for system outputs. The speech interface is designed as a simple one-question-one-response strategy to accomplish daily guidance for users without time delays. A user speech is captured through a microphone, and a response is output by a synthesized voice generated from a Text-To-Speech program. Animated gestures and related Web pages are also presented on two displays. The architecture design is based on a blackboard model comprised of four modules: Main (speech recognition and response selection), Agent, Web browser, and Speech Synthesizer. They communicate with each other through a status server that shares the states of all modules via TCP/IP. Each module operates independently and can stop and start at anytime.

To make appropriate response choices automatically, an example question-answer (QA) database is prepared beforehand. It consists of actual questions queried to the system, which are morphologically analyzed from transcriptions of user utterances. Each question is attached to a suitable answer. After decoding a user speech to a text, the number of matched morphemes between an example question and the text is totaled for all pre-stored questions. In this procedure, N-best outputs for to complement recognition errors.

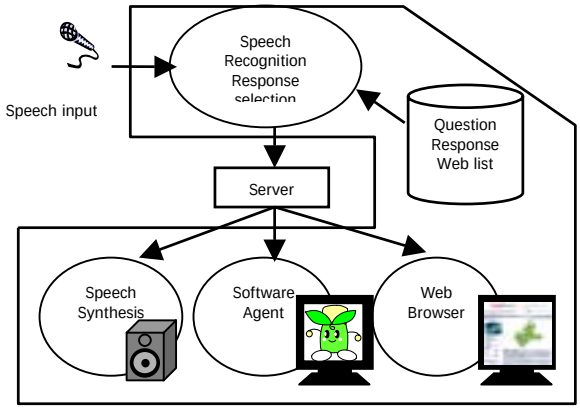

Fig. 2. System structure of "Takemaru-kun"

Then each score is determined by dividing the number of matched morpheme by words in the questions. A response candidate attached to the best-matched example is selected as a response. An advantage of this approach is that it generates a certain response concerned with a guidance task if the QA database is satisfied. The QA database is prepared that can cover a wide area that relates to topics of users' interests. In addition, after practical system operation starts, user input logs are examined and important ones are added to the QA database.

This modularity simplifies the development of each part. The main module converts input speech to texts and generates a proper response by choosing a suitable one from prepared response candidates. In addition, rejection of unnecessary inputs and estimation of user age groups are also performed in the same module. The Synthesis module synthesizes output speech according to the generated response.

The Agent module displays gestures of a CG agent characters, Takemaru-kun synchronizing with response speech. Agents can also indicate the detection of the start of an utterance to a user by just nodding. Visual information such as Web pages, maps are displayed by Web browser module. For further Web retrieval, manual operation with a mouse is also possible.

2.2. Speech recognition Architecture

A detailed procedure of the speech recognition module is illustrated in Fig.3. Two major features of recognition system are briefly described here with its practical system installation period.

2.2.1. Child and adult discrimination

Around December 2003, to make the system interface more friendly to a child and an infant user, two parallel speech recognizers were installed in the system. Each has an age group-dependent language model (LM) and an acoustic model (AM) suitable for adult or child users. Outputs are chosen based on comparisons between the two likelihood scores from each decoder [Nisimura 2004].

This service is welcomed by children, however, system instability has occurred after this revision. This instability is supposed to effect to number of total utterance, as we mention in Section 4.

2.2.2. Speech and Noise discrimination

To reject unnecessary inputs, speech verification to determine whether a system input is a intended input or not by comparing acoustic likelihoods given by GMMs. To discriminate an utterance-like but not a user intended input such as laughter, coughing, and wrongly triggered background speech, they are also used to model "laughter", "coughing" and "other noise" GMMs [Lee 2004].

In addition to the GMM-based verification, short inputs removal is also implemented in practically operated "Takemaru-kun" system in April 2003.

Furthermore, the whole system stability is also checked in this implementation. As a result, reliability of the system is established after this revision.

