Acoustic-prosodic and paralinguistic analyses of "uun" and "unun"

Carlos T. Ishi [1], *Hiroaki Hatano*[1], *Miyako Kiso*[1]

[1] Intelligent Robotics and Communication Labs., ATR, Kyoto, Japan

carlos@atr.jp, hatano.hiroaki@atr.jp, miyakokiso@atr.jp

Abstract

The speaking style of an interjection contains discriminative features on its expressed intention, attitude or emotion. In the present work, we analyzed acoustic-prosodic features and the paralinguistic functions of two variations of the interjection "un", a lengthened pattern "uun" and a repeated pattern "unun", which are often found in Japanese conversational speech. Analysis results indicate that there are differences in the paralinguistic function expressed by "uun" and "unun", as well as different trends on F0 contour types according to the conveyed paralinguistic information.

Index Terms: interjections, acoustic-prosodic features, paralinguistic information, spontaneous conversational speech.

1. Introduction

In spontaneous dialogue speech, repeated "un" utterances ("unun…") or lengthened "un" utterances ("uu…n") are often used as variations of the (short/single) backchannel "un". These two patterns may cause different impressions to the interlocutor depending on the situations they are used. Therefore, in order to achieve a smooth communication between humans and machines, these two patterns should be discriminated.

In our past works, we have focused on interjections appearing in Japanese natural conversational speech, and analyzed the relationship between speaking style and the paralinguistic information (intentions, attitudes and emotions) conveyed by the interjections [1-4]. So far, several monosyllabic interjections (such as "un", "ee", "oo", "ha", "he", "ya") have been analyzed. However, utterances where interjections are repeated in sequence have not been focused.

So far, several works have been conducted regarding the interjection "un" in Japanese (including its variations in speaking styles) [1-8]. However, there are only few studies devoting attention to the functions of "unun" comparing to "un" in discourse. For example, the relations between acoustic features and functions of interjection "un" have been investigated using speech data extracted from TV drama [5]. It is mentioned that "uun" indicates "unknown information is being stored", "embarrassment" or "hesitation". In [6], it has been described that "unun" functions as a marker that the interlocutor's last utterance evoked the speaker's attention for the current conversation topic. In [7], the distribution of backchannel expressions appearing in natural conversations has been analyzed. It has been concluded that "un" is commonly used by both information providers and followers, while "unun" is only used by information followers.

Regarding F0 pattern analysis, the relationship between speaker's attitudes and F0 patterns of "un" have been analyzed in [8]. They found that mean F0 is higher when attitudes of "activation, acceptance, confidence" are expressed, while durational change structures are related with "affirmation/negation" expression. In our past works [1-4], we have shown that the tone type (rising, falling, flat tones) is useful for discriminating between functional speech acts (such as positive reactions, asking for repetition, and thinking), while voice quality features are useful for discriminating emotion expression (such as surprise, admiration, and disgust). However, the speaking styles and paralinguistic functions of "unun" utterances have not been clarified so far.

Regarding backchannel analysis in other languages, it is reported that occurrence rates and environmental location of Japanese backchannels differ from Mandarin and English [9]. It is reported that, in general, appropriate forms and timings of backchannels ("reactive tokens" in their term) show the interest for the speakers and encourage the speaker to keep talking. However, a relatively heavy usage of backchannels in Japanese provides emotional support for speakers [9]. In [10], a large variety of non-lexical tokens, mainly backchannels, uttered in English conversations was examined and the relationship between prosodic features and its meanings was reported. It was stated that duration lengthening (equivalent to "uun" in Japanese) means "amount of thought", while syllabification (equivalent to "unun") means "lack of desire to talk".

Another motivation for the present work is the generation of head motion of robots synchronized with speech utterances [11,12]. Our analysis of head motion during speech utterances has revealed that backchannels are often accompanied by nods, and a sequence of repeated backchannels, such as in "unun" are usually accompanied by multiple nods, approximately one nod per "un" repetition [11]. On the other hand, it was also found that in "uun" utterances where the speaker is thinking, a head tilting is often accompanied. Thus, in the synchronization of head motion with speech, the discrimination of these two types becomes important.

