# Analysis of acoustic-prosodic features related to paralinguistic information carried by interjections in dialogue speech

*Carlos T. Ishi* [1], *Hiroshi Ishiguro* [1,2], *Norihiro Hagita*[1]

[1] Intelligent Robotics and Communication Labs., ATR, Kyoto, Japan
[2] Dept. of Adaptive Machine Systems, Osaka University, Japan
carlos@atr.jp, ishiguro@sys.eng.osaka-u.ac.jp, hagita@atr.jp

## Abstract

Interjections are often used in dialogue communication for expressing a reaction (such as agreement, surprise and disgust) to the interlocutor. Thus, a correct interpretation of the paralinguistic information (intention, attitude or emotion) carried by interjections is important for achieving a smooth dialogue interaction between humans and robots. In the present work, analyses are conducted on several interjections appearing in spontaneous conversational speech databases to investigate the relationship between acoustic-prosodic features (related to intonation and voice quality) and their paralinguistic functions in dialogue speech. It is found that there are common and interjection-dependent relationships between acoustic features and paralinguistic information. Regardless of the interjection type, non-modal voice qualities, such as whispery, harsh and pressed voices, are shown to be important cues for the expression of emotions and attitudes.

**Index Terms**: interjection, prosody, voice quality, paralinguistic information, speech act, emotion

## 1. Introduction

Besides the linguistic information, the understanding of paralinguistic information (such as intentions, attitudes and emotions) is also important for achieving a smooth dialogue communication between humans and robots. Interjections like “eh” and “un” are often used to express a reaction to the interlocutor’s utterance in spontaneous dialogue speech, and usually convey some paralinguistic information related to intention, attitude, or emotion, such as agreement, surprise or disgust. Also, most of the paralinguistic information is conveyed by prosodic features, including variations in intonation and voice quality.

Most works regarding paralinguistic information extraction, have focused on prosodic features related to intonation and rhythm, such as F0 (fundamental frequency), power and duration. However, it has been shown that voice quality information (caused by non-modal phonations, such as breathy, whispery, creaky and harsh [1]) also plays important roles, when analyzing natural conversational speech data, mainly in expressive speech utterances [2-6].

In our previous work [6], we have proposed a framework for paralinguistic information extraction considering the speaking styles represented by prosodic features related to intonation and voice quality, as shown in Fig. 1. We analyzed the roles of the speaking styles, for discrimination of several paralinguistic information items carried by “e” and “un”, which are the most commonly used interjections in Japanese. However, although some relationship was found between acoustic-prosodic features and paralinguistic information for “e” and “un”, it is not guaranteed that these relationships can be straightly extended to other interjections.

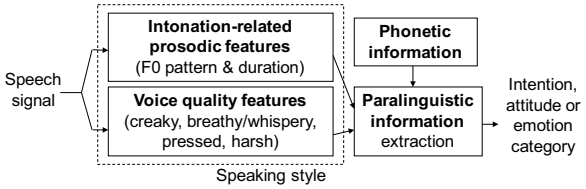


Figure 1: *A framework for paralinguistic information extraction considering prosodic and voice quality features.*

Thus, we also have conducted analyses on the variations in speaking style and paralinguistic information carried by several interjection types appearing in Japanese spontaneous dialogue speech [7]. It was found that although most of the paralinguistic information carried by an interjection depends on its speaking style, there is also dependency on the interjection type, (i.e., on its phonetic contents).

In the present work, we conduct a deeper analyses on the relationship between acoustic-prosodic features and the paralinguistic information, by comparing the similarities and differences in four groups of interjection: “o/oo”, “on/oon”, “ya/yaa”, and “wa/waa”.

## 2. Speech data and labeling

### 2.1. Speech data

The data consist of interjection samples extracted from three databases of Japanese spontaneous dialogue speech. The speech databases used in the present work are the same used in our previous work [7].

One is the JST/CREST ESP expressive speech database [8]. Data of ten speakers (seven female and three male speakers) were used for analysis. This database has in total about 200 hours of spontaneous conversational speech, and is very rich in terms of interjections.

The second database is the dialogue data set (D01-D04) of the CSJ (“Corpus of Spontaneous Japanese”) speech corpus [9]. The dialogues are between speakers which are familiar and not familiar with each other. Speakers are male and female aging from 20s to 50s.

The third database is a multi-modal conversational database newly recorded at the ATR IRC Labs. This database contains 42 free dialogue conversations of 10 to 15 minutes each, by 4 male speakers (from 20s to 40s) and 3 female speakers (from 30s to 50s). For the present work, only the transcriptions and the speech data were used.

The interjection tokens were obtained by executing a text search in the text transcriptions of each database. Only interjections in the beginning of an utterance were searched.

In the present work, we focused on the interjections which carry multiple paralinguistic information, according to their

speaking styles. Samples of the interjections can be listened at <http://www.irc.atr.jp/~carlos/interjection/>. In our previous work [6], we have conducted acoustic analyses for the interjections "e, ee" and "un, uun". In the present work, we focus on four other interjection types: "o, oo", "on, oon", "ya, yaa", and "wa, waa".

