Acoustic correlates of monosyllabic utterances of Japanese in different speaking styles

Akemi Iida[†‡], Parham Mokhtari[†] and Nick Campbell*

† JST/CREST Expressive Speech Processing

‡ Keio Research Institute at SFC

*ATR Human Information Science Research Laboratories

E-mail: akeiida@sfc.keio.ac.jp, {parham, nick}@atr.co.jp

ABSTRACT

This paper describes our research into the relation between paralinguistic information and acoustic features of monosyllabic, typically backchannel and filler utterances in Japanese. For this study, we extracted 141 examples of "hai", "un", and "ah" from recordings of spontaneous conversational speech from one Japanese female. These utterances can function differently according to the speaker's intention and tone of voice. We show in this paper that prosodic differences can affect the perceived meaning of each of these three utterance types. We first subcategorized them into three types each (i.e., affirmative, reflective, and turn-holding) based on the perceived intention of the speaker and then performed an acoustic analysis for each occurrence, and compared their acoustic characteristics, i.e., F0, duration, energy, and a normalized amplitude quotient (NAQ) that measures breathiness. The results showed that affirmative utterances were typically produced with a higher fundamental frequency and with more energy than reflective and turn-holding ones. The latter had higher values of duration and NAQ, enabling all three types to be distinguished, with respect to functional intention, on the basis of their prosody alone.

1. INTRODUCTION

Speech signals not only linguistic content, but also other types of inter-personal and communicatively-relevant information, such as the identity, intention, attitude, and mood of the speaker, as well as urgency of the message. Such *paralinguistic information* cannot easily be found in "lab speech", where no listener is immediately present, but is common in naturally-occurring spontaneous speech. We have been investigating the relation between paralinguistic information and acoustic features using several hours of recordings from one Japanese female, a volunteer who wore a small head-mounted studio-quality microphone and recorded her day-to-day speech to a small minidisk recorder over a period of several months [1]. In this paper, we describe our investigation into the prosodic characteristics of short monosyllabic utterances from this data, typically fillers or backchannel utterances, that were extracted from her telephone conversations.

Backchannels are short utterances produced by a participant in a conversation while the other is talking and are sometimes known as 'listeners responses'. Typical English examples include "yeah", "uh-huh", "hm", and "aw" while typical Japanese examples include "hai", "ah", "un", and "so". They are particularly frequent in Japanese speech, where the listener takes a very active role in the interaction. According to Maynard [2], they function as 'continuers', 'display of understanding the content', 'support toward the speaker's judgment', 'agreement,' 'strong emotional response', 'correction', 'request for information', and 'adding information'. When produced with a specific intonation and voice quality, backchannels can function to show 'disapproval' or 'objection', as can be seen in "aww" in English and "so?" in Japanese. In addition, backchannels also function as signals for turn holding and turn taking in a discourse.

We selected three kinds of monosyllabic backchannel utterance, "hai", "un", and "ah" which appeared either alone, or in sentence-initial position. Some of these utterances comprised an entire sentence, whereas others simply formed the initial part of a longer utterance. There were 141 of these words in our data. We did not attempt to subcategorize them using all of the functions listed above, but rather into three basic functional types which are used to express the listener's intentions and attentional state: *affirmative*, *reflective*, and *turn-holding*. We performed acoustic analyses for each occurrence and compared their acoustic-prosodic characteristics (F0, duration, and energy). Further we calculated the normalized Amplitude Quotient (NAQ) to quantify phonation-type. This has been found to correlate with the perception of breathiness [3]. In the following section, we describe our procedures and the results of our acoustic analyses.

Table 1. Counts of tokens by feature.

	Affirmative	Reflective	Turn-holding	Sum
"ah	13	20	23	56
"un	22	21	23	66
"hai	12	6	1	19
Sum	47	47	47	141