Another issue is that "unun" and "uun" utterances may have similar spectral features, so that acoustic features commonly used in speech recognition, such as MFCC (Mel-Frequency Cepstral Coefficients), would not be enough for distinguish these two patterns. Thus, acoustic analyses are conducted on spectral features and F0 contours.

2. Analysis of "unun" and "uun" utterances

2.1. Speech data

The dataset for analysis was extracted from the ATR multi-modal natural dialogue database. The database contains 65 dialogue sessions of 10 ~ 15 minutes, including 11 male speakers and 14 female speakers with ages from 10s to 60s. The conversation topics are free, including past experiences, future plans, topics about a common known person.

Firstly, a text search was conducted on the transcriptions in the database for collecting "unun" (including two or more repetitions of "un") and "uun" (including two or more repetitions of the vowel lengthening) utterances. In Japanese a vowel lengthening symbol is used for long vowels. As the transcriptions in the database were conducted by multiple

10.21437/SpeechProsody.2014-8

annotators, the transcription criteria might not be unified. Thus, we asked three native speakers (research assistants) to check the consistency of the transcriptions. As a result, 6% of the utterances were corrected, resulting in a total of 342 "unun"-type utterances, and 926 "uun"-type utterances.

The left panel in Fig. 1 shows the distribution of the utterance duration in "unun" and "uun" utterances. Due to a big overlap in the distributions, segmental duration cannot be used to distinguish these two utterance types. In the right panel of Fig. 1, the distribution of the number of "un" repetitions are shown for the "unun" utterances. It can be noted that the most frequent was the pattern with two repetitions (45%), followed by the pattern with three repetitions, i.e. "ununun" (32%), four repetitions (16%) and five or more repetitions (7%).

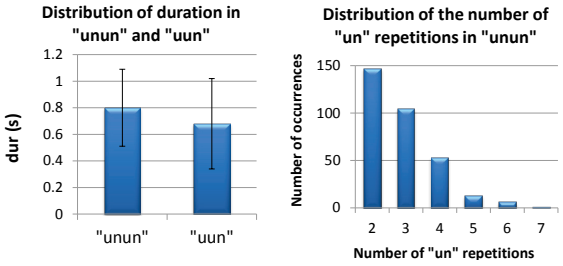

Figure 1: *Distribution of duration in "unun" and "uun" utterances (left panel). Distribution of the number of "un" repetitions in "unun" utterances (right panel).*

2.2. F0 contour type analysis

For each utterance, tone labels were annotated by the first author (which is experienced in prosody annotation), based on F0 curve displays and auditory impression. As tone categories, the following labels were used for monosyllabic utterances:

- "Fa": falling tones
- "Ft": flat tones
- "Rs": rising tones
- "Rt": pitch reset
- "FtFa": pitch remains flat and ends with a falling tone.
- "?": F0 cannot be observed due to vocal fry or low power.

For utterances with two or more syllables, such as in "unun", the following labels were used:

- "FaFa": sequence of falling tones
- "Fa_Fa": sequence of falling tones separated by a short pause

Fig. 2 shows the distributions of different tone types in "unun" and "uun" utterances.

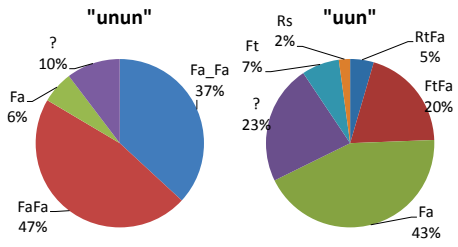

Figure 2: *Distribution of tone types in "unun" and "uun" utterances.*

Flat tones were observed in 7% and rising tones in 2% of the "uun" utterances, while such tone types were not observed in "unun" utterances.

Figs. 3 and 4 show examples of F0 contours and MFCC-smoothed spectrograms of "unun" and "uun" utterances found in our dataset.

