ACOUSTICAL LETTER

Dynamic aspects of aizuchi and its influence on the naturalness of dialogues

Hiroki Mori*

Graduate School of Engineering, Utsunomiya University, 7–1–2, Yoto, Utsunomiya, 321–8585 Japan

(*Received 30 July 2012, Accepted for publication 24 September 2012*)

Keywords: Backchannel, Spontaneous speech, Paralinguistic information, Perception

PACS number: 43.71.An [doi:10.1250/ast.34.147]

1. Introduction

To realize a *humane* spoken dialogue system, generating natural spoken feedbacks is crucial. Unlike speech synthesis for read-aloud applications, dialogue speech synthesis is expected to handle the interactive, communicative, and real-time aspects of spoken dialogue. Among them, it is important to take into account paralinguistic aspects of synthesized speech (how to speak) as well as linguistic aspects (what to speak).

Aizuchi (backchannel) is a phenomenon that is frequently observed in conversations by Japanese speakers. It plays important roles in monitoring interlocutor's state and maintaining smooth turn-taking. Therefore, generating appropriate aizuchi by spoken dialogue systems is important to make human-computer interactions natural and smooth. To this aim, a number of studies were involved in generating aizuchi by spoken dialogue systems, most of which are related to timing determination for aizuchi responses [1–5]. On the other hand, some studies investigated the relationship between paralinguistic information conveyed by natural aizuchis and their acoustic characteristics [6,7]. However, there have been few studies on controlling paralingusitic information of machines' synthetic aizuchi responses by manipulating their speech parameters. One of the difficulties in expressive and "rich" aizuchi synthesis is that there is no wide-accepted framework to evaluate synthesized aizuchis.

This letter describes a preliminary study on aizuchi synthesis for spoken dialogue systems. Specifically, it focuses on human perception of the naturalness of consecutive aizuchis. The goal is to understand perceptual effects of aizuchi's acoustic dynamics on overall naturalness of spoken dialogue.

In this letter, the unit of experiment is a discourse segment which is called 'turn' here. We hypothesized the following two perceptural effects:

Hypothesis 1 *Replacing aizuchi waveforms in a turn by some identical aizuchi waveforms gives less natural impression.*

Hypothesis 2 *Replacing aizuchi waveforms in a turn by randomly selected aizuchi waveforms (i.e. irrespective of local context) gives less natural impression.*

2. Corpus and aizuchi characteristics

To test these hypotheses, a set of stimuli was created by using a spontaneous dialogue speech corpus, the Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies (UUDB) [8], as a material. The UUDB includes natural, spontaneous and expressive dialogues between seven pairs of college students (12 females, 2 males). Because both participants in each pair were close friends, and because the "4-frame cartoon sorting task" used in the corpus collection was very amusing and strongly motivated the participants, the corpus covers a wide variety of expressive utterances as well as aizuchis.

As for the dialogues included in the UUDB, aizuchis occured in relatively early timings. The mean switching pause duration was 361 ms for normal transitions, whereas it was 55 ms for aizuchis [9]. In addition, 43% of aizuchis started earlier than the endings of interlocutor's preceding utterances. These aizuchi timings were earlier than those of the telephone shopping task [3]. This implies that a considerable number of aizuchis were produced in a somewhat automatic way, without deep understanding of the interlocutor's utterances, in the case of the UUDB.

Figure 1 shows the distributions of acoustic parameters (f_0 range, peak intensity, spectral tilt and duration) of utterances "uN" in the UUDB. "uN" is one of Japanese aizuchi expressions and almost all (98%) aizuchis are "uN" in the UUDB. From Fig. 1, it is understood that acoustic parameters of aizuchis distributed over a broad range even though their linguistic contents were identical.

3. Method

The unit of speech stimulus is called 'turn' in this paper. A turn is a short interaction of two parties, which consists

- exactly one whole utterance of speaker A, and
- equal to or more than three "uN" utterances of speaker B.

A turn starts with speaker A's utterance, and ends with either speaker A's utterance or speaker B's final "uN." Although not all of "uN" utterances are aizuchi [10], we did not distinguish them and will henceforce regard all of them as aizuchi, because the acoustic parameter distributions of aizuchi and non-aizuchi "uN"s were hardly distinguishable in the case of the UUDB*.

In this way, 60 turns of 12 speakers were extracted from the UUDB, where 9 of them were discarded because of

*e-mail: hiroki@speech-lab.org

*The only acoustic parameter shown in Fig. 1 with a significant mean difference between aizuchi/non-aizuchi groups was peak intensity, where the difference was not more than 1.1 dB.