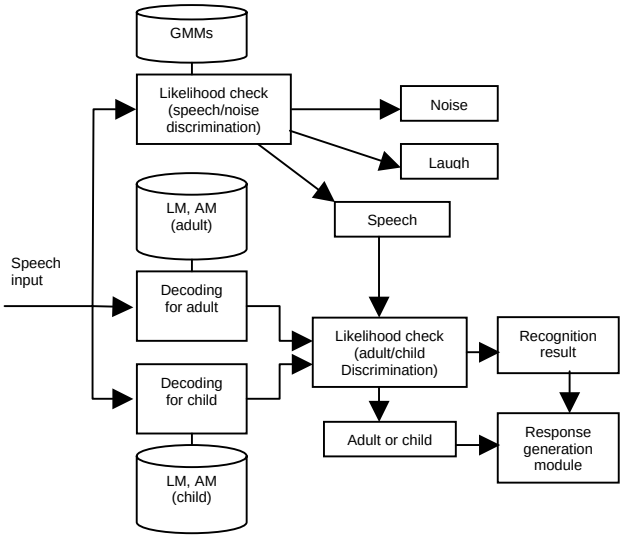

Fig.3. Architecture of a speech recognition module

3. Speech Database

All triggered system inputs are recorded and those recorded between November 2002 and October 2004 have already manually transcribed and labeled by four matured labelers. In this paper, 22 month (from Nov. 2002 to Aug. 2004) user utterances are used for analysis.

3.1. Labeling

The collected inputs are classified manually to examine how unintended inputs occur. Noise input that contains no speech part is classified into the "Noise" category. On the other hand, an utterance that is clear and can be easily transcribed by a human, and the speaker has the obvious intention to talk to the system, it is classified into the "Clean utterance" category. Fillers and disfluencies also belong to this category. Other inputs should be further classified into the "Wrongly triggered utterance" and "Non-verbal" categories. The former contains unintentionally triggered utterances (i.e., background speech) and incomprehensible speech (level underflow, level overflow and vocal sounds which are impossible to transcribe). The latter consists of "laughter" and "coughing". The utterances in the "Clean utterance" category are further classified. They're labelled by listening for its age group (Infant / Lower grade child (up to about 9 years old) / Higher grade child (about up to 15)/ Adult / Elderly Person / Uncertain) and by gender (Male /

Female / Uncertain). These are perceptual labels therefore severe boundary of neighbouring categories cannot be set, so we focus on temporal changes of each group.

3.2. Numbers of utterances

Total numbers of utterances summed up by month are illustrated in Fig. 4. The intended input during 22 months is 133,179 utterances in total. Percentage of children utterance increases through the period although absolute number of adult utterance doesn't decrease. In general, during spring and summer holiday seasons, the numbers of utterance increase. In addition, they reflect condition of the service. For example, we revised the system fundamentally about December 2003 with introducing child/adult speech discrimination. By this module installation, the system becomes more familiar especially to children because it comes to output different response to adults (polite sentence (conventional)) and children (friendly sentence), however some troubles on stability were also occurred in the revised system. The decrease of the following months can be considered that users' interest also decrease. From the next major revision in April 2004, stability has established and speech/noise discrimination module is also introduced. The increase of utterance is supposed the result of this improvement.