Among the "unun" utterances, short pauses smaller than 100 ms between successive "un" syllables were found in 37% of the utterances. In this type, the F0 breaks between successive "un" syllables, being observed as a sequence of falling tones, as the example shown in Fig. 3a.

Most of the "unun" utterances appeared with the pattern without pauses between the "un" syllables (47%). Among them, we observed patterns where the nasal "n" portion is identifiable in the spectrogram and patterns where the (smoothed) spectral features do not change clearly, as the example shown in Fig. 3b. This last pattern, in particular, has little difference to the "uun" utterance spectrogram, so that the only use of spectral features (such as MFCC) would not be enough for their discrimination.

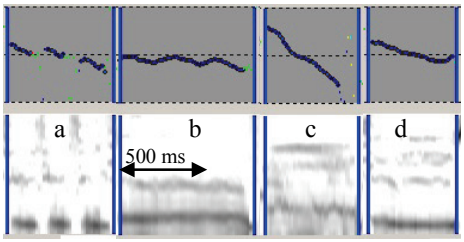

Figure 3: *Examples of F0 contours and MFCC smoothed spectrograms for "unun". F0 is in log scale from 110Hz ~ 440Hz (center dashed line at 220 Hz); spectrogram is in mel-scale 8kHz band. a) Fa_Fa_Fa ; b) FaFaFa ; c) FaFa? (pitch reset in the second "un" is unclear) ; d) Fa (F0 up-down motion is unclear)*

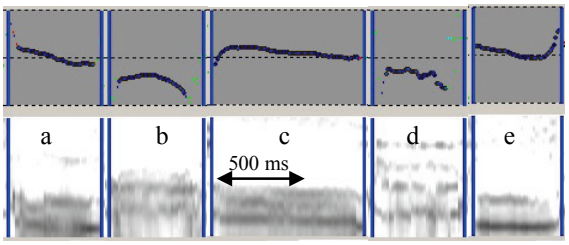

Figure 4: *Examples of F0 contours and MFCC smoothed spectrograms for "uun". a) Fa; b) FtFa; c) RtFa; d) FtFa with laughing; e) Fa+final rising.*

The phonetic transcriptions of "unun" utterances are usually represented as /uNuN/ (where /N/ is the syllabic nasal). However, in practice, the vocal tract shape does not change much its shape, and only pitch changes upward and downward. This change in pitch is thought to be used in the perceptual distinction of "uun" and "unun" by native speakers.

Another often observed pattern was the one shown in Fig. 3c, where the pitch reset (upward F0 motion) between successive "un" syllables is not clear. Nonetheless, visually one can still observe that the gradient of the F0 contour changes between the "un" syllables.

However, utterances where the F0 resets were almost non-identifiable were also observed in 6% of the "unun" utterances, as the example shown in Fig. 3d. In such cases, the tone label "Fa" was attributed. Although no clear F0 changes can be observed, spectral changes in the /N/ portion can be observed (note that the second formant is broken in Fig. 3d). This means that "unun" utterances are not necessarily accompanied by strong upward-downward F0 changes.

On the other hand, falling tones were observed in most of the "uun" utterances (43% for Fa, and 20% for FtFa). Examples of these patterns are shown in Fig. 4a ~ c. Fig 4d shows an example of "uun" utterance accompanied by laughing, where F0 considerably fluctuates upward and downward.

2.3. Paralinguistic information analysis

Three native speakers annotated the paralinguistic information conveyed by "uun" and "unun" utterances. The paralinguistic items were attributed according to the list below, prepared based on past research on paralinguistic information annotation of interjections. The original terms in Japanese are shown in brackets.

- backchannel ("aiduchi"): "I'm listening."
- affirmation ("koutei"): "Yes, that's right."
- agreement ("dooi"): "Yes, I agree."
- denial ("hitei"): "No, that's wrong"; "No, I disagree."
- negative reaction: dissatisfaction, blame, suspicion ("hiteiteki: fuman, hinan, utagai"): "I'm not satisfied"; "I'm suspicious about"; "I can't accept immediately."
- disgust ("ken-o"): "That's disgusting."
- understanding ("rikai"): "I see", "Yes, I understand."
- admiration ("kanshin"): "I'm admired"; "I'm impressed."
- embarrassment, hesitation ("tomadoi", "chuucho", "konwaku"): "I'm embarrassed/hesitated (on how to react to your utterance)."
- thinking ("kangaechuu"): "I'm preparing my next utterance."
- sympathy, compassion, pity ("kyoukan, doujou, zannen"): "I feel the same"; "It's a pity".