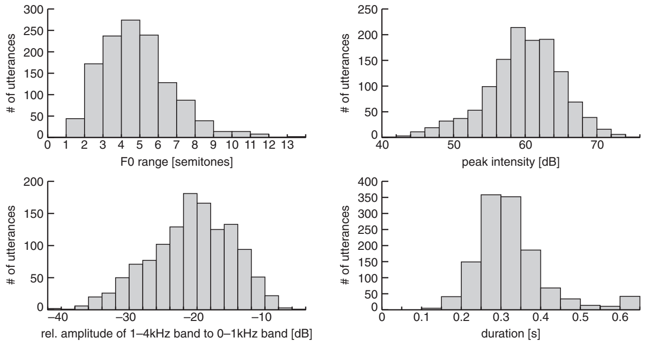

Fig. 1 Distribution of acoustic parameters of utterances "uN."

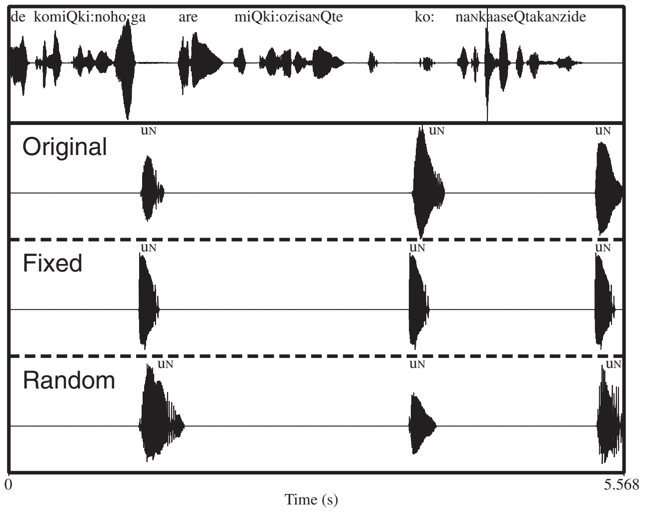

Fig. 2 Manipulation of aizuchi waveform.

including too characteristic aizuchi or big noise. Then, for each speaker, "uN" utterances were clustered using K-means ($K = 8$) with 3 speech parameters (peak f_0, duration, peak intensity). The clustering was necessary to ensure that original and replacing aizuchis (see the descriptions of Original, Fixed, and Random below) constituting an utterance have similar acoustic parameters. Without the clustering process, extremely prominent aizuchis could replace normal aizuchis, and vice versa. Based on the clustering result, turns that contain "uN"s belonging to different clusters were discarded. In other words, the set of stimulus was constructed from such turns in which all "uN"s were similar to each other. Finally, eight female speaker's 34 turns that meet the above conditions were selected as the set of stimulus.

Next, aizuchi waveforms contained in each turn were manipulated according to the following three patterns (Fig. 2).

Original Speaker B's aizuchi waveforms were left as they were.

Fixed Speaker B's all aizuchi waveforms were replaced by an identical aizuchi waveform.

Random Speaker B's aizuchi waveforms were replaced by randomly selected aizuchi waveforms.

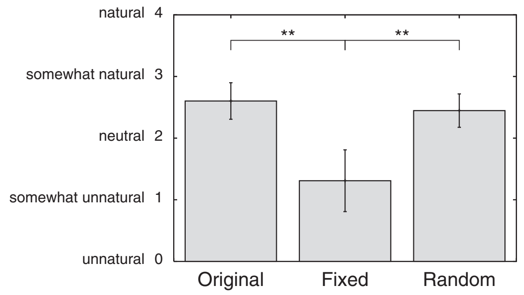

Fig. 3 Mean scores for the manipulation conditions (**: $p < 0.01$).

For the Fixed pattern, one aizuchi waveform was randomly selected from the cluster to which the original aizuchis belonged. Similarly for the Random pattern, multiple aizuchi waveforms were randomly selected from the same cluster, just as many as Speaker B's aizuchi utterances, without getting duplicates. The replacing waveforms were aligned so as to keep the start points unchanged. For each of the 34 turns, manipulation of the three patterns described above was made, which yields 102 two-channel speech waveforms for the stimuli.

Ten college students (5 females, 5 males) participated in the perceptual experiment. None of them had specific knowledge of speech science. They were asked to wear headphones (MDR-Z600, SONY) to listen the stimuli. Speaker A's utterance was played through one channel, and Speaker B's aizuchis were played through another channel. Each of 102 stimuli was presented in a random order. After finishing each playback, they were asked to evaluate the overall naturalness of the dialogue with 5-level opinion score (0: unnatural, 1: somewhat unnatural, 2: neutral, 3: somewhat natural, 4: natural), with paying attention to aizuchis. They were instructed to evaluate each stimulus as a whole, rather than judging each aizuchi individually. This was due to the presumption that concentrating to individual aizuchis leads to neglect of context in the judgment.

