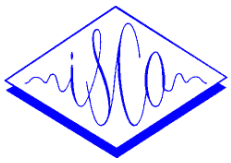

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

PROSODIC PATTERN OF UTTERANCE UNITS IN JAPANESE SPOKEN DIALOGS

Kazuyuki Takagi and Shuichi Itahashi

Institute of Information Sciences and Electronics, University of Tsukuba
Tsukuba, Ibaraki, 305 Japan

ABSTRACT

This paper reports a study to classify spontaneously spoken utterances by their prosodic features only. First, input speech was segmented into utterance units, each of which is a speech interval segmented by pauses or distinctive utterances such as interjections and fillers. Secondly, the cluster analysis was conducted using seven prosodic parameters of an utterance unit, i.e., speech rate, F_0 onset, final, values and the positions of the maximum and the minimum F_0. On the 10 spoken dialogs uttered by 9 speakers selected from ASJ Continuous Speech Corpus, the utterance units were classified into 5 distinctive clusters: "Initial & Middle", "Final", "Slow", "Devoiced", and "Others". Each cluster was characterized by syntactic and prosodic features of utterance units contained in the cluster. Results of this paper can be used for a preprocessing of spontaneous speech understanding systems.

1. INTRODUCTION

Spontaneously spoken language has distinctive phenomena, which make machine processing of spontaneous speech a very difficult task, e.g., interjection words, filled pauses, false starts, interruptions, repeated phrases, elliptic sentences, inversions of word order, and so on. It is well known that prosodic cues have significant effect on the naturalness and intelligibility of spoken language. Especially for continuous speech processing systems that have to deal with syntactically loosely constrained speech input, prosodic information should be beneficial to remedy syntactic and semantic ambiguity of the speech.

There have been several studies on the utilization of prosodic information for spoken language understanding and dialog management (e.g., [1], [2], [3]). An interesting approach is a data-driven approach, such as prosodic phrase segmentation method proposed in [4], where no prior knowledge of prosodic structure is given.

The aim of this paper is to get some meaningful relationship between the syntax and semantics of spoken dialogs and the prosodic features of the utterance. Prosody of real speech data is dynamically affected by contexts, interaction between the speakers, emotion, etc., as well as the accent patterns and phrase structure, thus it is not well known whether model-based approaches are effective to spontaneous speech data. We do not assume any prosodic model for that reason.

Various guide task dialogs (e.g., tourist guides, restaurant guides, geographical guide) of ASJ Continuous Speech Corpus were used for the analysis, where two participants were asked to make a conversation freely with knowledge about the task given prior to the recording [5]. Utterance units (i.e., speech intervals segmented by pauses, and utterances such as interjections, fillers) in the dialogs were classified by their prosodic parameters related to speech rate and F_0 contour. The k-means clustering analysis gave five clusters of utterance units, each of which was characterized by syntactic and prosodic features of utterance units contained in the cluster.

2. SPOKEN DIALOG MATERIAL

Various simulated spoken dialogs of guide tasks were used for analysis, in which roles of two speakers were fixed to the client and the agent. Spoken dialog materials were selected from ASJ Continuous Speech Corpus for Research Vol.7 published as a CD-ROM disc from Japan Information Processing Development Center, which consists of 32 guide task dialogs and 5 free conversations [5]. Simulated dialogs were collected at 8 institutions, recorded with DAT in a quiet room or through telephone line. Scenarios, situations, specific goals and outlines of the dialog, and necessary information, such as maps and time tables, was given to the participants before the recording, then the participants were asked to make a conversation freely to achieve the task. The corpus provides detailed kanji-kana transcriptions of the dialogs. Speech data is neither segmented nor labelled.

We picked up 10 spoken dialogs of guide tasks for analyses whose topics are restaurant guide, tourist guide, inquiry on passport application, geographical guide, and so on. The dialogs were spoken by 9 different speakers (8

male speakers and 1 female speaker) in face-to-face situation. Speech data were sampled at 16 kHz and quantized into 16 bits. Two speaker's voices were mixed in a single recording channel.

