5V－1

Prosodic Features in User Response to a Speech Dialogue System *

L.J. Mayfield, K. Itou, S. Hayamizu, and K. Tanaka
Electrotechnical Laboratory

1 Introduction

This paper discusses prosodic features characteristic of English in user responses to a speech dialogue system and examines how this information can be extracted from the speechwave.

Feeling that the ability to recognize sub-lexical information is important for speech recognition systems, we looked at the way users of English and Japanese speech dialogue systems modified their utterances when asked for an exact repetition. By comparing data from the two systems, we tried to identify those English prosodic features which have the greatest influence on meaning. We outline here the experiment performed and present the results of acoustic analysis.

2 Dialogue System

The English speech dialogue system is an expansion of the previously-existing Japanese and English continuous speech recognition systems [1-2] at ETL, incorporating elements of the Japanese-language dialogue system [1]. The dialogue manager of the English system has the same structure as the Japanese system, modified to accept English input[3-4]. Phoneme models were trained using the TIMIT database [2].

Interacting with the ETL speech dialogue system[3], users asked questions about the Tokyo train system. When their utterance was unprocessable by the system, users were asked to repeat themselves. It was these lexically identical utterances that we studied, searching for evidence of prosodic influence on meaning.

3 Experiment

The experiment was designed to force the user to spontaneously repeat exactly the same sentence one or more times. By comparing these utterances, we hoped to be able to identify what tools native speakers use to emphasize meaning.

A total of 67 dialogues from 10 users were recorded. Among these were 49 repetitions of exactly the same sentence. A brief examination of the ways in which users responded to rephrasement requests is in itself revealing. There were 13 instances of blatant ungrammaticality, seven of which were abbreviations of the utterance immediately preceding to the core word/phrase. (How long, transfer, etc.) Stumbling and insertion of filler words (um, er, ah, yeah, so) occurred as shown in Table 1.

Speaker#	1	2	3	4	5	6	7	8	9	10
Stumbling	3	1	3	2	3	-	2	1	1	1
Filler	1	-	-	-	5	7	-	2	1	-

表 1: Occurrence of stumbling and filler words

The Japanese data used for comparison was taken in a similar experiment in November of 1991. A total of 33 repeated phrases from the 62 dialogues were examined.

4 Analysis of User Response

Prosodic features of lexically identical utterances obtained using the English system were subjectively evaluated by native listeners. Three dialogues, chosen as representative of each of the three main types of prosodic modification, were then analyzed.

Sentence pairs for the analysis were those agreed on by three independent native speakers to have as their primary distinguishing element stress. Speakers listened to neighboring repeat sentence pairs in isolation, without hearing the full dialogue.

Dialogue 1

a. What will the fare be?
b. What will the **fare** be?
fare=/feyer/ ; p_1=/f/, p_2=/ey/, p_3=/er/

Dialogue 2

a. How long is the trip?
b. How **long** is the trip?
long=/laang/ ; p_1=/l/, p_2=/aa/, p_3=/ng/

Dialogue 3

a. How long does it take?
b. How **long** does it take?
long=/laang/ ; p_1=/l/, p_2=/aa/, p_3=/ng/

Dialogue 4

a. How much is it?
b. How **much** is it?

*音声対話システムへのユーザ発話における韻律的特徴
ローラ・メイフィールド, 伊藤克亘, 速水 悟, 田中和世
電子技術総合研究所

Pair	F_0phr	F_0wor	Lp_1	Lp_2	Lp_3
1-a	4.13	4.13	190	80	280
1-b	−11.22	−11.22	200	160	290
2-a	15.73	11.32	110	100	70
2-b	25.45	29.68	420	30	90
3-a	32.44	−7.47	220	30	30
3-b	42.63	−18.35	270	50	30
4-a	−26.99	—	60	110	150
4-b	−35.77	—	150	150	310

表 2: Phonetic characteristics of stressed vs. nonstressed words (F_0phr = F_0 variances of phrases in Hz, F_0wor = F_0 variances of words, Lp_1-Lp_3 = Length of phonemes in msec)

much=/mahch/ ; p_1=/m/, p_2=/ah/, p_3=/ch/

Length of phonemes in stressed word of each pair were compared, as were the pitch variances of both individual stressed words and the phrases containing those words. Pitch extraction and segmentation were executed using a process based on voice fundamental wave filtering developed at ETL[5]. Length of phonemes is automatically estimated from alignment data using the recognition system. These phonetic characteristics of stressed vs. nonstressed words are summarized in Table 2.

Examination of the 49 instances of repetition revealed general trends so distinct that results of fine quantitative analysis may seem inconsequential by comparison. Comparing word length and pause length showed slight trends towards longer total sentence length and longer pauses between words, but compression of all but the stressed syllable; which actual consonant phones of the stressed syllable were lengthened varied from person to person.

General trends in the examination were in order of frequency:

- Increasing stress in repetitions
- Shift of word stress
- Rising intonation in repetitions
- Clear enunciation of each word

The repetitive sentences obtained using the Japanese sytem were subjectively judged by a native speaker of Japanese. The listener categorized the repetitions into those containing stress, those that were repeated slowly or emphasizing each syllable, and those that were the same. Of the 21 repetitions, 10 contained additional stress, seven were slowed down, one was repeated syllable by syllable, and 10 underwent no change.

The instances of stress, however, were for the most part undetectable by the English speaker. Japanese listeners do *hear* stress in fluent speech, but it is used solely for emphasis. Rudimentary examination of a representative dialogue including both additional stress and slowing of pronounciation, as judged qualitatively, revealed a slight increase in pitch drop between the unstressed and stressed instances, as well as significant lengthening of both repetitions.

5 Future Direction

The issues and results discussed in this paper represent only the beginnings. By designing speech recognition systems which exploit the unique aspects of different languages we are not precluding teaching machines what all languages have in common. Just as there are two components of acquired language, that which we know from birth and that which we absorb through exposure, there are two ways in which we can develop electronic systems *which are not mutually exclusive.*

Still unresolved is the question of corresponding phenomena in Japanese. Also, these results, although possibly attributable to personal differences, suggest that prosodic elements in English may not be common to all dialects. Differences in the perception of stress by Japanese and English speakers indicate that we may be fooled in our conception of what our languages share.

Acknowledgement

The authors wish to express their sincere thanks to Dr. Hiroshi Ohmura, whose programs were used in acoustic analysis, and all the members of the Speech Processing Section at ETL.

References

[1] K. Itou, S. Hayamizu, K. Tanaka and H. Tanaka. "System Design, Data Collection and Evaluation of a Speech Dialogue System", IEICE Trans. Inf. & Syst., Vol.E76-D, No.1, pp. 121-127, 1993.

[2] J. Kuo and S. Hayamizu. "An English Speech Recognizer for the Task of Railway Transportation Guidance", ETL Technical Report, TR-92-29, 1992.

[3] L.J. Mayfield, K. Itou, S. Hayamizu and K. Tanaka. "Examination of User Response to a Speech Dialogue System and Development of the System Used", ETL Technical Report, TR-93-8, 1993.

[4] L.J. Mayfield, K. Itou, S. Hayamizu and K. Tanaka. "Language Specific Influences in User Presentation of Response to Speech Dialogue Systems", IPSJ 93-SLP-1-6, pp. 27-33, 1993-07.

[5] H. Ohmura. "Pitch Extraction and Segmentation by Voice Fundamental Wave Filtering", Technical Report of IEICE, SP92-77,DSP92-62, 1992-10.