The annotations were conducted by taking contextual information into account, by listening to utterance intervals of both dialogue partner voices, including five seconds before and five seconds after the target interjection part. Annotators were also allowed to include a new item, if they find the items in the list do not fit to the paralinguistic information conveyed. As a result, new labels were included:

- self-affirmation ("jiko-koutei"): affirmation-like interjection right after and directed to the speaker's own utterance.
- modest affirmation ("shoukyokuteki koutei"): affirmation-like interjection, but the speaker does not express it a straightforward manner.

The inter-annotator agreement rates (in terms of kappa values) were 0.59, 0.65 and 0.80 for each pair of annotators.

Fig. 5 shows the distributions of the paralinguistic information items for "unun" and "uun" utterances. A paralinguistic item was attributed to an utterance if two or more annotators agreed. Utterances where the number of utterances for a specific paralinguistic item was smaller than 10, or where agreement was not achieved among the annotators are included in the "others" category. Agreement was achieved in more than 90% of the "unun" utterances, and in more than 75% of the "uun" utterances.

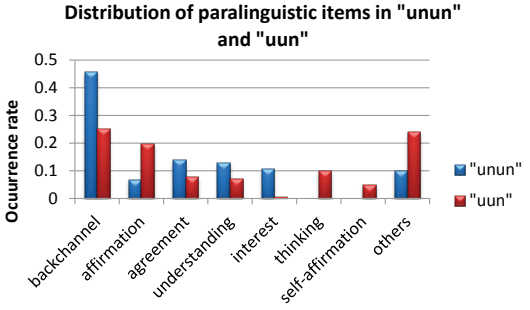

Figure 5: *Distributions of the paralinguistic information items in "unun" and "uun" utterances.*

It can be observed from Fig. 5 that both "unun" and "uun" utterances are used to express "backchannel", "affirmation", "agreement" and "understanding". However, the paralinguistic items "thinking" and "self-affirmation" were observed only in "uun" utterances, while the item "interest" was observed mainly in "unun" utterances.

Analyses were then conducted on the relationship between the tone types and the conveyed paralinguistic information.

Fig. 6 shows the distributions of the paralinguistic information items in "uun" utterances, for different tone types. The occurrence rates are normalized by the total number of utterances in each tone type, which are shown within brackets.

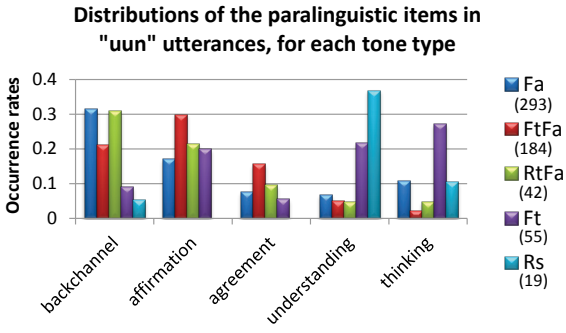

Figure 6: *Distributions of the paralinguistic information items in "uun" utterances, for different tone categories. The occurrence rates are normalized by the total number of utterances in each tone category.*

It can be observed from Fig. 6 that falling tones ("Fa" + "FtFa" + "RtFa") appear with high occurrence rates in backchannel, affirmation and agreement, flat tones ("Ft") are predominant in thinking, and rising tones ("Rs") are predominant in understanding. This is in agreement with the trends reported in past works for monosyllabic "un" utterances.

Regarding the polysyllabic "unun" utterances, no clear differences could be found between tone type and paralinguistic information, since almost all "unun" utterances have a sequence of falling tones as was shown in Section 2.2. Instead, we observed that the number of "un" repetitions could cause different impressions in the dialogue flow.