### 2.2. Annotation of paralinguistic information and speaking styles

The paralinguistic information items carried by an interjection were annotated for each utterance, by listening to the audio segment including 5 seconds before and after the interjection. The audio of the conversation partner was also available to allow contextual information. The paralinguistic information items for each interjection are based on the online dictionary [10,11], and on results of past publications [7,12,13]. However, new items were allowed to be freely added by the subjects. The selection of multiple items for an utterance was also allowed, since the items are not completely mutually exclusive.

The used label items, with their corresponding meaning/nuance, are shown below.

- Backchannel: feedback to the interlocutor "I'm listening."; "Uhm."
- Unexpected: "I was not expecting that."
- Admired: "That's impressive."
- Surprised: "That's amazing."
- Ask for repetition: "Say again."
- Notice: "I just noticed/realized/remembered."
- Affirm/agree: "Yes."; "I think so."; "Right."
- Deny/negate: "No."; "I don't think so."; "Wrong."
- Negative reaction (dissatisfied/blame/suspicious): "I can't agree/accept without objecting."
- Disgusted: "This is disgusting."; "This is unpleasant."
- Long backchannel: "I'm listening; keep talking."; "Uhmm."
- Doubt: "I wonder if..."
- Sigh/tired: "I'm tired."
- Sigh/relieved: "I'm relieved to hear that."
- Suffer (moan): "I feel pain."
- Understanding: "I see; I know what you mean."
- Deny(modesty): "I don't say yes, but I appreciate about what you said."
- Hesitated/embarrassed/confused: "I'm embarrassed about what you said/asked."
- Sympathy, pity, disappointed: "It is a pity about that."; "I'm disappointed with that."; "That's awful!"
- Happy: "It is nice to hear that!"
- Envy: "How I envy you!"
- Thinking/Fillller: "I'm thinking what to utter next. Please wait."
- Short filler (conjunction, opener): used as a cushion for the next utterance. "So, …", "I mean, …"

The paralinguistic information items were annotated by four subjects, and utterances where two or more subjects agreed were used for subsequent analysis. Table 1 shows the subjective agreement in terms of the percentage of utterances where at least two subjects annotated the same label.

Table 1. *Subjective agreement on paralinguistic information labels for each interjection group.*

| Interjection type | o/oo | on/oon | wa/waa | ya/yaa |
|---|---|---|---|---|
| N. of utterances | 371 | 157 | 289 | 728 |
| Subj. agreement | 74.4% | 78.4% | 73.0% | 72.9% |

The labels related to prosodic features related to intonation and voice quality were annotated by a subject with experience in prosody and voice quality (the first author).

The following set was used for intonation annotation.

- *Rs* (rise): pitch rising within a syllable.
- *Fa* (fall): pitch falling within a syllable.
- *Ft* (flat): flat tone; no pitch change.
- *Rt* (reset): pitch reset (often confused with *Rs*; *Rt* corresponds to a perceptual raising of pitch between syllables)
- combination of the above tones, such as *FaRs*, *FtFa*, *RtFtFa*

The duration of the interjection was automatically estimated from the speech signal.

For voice quality, the following set was used.

- *w* (whisper): (unvoiced) whisper.
- *wv* (whispery voice): whispery or breathy voice (presence of aspiration noise during phonation).
- *c* (creaky, vocal fry): very low fundamental frequencies with impulse-like pulses, usually with irregularity in periodicity.
- *h* (harsh): rasping noisy voice.
- *p* (pressed): pressed voice.
- *a* (aspirated): aspiration in the end of utterance.
- combination of the above voice qualities, such as *pc*, *hwv*.

## 3. Analysis results

The left panel of Fig. 2 shows the distribution of different voice qualities for each interjection type, after the annotation results. The item *w* includes *w* and *wv*, *p* includes *p* and *pc*, and *h* includes *h* and *hwv*. It can be observed from the figure that more that 40% of the tokens in "wa/waa" were accompanied by a pressed voice quality (*p*), while almost 40% of the tokens in "ya/yaa" were accompanied by a whispery voice quality (*w*). The right panel of Fig. 2 shows the distributions of duration for each interjection type. Table 2 shows the distribution of the paralinguistic items for each interjection group. These results will be discussed in this section, for each interjection type.