3. ANALYSIS METHOD

3.1 Utterance unit

Syntactic boundaries are not clearly observed in spontaneous speech data, because of the disfluencies such as false starts, repetitions, interjection words and filled pauses, and because an utterance of one speaker is often interrupted by the other speaker. Therefore a unit of utterance should be defined as a basis for an analysis of spontaneous speech.

We segmented speech data into utterance units (UU), i.e., speech intervals segmented by pauses. Each interjection word, filled pause, and false start were segmented as a single utterance unit even if they were not bounded by pauses. An utterance unit does not necessarily correspond to syntactic segments in transcribed texts because pauses in spontaneous speech may occur within a word or a phrase, and because there may be no pause at a syntactic boundary.

Segmentation of speech data into utterance units was conducted manually through observation of speech waveform and spectrum. When two speakers' utterances overlapped each other, they were separated into two utterance units. The time measuring unit for start and end points of an utterance unit was 10 ms. Then every utterance unit was labelled with start and end time of the utterance unit, and transcription text.

There were 2965 utterance units whose duration ranged from 20 ms. to 4.4 seconds with the average duration 78.3 ms. One utterance unit generally consisted of several Japanese syntactic phrases, or "bunsetsu". The average number of syntactic phrases per one utterance unit was 1.55 including interjections, fillers, response words, and false starts, and 1.93 excluding them. There were 9.22 moras† in a single utterance unit on the average (min.=1, max.=43), and 92.7% of utterance units were shorter than 21 moras consisting of 1 to 5 syntactic phrases (or "bunsetsu").

† mora: Japanese syllabic unit, which basically corresponds to one Japanese syllabary character, or "kana". A mora is composed of five types of phoneme concatenation: (1) V, (2) CV, (3) CYV, (4) YV, or (5) a special phoneme (syllabic nasal /N/, geminate consonant /Q/, long vowel /-/) (C: consonants /p, t, k, b, d, g, m, n, s, z, r, c, h/, Y: semivowels /y, w/, V: vowels /a, i, u, e, o/).

3.2 Prosodic parameters

We selected the following seven acoustic features illustrated in Figure 1, which seemed to be meaningful for the classification of spontaneously spoken utterances.

- speech rate: number of moras uttered per second
- fundamental frequency (F_0) contour
  - F_0 onset value
  - F_0 final value
  - F_0 maximum value
  - F_0 minimum value
  - Position of F_0 maximum value
  - Position of F_0 minimum value

Fundamental frequency contours were calculated by the AMDF (average magnitude differential function) method. Estimation errors were corrected manually.

The first five parameters in the above list were normalized by the mean value and the standard deviation of each speaker. The duration of every utterance unit was normalized to unity, so that the last two parameter values were bounded to [0,1]. For devoiced utterance units (i.e., utterance units where no pitch was detected) all of the six F_0 parameters were set to zero.

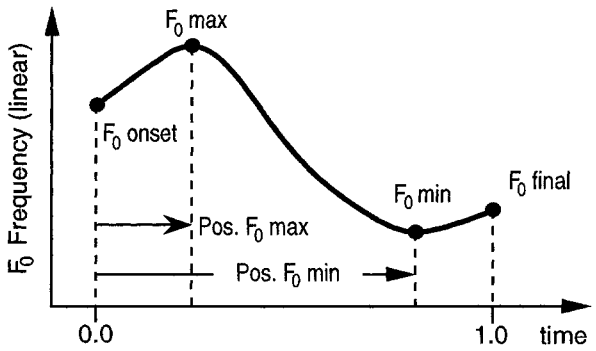

Fig. 1. Prosodic parameters selected for a classification of utterance units, which relate to the fundamental frequency contour. Values of F0 onset, F0 final, F0 max, F0 min were normalized by the speaker's mean F0 value and the standard deviation. The duration of an utterance unit was normalized to unity. Values of Pos.F0max and Pos.F0min were therefore bounded to [0, 1]. Note that for devoiced utterance units all the F0 parameters were set on zero.