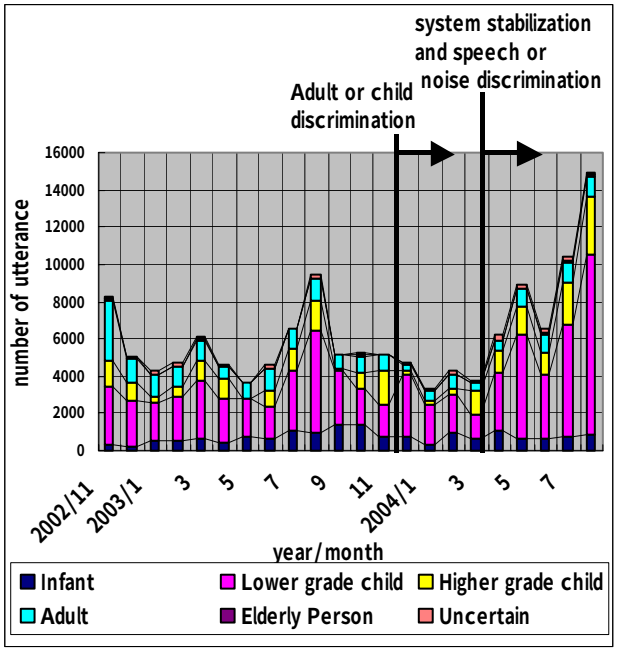

Fig. 4. Number of utterance in each month

4. Prosody analysis

To analyze long-term prosodic tendency, monthly averaged F0 and speech rate are calculated.

In this paper, as a first step to investigate cues to estimate user's state from prosodic features of input speech, we focus on the relationship between average prosody and system revision including its first appearance in November 2002. As a matter of course, many and unspecified (supposed to living in the city) persons use this system, a speaker of each utterance cannot be identified. However, it is assumed that users' average condition for the system is expressed in average prosodic features.