ISBN 1-876346-48-5 © 2003 UAB

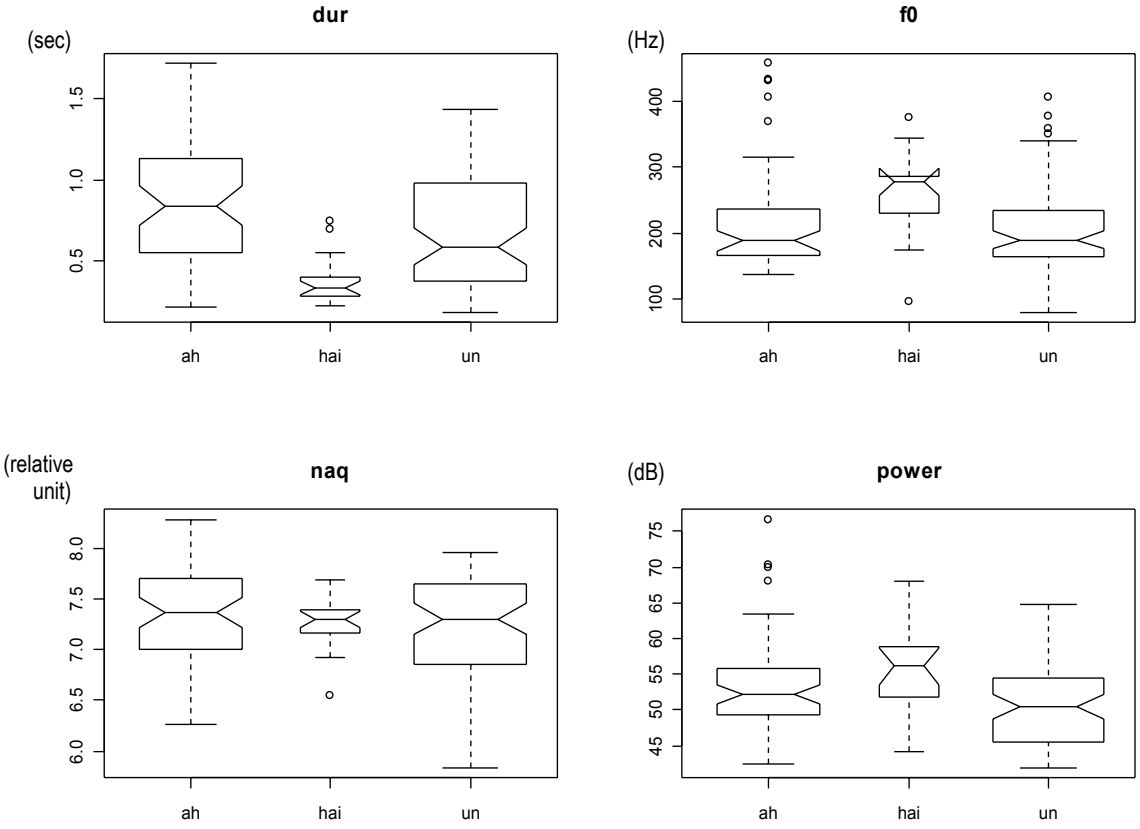

Figure 1. Prosodic characteristics or the three words.

2. ACOUSTIC ANALYSIS

2.1 Speech Data

Table 1 lists the distribution of the 141 stimuli in each of the three utterances in each of the three function types. "Hai" did not appear in the data as often as "ah" and "un" but since it is lexically marked for affirmation, we consider it as a reference for the prosody of an affirmative type of speech act.

2.2 Acoustic Analysis

The following acoustic features were extracted using custom-developed software in order to determine prosodic and voice-quality parameters. The fundamental frequency of voicing (henceforth "F0") was measured using a subharmonic-summation method based on Hermes [4]. The root-mean-square (rms) energy was measured (in dB) from the speech waveform at each analysis frame. The duration of each utterance was measured on the basis of manually-determined labels, giving timing information relating the speech waveform with the orthographic transcription. For each utterance a glottal amplitude quotient was measured at automatically-located centers of reliability in the speech stream, where the results of formant estimation and therefore inverse-filtering are most reliable [3, 5].

AQ is defined as the ratio of the peak-to-peak amplitude of the glottal waveform and the largest negative-peak of the glottal waveform derivative. We here use the normalized-AQ (NAQ) which reduces the F0-related variations of AQ, and which therefore better reflects changes in glottal voice quality.

2.3 Acoustic Characteristics

We analyzed four prosodic or speaking-style features: F0, Duration, RMS energy, and breathiness. We then performed means comparisons tests between particular pairs using Student's t test. In this section, the results of the means comparisons tests are reported for each feature. The means and standard deviations are summarized in Table 2, which gives a breakdown according to each of the three utterances.

3. RESULTS

Figure 1 shows the prosodic features of each word. We can see that "hai" has a higher F0 and a greater RMS energy than "ah" or "un", which can also be used as fillers (indicating the speaker's reflective state) or as turn-holders, preventing a turn exchange in the discourse. The main feature of this figure is that it shows very little difference in the prosodic characteristics of "ah" and "un", and reveals that they show a wider range of values in each dimension compared with "hai".