3.3 Clustering

An experiment to classify utterance units was conducted by using the seven prosodic parameters described above. Some utterance units were simultaneously spoken with the other speaker's utterance. Prosodic parameters on the overlapped portion of these utterances are fairly difficult to estimate. In the ten dialogs selected from the ASJ Corpus, where the speech signals of the two participants were not separated, 471 of 2965 utterance units (15.9%) were simultaneously spoken. The overlapped utterance units

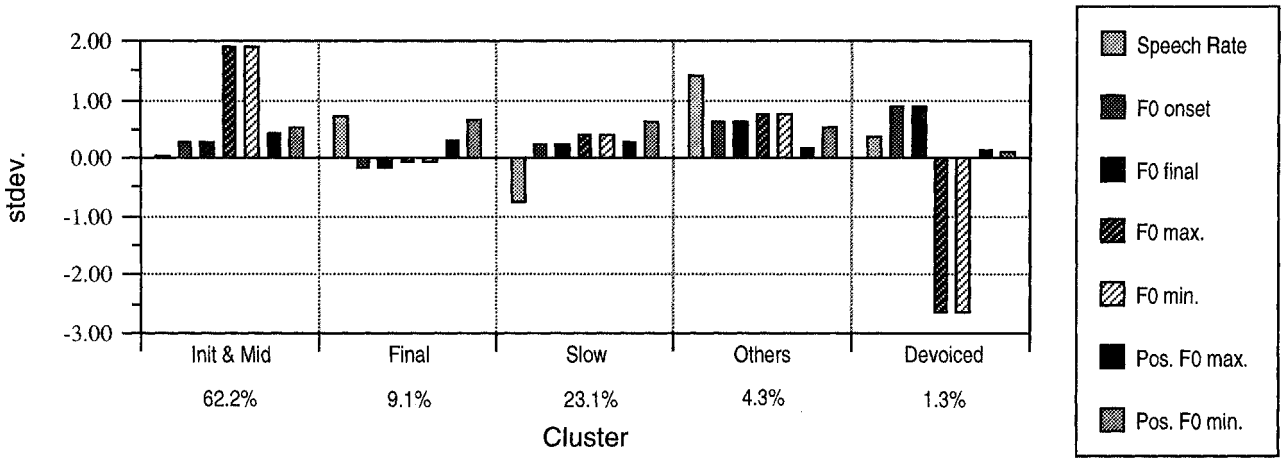

Fig. 2. Mean value of prosodic parameters of five distinctive clusters of utterance units. Parameter values are depicted as the deviation from the mean values of all utterance units. Each cluster was named according to some sentential and prosodic characteristics of utterance units contained in the cluster. **"Init & Mid"**: sentence-initial and mid-sentence utterances, **"Final"**: sentence-final uterances, **"Slow"**: utterance units spoken slowly, **"Devoiced"**: devoiced utterances, **"Others"**: the remainder. Note that these cluster names do NOT mean that all utterances of the kind indicated by the name were gathered in the clusters. Percentages are the population ratios of the clusters.

were excluded from the analysis.

The k-means clustering technique was used to classify utterance units using their seven prosodic parameters described in the previous section. The distance measure between two utterance units was the Euclid distance, and between group distance was measured by the centroid distance.

4. Results

We performed several sessions to cluster 2494 utterance units, by changing clustering conditions, such as initial seeds and distance thresholds. Every clustering session produced four to seven clusters, but the population ratio and the prosodic parameter values of each clusters were stable.

Figure 2 illustrates the representative result, where the utterance units were classified into five distinctive clusters. Figure 2 shows the mean value of seven prosodic parameter of each cluster. We will name each cluster according to sentential feature (e.g., sentence-initial, sentence-final) and prosodic characteristic (e.g., slowly uttered, devoiced) characteristics of utterance units contained in the cluster. They were "Initial & Middle", "Final", "Slow", "Devoiced", and "Others", which has 62.2%, 9.1%, 23.1%, 4.3%, and 1.3% of all utterance units respectively. The following sections will describe characteristics of each cluster.

4.1 Cluster "Initial & Middle"

This is the largest cluster, which has more than 60% of utterance units. The value of F0 maximum and minimum are higher than the other clusters, as plotted in Figure 2. Speech rate is about the mean value of all utterance units.