4. Result and discussion

Mean opinion scores over all subjects for each aizuchi manipulation pattern are shown in Fig. 3. The naturalness for the Fixed pattern was noticeably lowered than the original. It is also understood that the naturalness for the Random pattern seems not so lowered. A three-way ANOVA (gender of subject (2) × turn (34) × manipulation pattern (3)) revealed a significant main effect of turn ($F(4.59, 36.7) = 4.39$, $p < 0.01^{\dagger}$) and manipulation pattern ($F(1.18, 9.43) = 31.0$, $p < 0.01^{\dagger}$). The Bonferroni post hoc analysis for manipulation pattern showed that there was a significant difference between Original and Fixed ($p < 0.01$) and between Fixed and Random ($p < 0.01$), but the difference between Original and Random was not significant ($p > 0.05$). The main effect of gender was not significant ($F(1, 8) = 4.05$, $p > 0.05$) and the interaction between gender and manipulation pattern was not significant ($F(1.17, 9.43) = 0.93$, $p > 0.05^{\dagger}$).

The naturalness degradation for Fixed compared to Original is a result that supports the Hypothesis 1. This

[†] Degrees of freedom were adjusted using the Greenhouse-Geisser correction.

implies that the subjects felt strange about hearing exactly same aizuchis.

On the other hand, the same degree of naturalness for Random compared to Original is a result that rejects the Hypothesis 2. One possible interpretation for this would be the slight change of aizuchis within a turn. It was common both for the Fixed and Random pattern that all aizuchis were acoustically similar, and that the choice of aizuchis was irrespective of local context. The only difference was the change from aizuchi to aizuchi, however slight.

From the viewpoint of aizuchi synthesis, the result that Random aizuchis were as natural as Original ones implies the importance of changing, rather than controlling, the acoustical properties of aizuchi. This is analogous to the fluctuation of glottal pulse, a dominant factor for the naturalness of synthetic speech. Ifukube *et al.* [11] showed that repetition of identical pitch waveforms made synthetic voice unnatural, and that random shuffling of pitch waveforms within a certain time range did not cause naturalness degradation.

5. Conclusion

As a preliminary investigation for synthesizing paralinguistic features of aizuchis for spoken dialogue systems, this letter described a perceptual experiment to assess the effect of manipulating acoustical properties of aizuchi. The order of overall naturalness for three manipulation patterns was: Original = Random > Fixed. This implies that mere changing aizuchi waveforms randomly gives pretty natural impression.

This finding serves as a ground for designing and evaluating aizuchi synthesis functionality, which is expected for future spoken dialogue systems to provide.

Acknowledgments

The author would like to thank Mr. Yuuki Saito for his great assistance.

References

[1] N. Ward, "In Japanese a low pitch means 'back-channel feedback please'," *IPSJ SIG Notes*, 1996-SLP-11, 7–12 (1996).

[2] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa and Y. Den, "An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs," *Lang. Speech*, **41**, 295–321 (1998).

[3] Y. Okato, K. Kato, M. Yamamoto and S. Itahashi, "Giving 'aizuchi' using prosodic information," *Trans. Inf. Process. Soc. Jpn.*, **40**, 469–478 (1999) (in Japanese).

[4] J. Hirasawa, M. Nakano, T. Kawabata and K. Aikawa, "Effects of system barge-in responses on user impressions," *Proc. Eurospeech '99*, pp. 1391–1394 (1999).

[5] S. Fujie, K. Fukushima and T. Kobayashi, "Back-channel feedback generation using linguistic and nonlinguistic information and its application to spoken dialogue system," *Proc. Interspeech 2005*, pp. 889–892 (2005).

[6] J. Umeno, H. Kashioka and N. Campbell, "The aizuchi un's prosody expressing yes/no and para-linguistic information," *1st JST/CREST Int. Workshop Expressive Speech Processing* (2003) (in Japanese).

[7] M. Enomoto and Y. Ishimoto, "An investigation on acoustic features of 'un' in Japanese for automatic classification of answer, acknowledgement and aizuchi," *IPSJ SIG Notes*, 2009-SLP-77, 1–6 (2009) (in Japanese).

[8] H. Mori, T. Satake, M. Nakamura and H. Kasuya, "Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics," *Speech Commun.*, **53**, 36–50 (2011).

[9] H. Mori, "An analysis of switching pause duration as a paralinguistic feature in expressive dialogues," *Acoust. Sci. & Tech.*, **30**, 376–378 (2009).

[10] The Japanese Discourse Research Initiative, "Japanese dialogue corpus of multi-level annotation," *Proc. 1st SIGdial*, pp. 1–8 (2000).

[11] T. Ifukube, M. Hashiba and J. Matsushima, "A role of 'waveform fluctuation' on the naturality of vowels," *J. Acoust. Soc. Jpn. (J)*, **47**, 903–910 (1991) (in Japanese).