Fig. 7 shows the distributions of the paralinguistic information items in "ununn" utterances, for different "un" repetition numbers (i.e., "2" for "ununn", "3" for "unununn", and so on). The occurrence rates are normalized by the total number of utterances in each "un" repetition number category, which are indicated within brackets.

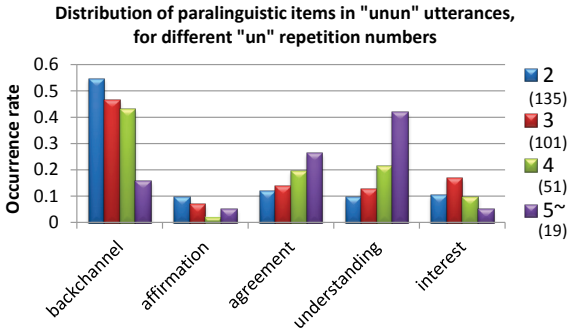

Figure 7: *Distributions of the paralinguistic information items in "ununn" utterances, for different "un" repetition numbers (i.e.,"2" for "ununn", "3" for "unununn", and so on). The occurrence rates are normalized by the total number of utterances in each "un" repetition number category.*

Among the "ununn" utterances, it can be observed from Fig. 7 that utterances with two repetitions ("2") occur with high rates in "backchannels" (more than 50%). The utterances with more than five repetitions ("5~") can be observed with predominant occurrence rates when expressing "understanding" or "agreement". When expressing interest, number of "un" repetitions around three ("3") is found to be predominant.

Finally, from the comments from the annotators, in general, "ununn" gives impression of actively pull out the interlocutor utterances by expressing interest, while "uun" gives impression of sharing the feelings of the interlocutor.

3. Discussion: Issues on the discrimination of "uun" and "ununn"

For the utterances where pauses are present between successive "un" syllables or where clear spectral changes can be observed in the /N/ portion, spectral features could be used for identification of "ununn" utterances. However, for other types, F0 patterns would be useful for discrimination. In this section the problems found in the discrimination of "ununn" and "uun" utterances based on F0 patterns are discussed.

The upward-downward F0 motions would be one strong cue for identifying "ununn" utterances. When F0 does not change, or when it moves upward and downward only once, the utterance is likely to be perceived as "uun".

However, it was shown in Fig. 4d that F0 can also move upward and downward in "uun" utterances accompanied by laughing. It would be difficult to discriminate such utterances from "ununn" utterances by only using F0 range information. Further, discrimination of "ununn" utterances accompanied by laughing would also be more difficult. Thus the dynamic features of F0 contours should be modeled for discriminating these features.

F0 rising in the end portion of the utterances was observed in 10% of the utterances, in 7 speakers (6 female and 1 male). Fig. 4e showed an example of "uun" utterance where the F0 rises in the end portion. This F0 rising was also observed (with less frequency) in the end portion of "ununn" utterances (as in Fig. 3d), and is thought to be unconsciously produced when the vocal folds stop vibrating. As this F0 rising is not well perceived, it should be removed for tone analysis.

The "?" label was annotated in 23% of the "uun" utterances and in 10% of the "ununn" utterances. Most of them were either due to low power in bad recording conditions or due to presence of vocal fry, so that F0 contours could not be obtained.

Regarding the recording conditions, in part of the database headset microphones are available, while in the other part only directional microphones on the table are available. For the table microphones, the distance to the mouth is 30 ~ 40 cm on average, so that both background air conditioner noise and the dialogue partner interference sound are strongly observed.

The problem is that "un" utterances have lower power even within the utterances of the same speaker, and their backchannel functions in dialogue make them highly probable of being overlapped with the interlocutor's utterances. Thus, the SNRs tend to become very low in "un" utterances if the microphone is not positioned close to the speaker's mouth, affecting both speech recognition and F0 extraction.