4.1. F0, speech rate and mora length

Figures, 5, 6 and 7 illustrate average F0 [Hz], speech

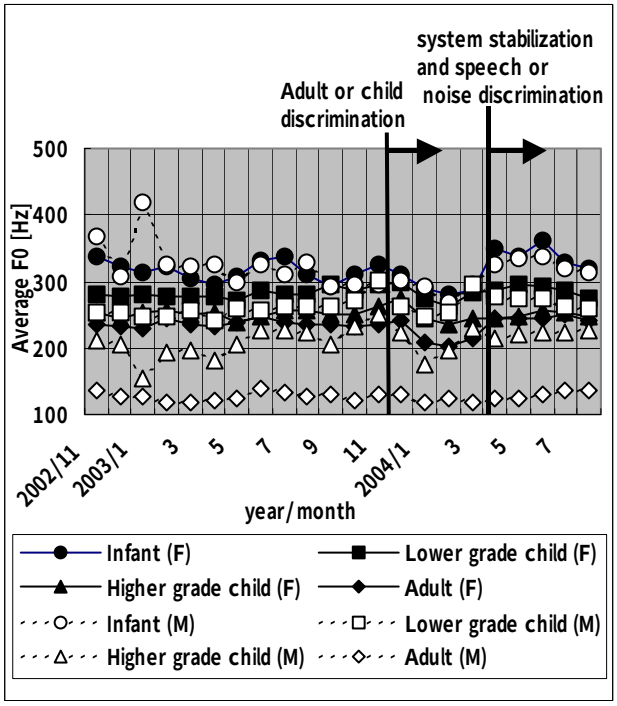

Fig. 5 Average F0

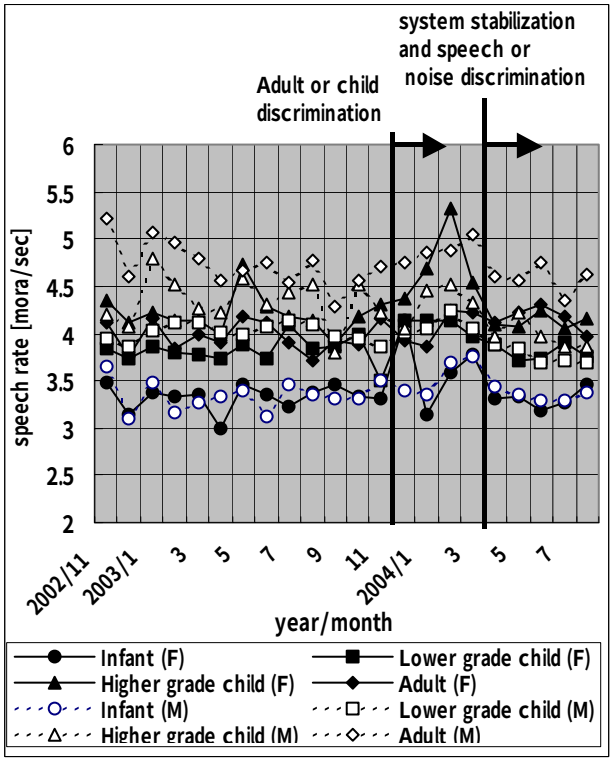

Fig. 6. Average speech rate

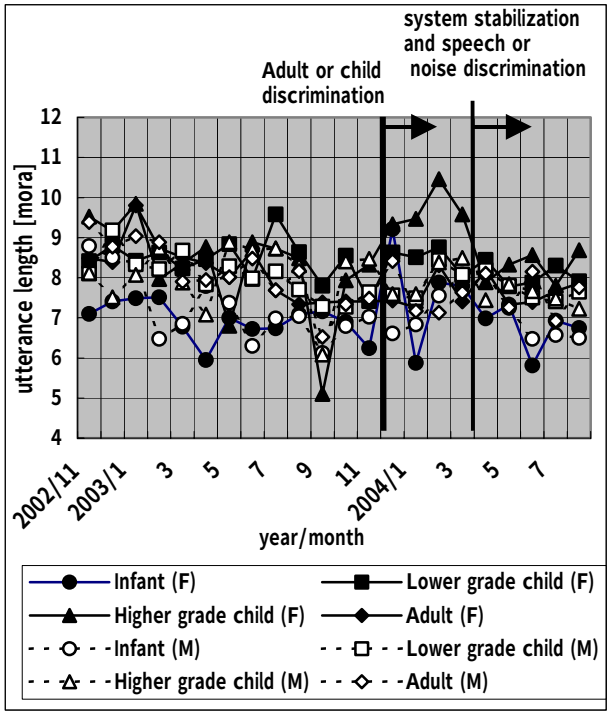

Fig.7. Average utterance length

rate [mora/sec] and utterance length [mora/utterance], respectively. Elderly person data are omitted because the number of utterance is small as shown in Fig. 4. F0 calculation is executed using STRAIGHT-Tempo method [Kawahara 1999] with 5 [msec] window shift, from 16 kHz sampling 16 bit speech samples. F0 averaging is done in log domain.

4.2. Discussion

In every figures, the child/adult discrimination and the speech/noise discrimination period are illustrated.
Same increase/decrease tendency is observed in F0 and number of utterance in Fig. 4. It is supposed that when a spoken system has troubles (from December 2003 to March 2004), a user feels some kinds of negative feeling and they're expressed as F0 falling regardless of age or gender. In addition to that, although they're not obvious characteristics but on speech rate and on utterance length, their increasing for every category are roughly observed in the same period. This

5. Conclusion

This paper conducts a long-term prosody analysis of user speech to find a clue to obtain user's attitude from a user's speech. Although this analysis is in a preliminary stage but the possibility of common prosodic tendencies in average as it is shown in Fig. 5 that is to say, user F0 decreases when the stability of the system is not satisfied.

In our future work, we will report quantitative analysis and test between speaker's attitude and prosodic features and evaluate estimating attitude from speech inputs.

6. Acknowledgment

A part of this study is supported by the e-Society project provided by MEXT (Ministry of Education, Culture, Sports, Science and Technology), Japan. The authors also greatly appreciate supports by the Ikoma City office and the Ikoma North Community Center.

7. References

R. Nisimura, et al. (2005). Operating A Public Spoken Guidance System In Real Environment. In *Proc. Interspeech2005 (EUROSPEECH2005)*, Lisbon, Portugal, pp.845-848.

A. Lee, et al (2004). Noise Robust Real World Spoken Dialogue System using GMM Based Rejection of Unintended Inputs. In *Proc. Interspeech2004 (ICSLP2004)*, Jeju, South Korea, pp.173-176.

R. Nisimura, et al. (2004). Public speech-oriented guidance system with adult and child discrimination capability. In *Proc. IEEE-ICASSP*,Montreal, Canada pp.433-436.

H. Kawahara, et al (1999), Fixed point Analysis of Frequency to Instantaneous Frequency Mapping for Accurate Estimation of F0 and Periodicity, In *Proc. EUROSPEECH*, Budapest, Hungary, pp.2781-2784.