ISBN 1-876346-48-5 © 2003 UAB

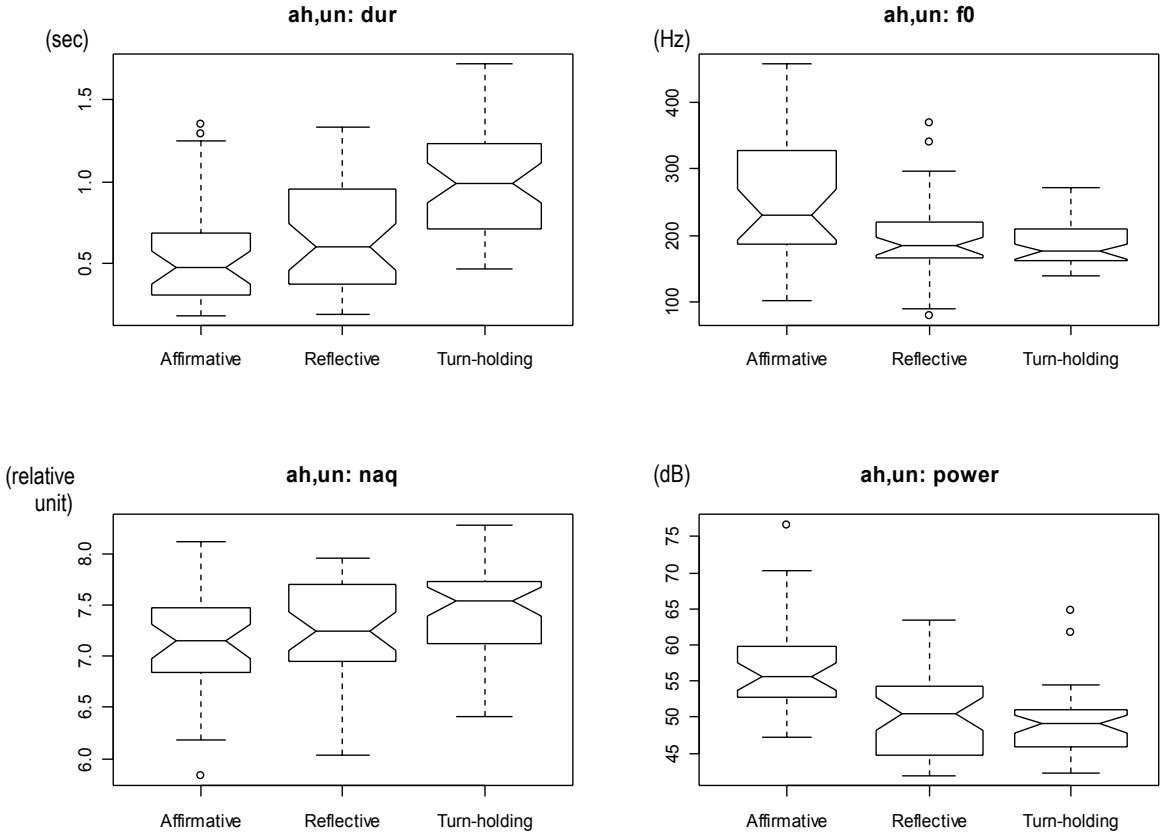

Figure 2. Prosodic characteristics of the three functions.

There is a significant difference with respect to duration between "ah" and "un" (t (df, 113) = 2.84, $p<0.01$) but there is no significant difference with respect to power (at $p<0.01$), NAQ, or F0, between "ah" and "un". It is interesting that these words can be used in functionally different contexts without misunderstandings on the part of the listener. We hypothesize therefore that they must be distinguished by separation in the prosodic space in conversation, to reveal the intentions of the speaker without obvious ambiguity. To test this hypothesis, we plot the same data, factored this time by functional type, in Figure 2. The data for "hai" has been excluded from this plot to reduce any statistical influence from this word which is relatively unambiguous in terms of functional use and marked in its prosodic characteristics.

We can see from figure 2 that there is a clear separation between "Affirmation" and "Turn-holding" in every case. The difference in NAQ, for example, is significant at $p<0.001$ (t (df, 68) = 3.45. For F0, t = 4.15, for power, t = 6.03, and for duration, t = 5.28.

4. DISCUSSION

The word "hai" means "yes", and like "yes" in English, it can also be used to indicate listening, thinking, or just function as a filler. The other words, "ah" and "un" are not so clearly lexically defined. They carry paralinguistic information and signal the speaker's attitude and intentions.

When "un" or "ah" are used to signal affirmation, the pitch is higher, the power stronger, the duration shorter, and the voice less breathy. When used to signal turn-holding, the pitch is lower, the power weaker, the duration longer, and the voice is more breathy. However, the distinction between "reflection" and "turn-holding" is not so clear. It is likely that duration is used as the main cue to this distinction.

It is interesting to speculate that breathiness in the voice may be used to signal speaker intention. We notice that breathiness is highest for the case of turn-holding. This is a social situation that places the speaker in a potentially aggressive position, and we posit that the aggression may be mitigated somewhat by the increase in softness of the voice. This would be in accordance with the results presented in [1]. When signaling affirmation, on the other hand, the voice is least breathy. Table 3 gives a breakdown of the prosodic characteristics for each word in each functional situation. We can see that "ah" and "un" behave similarly in each case.