Relatively long (i.e. 15 to 25 moras) sentence-initial and mid-sentence utterances that are preceded by a interjection or a filler, and the second interjection words of consecutive interjections occupied 35.4% of this cluster. Utterances of this cluster contains more than two thirds of content words and keywords occurred in the present dialog material.

4.2 Cluster "Final"

The third largest group of utterance units (9.1%) has more sentence-final utterances than the other clusters. This feature was reflected to the relatively fast speech rate and low F_0 frequency values. Almost all sentence-final euphemistic expressions of inquiries and requests were classified into this cluster.

4.3 Cluster "Slow"

The most outstanding characteristic of this cluster is that the average speech rate of this cluster alone is negative. This cluster has 32% of interjections uttered inside a sentence as fillers. In addition, this cluster contains slowly spoken clue words and phrases such as, ". . . iQpaNtoshitewa . . . " (in general), ". . . sorede . . ." (well then).

4.4 Cluster "Devoiced"

The mean of F_0 maximum and minimum of this cluster showed large negative value. All utterances gathered in this cluster were spoken with no pitch. Among them were particles of 1 to 3 moras, short interjection words, and chiming utterances that were uttered both ambiguously and quickly.

In a real speech recognition system, utterance units that fall into this cluster could be excluded from the parsing procedures, because these utterances bear no significant information relevant to the dialog contents.

4.5 Cluster "Others"

Both speech rate and F_0 values of this cluster are relatively high. Interjectory utterances inserted between two phrases occupy 35.9% of this group. Among them were interjections spoken lightly (e.g. "... ma ..."(well), "... eto ..."(say)), phrases that were spoken quickly (e.g. ". . . teyuunowa . . ."(that is to say)). There are no other distinctive syntactical and phonological feature of utterance units contained in this cluster.

5. CONCLUSION

We have conducted an analysis on prosody of spontaneously spoken dialogs. The analysis was done by bottom-up approach, which means that neither prior knowledge of the prosody nor a model was employed.

Utterance units in spoken dialogs were clustered into five clusters by seven prosodic parameters relating to speech rate and fundamental frequency contour of the utterance unit. These clusters given in the following list are distinctive in terms of the syntactic and prosodic features of the utterance units belonged to the cluster.

- "Initial & Middle" (62%): sentence-initial and mid-sentence utterances; 2/3 of content words and keywords
- "Final" (9.1%): sentence-final utterances
- "Slow" (23.1%): fillers, slowly spoken clue words
- "Devoiced"(1.3%): devoiced utterances
- "Others" (4.3%)

Both two large clusters, "Initial & Middle" and "Slow", contain utterances that do not fall into the categories indicated by the cluster name. Analysis involving the contexts of utterance units (e.g., difference of F_0 maximum value, or speech rate change) will be required for further examination of the present results. Prosodic features of the simultaneously spoken utterances were excluded in the present analysis, because the two speaker's utterances were not separately recorded in the corpus. Analyses on prosody employed in the interactions between the dialog participants are left for future works.

ACKNOWLEDGEMENTS

This work was supported in part by Grant-in-Aid for Scientific Research of No. 05241107 from Ministry of Education Science and Culture.

REFERENCES

[1] A. Komatsu, E. Oohira and A. Ichikawa, "Conversational speech understanding based on sentence structure inference using prosodics, and word spotting," Trans. IEICE, Vol.J71-D, No.7, pp.1218-1228, July 1998 (in Japanese)

[2] R. Kompe, A. Kieβling, T. Kuhn, M.Mast, H.Niemann, E. Nöth, K. Ott, and A. Batliner, "Prosody takes over: A prosodically guided dialog system," Proc. of EUROSPEECH'93, pp.2003-2006, Sept. 1993

[3] S. Nakajima and H. Tsukada, "Utterance pattern characteristics in task-oriented dialogues," Technical Report of IEICE, SP93-102, Nov. 1993

[4] H. Shimodaira and M. Nakai, "Prosodic phrase segmentation by pitch pattern clustering," Proc. of ICASSP'94, Vol.II, pp.185-188, Apr. 1994

[5] S. Hayamizu, S. Itahashi, T. Kobayashi and T. Takezawa, "Design and creation of speech and text corpora of dialogue," Trans. IEICE, Vol.E76-D, No.1, Jan. 1993