The other problem is the presence of vocal fry (or creaky phonation), where the vocal fold vibrations become irregular, so that the measured F0 would not correspond to perceived pitch contours. Vocal fry was observed mainly in 4 male subjects, where robust F0 contours could not be obtained. In such cases, the presence of short pauses between "un" utterances become more important for identification of "ununn" utterances, rather than the F0 information.

4. Conclusions

We conducted analyses on acoustic-prosodic features and paralinguistic functions of "ununn" and "uun" utterances, which are repeated and lengthened variations of the interjection "un", commonly appearing in conversational speech.

Analysis results indicated that both "ununn" and "uun" appear in the expression of backchannels, affirmation, agreement and understanding, while "ununn" appears more frequently when expressing interest, but does not appear for expressing thinking or self-affirmation, in contrast with "uun". Further, in "ununn" utterances, the number of repetitions of "un" tends to be higher when expressing agreement or understanding.

Regarding the patterns of F0 contours in "ununn" and "uun" utterances, it was shown that part could be discriminated by the up-down F0 movements or by spectral changes in the nasal part. However, it was also shown that utterances accompanied by laugh, final rising and vocal fry can be problematic. Future work includes evaluation of discrimination of "ununn" and "uun" based on acoustic features.

5. Acknowledgements

This work was partly supported by the Ministry of Education, Culture, Sports and (MEXT Kakenhi) and Japan Science and Technology Corporation (JST). We thank Mika Morita and Kyoko Nakanishi for helping in the data analysis.

6. References

[1] Ishi, C.T., Ishiguro, H., Hagita, N., "Automatic extraction of paralinguistic information using prosodic features related to F0," duration and voice quality. Speech Communication 50(6), 531-543, June 2008.

[2] Ishi, C.T., Ishiguro, H., and Hagita, N, "The meanings of interjections in spontaneous speech," Proc. Interspeech' 2008, 1208-1211, 2008.

[3] Ishi, C.T., Ishiguro, H., and Hagita, N., "Analysis of acoustic-prosodic features related to paralinguistic information carried by interjections in dialogue speech," Proc. Interspeech' 2011, 3133-3136, 2011.

[4] Ishi, C.T., Hatano, H., Hagita, N., "Extraction of paralinguistic information carried by mono-syllabic interjections in Japanese," Proceedings of The 6th International Conference on Speech Prosody (Speech Prosody 2012), 681-684, 2012.

[5] Sudo, J., "The Japanese interjection un: From its meanings and functions to an analysis of its phonetic features", Journal of the Phonetic Society of Japan, Vol.11 No.3, 94-106, 2007 (in Japanese)

[6] Togashi, J., "Aizuchi hyougen keishiki-ni miru shinnai-no jouhou syori-ni tsuite", Working papers for special project of Tsukuba university "touzai gengo bunka-no ruikeiron", 27-42, 2002 (in Japanese).

[7] Yoshida, E., "Detecting patterns of sequences by coding scheme and transcribed utterance information: An analysis of Japanese reactive tokens as non-primary speaker's role", Proceedings of The 3rd workshop of Japanese corpus, 435-440, 2013 (in Japanese).

[8] Kokenawa, Y., Tsuzaki, M., Kato, H. and Sagisaka, Y., "An analysis of speaking attitude manifesting as fundamental frequency characteristics", Technical report of IPSJ SIG, 87-92, 2004 (in Japanese).

[9] Clancy, P. M., Thompson, S. A., Suzuki, R. and Tao, H., "The conversational use of reactive tokens in English, Japanese, and Mandarin", Journal of Pragmatics, 26, 355-387, 1996.

[10] Ward, N., "Non-lexical conversational sounds in American English", Pragmatics and Cognition, 14, 113-184, 2006.

[11] Ishi, C.T., Liu, C., Ishiguro, H., and Hagita, N. (2010). "Head motion during dialogue speech and nod timing control in humanoid robots," Proceedings of 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2010), 293-300.

[12] Liu, C., Ishi, C., Ishiguro, H., Hagita, N. (2013). Generation of nodding, head tilting and gazing for human-robot speech interaction", International Journal of Humanoid Robotics (IJHR), vol. 10, no. 1, January, 2013.