ISBN 1-876346-48-5 © 2003 UAB

Table 2. Means and standard deviations of prosodic features in the three utterances.

	F0 (Hz)	Duration (ms)	RMS Energy (db)	NAQ
"ah"	216.52 ± 78.2	84.77 ± 36.97	53.29 ± 6.88	7.35 ± 0.45
"un"	204.27 ± 66.30	66.32 ± 34.37	50.63 ± 5.72	7.23 ± 0.51
"hai"	261.43 ± 67.88	37.30 ± 14.90	55.67 ± 5.94	7.27 ± 0.27

Table 3. Means and standard deviations of the prosodic features for each word and function type.

		Affirmative	Reflective	Turn-holding
F0 (Hz)	All three	254.16 ± 91.96	210.77 ± 66.50	185.58 ± 32.98
	"ah"	283.43 ± 115.33	206.45 ± 63.84	187.45 ± 29.93
	"un"	239.84 ± 83.20	190.91 ± 58.60	182.45 ± 36.42
	"hai"	248.72 ± 77.34	294.68 ± 35.69	214.40 ± ---
Duration (ms)	All three	51.90 ± 30.95	61.83 ± 32.96	95.47 ± 31.73
	"ah"	78.62 ± 38.34	71.83 ± 35.29	99.49 ± 33.73
	"un"	44.25 ± 22.73	60.91 ± 30.63	92.36 ± 30.34
	"hai"	36.98 ± 14.51	31.70 ± 4.14	74.80 ± ---
RMS energy (dB)	All three	56.88 ± 6.42	51.01 ± 5.87	49.20 ± 4.25
	"ah"	60.06 ± 8.64	52.57 ± 5.35	50.09 ± 3.85
	"un"	55.22 ± 4.52	48.32 ± 5.18	48.34 ± 4.61
	"hai"	56.49 ± 5.88	55.22 ± 6.24	48.60 ± ---
NAQ	All three	7.16 ± 0.43	7.24 ± 0.50	7.45 ± 0.41
	"ah"	7.30 ± 0.43	7.21 ± 0.45	7.51 ± 0.43
	"un"	7.00 ± 0.46	7.29 ± 0.60	7.40 ± 0.40
	"hai"	7.31 ± 0.29	7.20 ± 0.26	7.32 ± ---

5. CONCLUSION

This paper described the correlation between paralinguistic information and acoustic features of three Japanese monosyllabic words which are potentially ambiguous in their usage. We selected 141 segments of "hai", "un", and "ah" from the recordings of spontaneous daily conversational speech from one Japanese female. These words function differently according to the speaker's intention and the context of the discourse. We subcategorized them into three types based on the perceived intention of the speaker in her role as a listener: i.e., as affirmative, reflective, or turn-holding markers.

We performed an acoustic analysis for each occurrence and compared the F0, duration, energy, and breathiness. The results of our analyses showed that affirmative utterances were produced with higher frequency and with more energy than reflective and turn-holding ones. The latter in turn can be distinguished by duration, power, and voice quality. It is thus clear that prosody plays an important role in portraying speaker intentions and that an understanding of speech based on the lexical content alone would be missing important paralinguistic information.

Future work will include analysis of a more extended list of words that are dependant on prosody for their interpretation, and a clarification of the relationship between breathiness in the voice and the other prosodic features.

ACKNOWLEDGMENTS

We are grateful for financial assistance from the Japan Science and Technology Agency via the CREST (Core Research for Evolutional Science and Technology) scheme for Advanced Media Technology. We thank Dr. Hartmut Pfitzinger and Dr. Carlos Ishi of JST CREST for their valuable advice and discussion.

REFERENCES

[1] N. Campbell and P. Mokhtari, "Voice quality: the 4th prosodic dimension," In Proceedings of the 15th ICPhS.

[2] Mayhard, S. K, "*Japanese conversation: Self-contextualization through structure and interactional management*. Norwood, NJ:Ablex, 1989.

[3] P. Mokhtari and N. Campbell, "Automatic measurement of pressed/breathy phonation at acoustic centres of reliability in continuous speech," IEICE Transactions on Information and Systems, vol. E86-D, no. 3, pp. 574-582, 2003.

[4] Hermes, D. "Measurement of pitch by subharmonic summation," J. Acoust. Soc. Am., vol.83, no.1, pp. 257-264, 1988.

[5] P. Alku, T. Bäckström and E. Vilkman, "Normalized ampltude quotient for parameterization of the glottal flow," J. Acoust. Soc. Am., vol. 112, no. 2, pp. 701-710, 2002.