Table 2. *Distribution of the paralinguistic items, for each interjection group*.

| Paralinguistic item | o/oo | on/oon | wa/waa | ya/yaa |
|---|---|---|---|---|
| backchannel | **67** | **78** | 5 | |
| affirm/agree | 4 | **15** | | |
| long backchannel | **35** | **19** | 3 | |
| understand | **31** | 2 | 1 | 1 |
| ask repetition | 4 | 1 | | |
| doubtful | | 3 | 1 | 4 |
| unexpected | 9 | 1 | **10** | 1 |
| admired | **28** | | **57** | 8 |
| surprised | **26** | 1 | **29** | 8 |
| noticed | **18** | | 5 | **13** |
| negative reaction | 1 | 1 | **29** | 8 |
| hesitated | 7 | 1 | **19** | **11** |
| sympathy | | | **36** | |
| thinking/filler | **11** | | 8 | 3 |
| short filler | **35** | | 7 | **81** |
| deny(short filler) | | | | **109** |
| deny | | | | **263** |
| deny(modesty) | | | | 6 |

For intonation-related acoustic-prosodic features, we use tone maps for each interjection type, as shown in Fig. 3, where the vertical axis is the segmental duration and the

horizontal axis is *F0move*, which is a parameter that quantifies the pitch movements within a syllable in semitone unit, based on human perception properties [14]. It is calculated by splitting a syllable in two parts, estimating an average F0 value in the first part, a target F0 value in the end of the second part, and calculating the difference between the representative F0 values in the two parts [14]. Positive *F0move* values indicate rising tones, negative *F0move* values indicate falling tones, and *F0move* around 0 (e.g., from -2 to 1 semitone) are flat tones. The tokens with more than one F0 movement within a syllable (such as *FaRs*), which cannot be plotted in the tone map will be discussed separately.

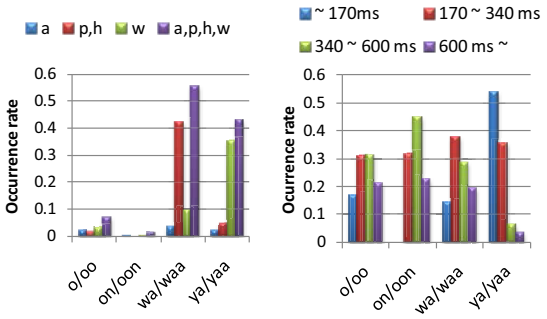


Figure 2 *Distribution of non-modal voice qualities (left) and duration (right), for each interjection group.*

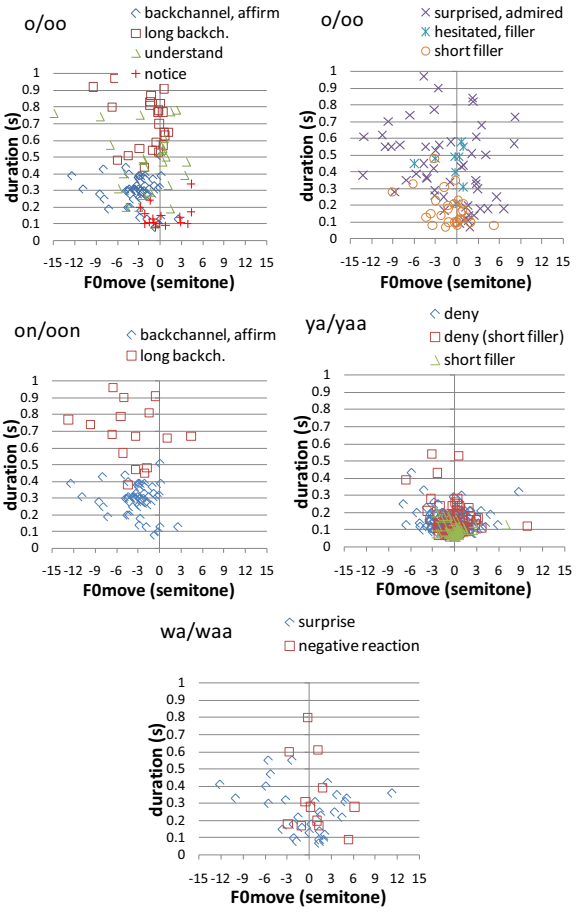


Figure 3 *Distribution of prosodic features for different paralinguistic items, and for each interjection group.*

### 3.1. o/oo

The interjection "o" is defined in the dictionary as a "voice sound occurring when a subject is surprised, suddenly notices" while the interjection "oo" is defined as "when a subject 1) is impressed; 2) is surprised or feels suspicious; 3) suddenly notices or suddenly remembers; 4) accepts or strongly answers."

A large variety of paralinguistic items was observed for "o/oo" in the present data, as shown in Table 2. Part of the "o/oo" tokens appeared to express surprise, admiration or noticing, while a large amount of the "oo" tokens appeared in casual situations mostly as conversational variations of "un", expressing backchannels or understanding. In this case, the vowel quality is not like a clear /o/ sound, but something between /oo/, /on/, /un/ and /an/. Also, it can be observed that the use as a short filler or opener (without specific emotion or attitude meanings) also appeared with a significant amount.

Regarding the prosodic features, two types of speaking style were predominant in the backchannel group, as shown in the tone maps of the top panels in Fig. 3. One is the common pattern of short-fall tones appearing in the "backchannel, affirm" group, being consistent with our past results for "un" [6]. The other pattern appearing in "long backchannel" is long fall or long flat tones, with segmental duration above 0.4 seconds. This type of backchannel is something in between a classical "agree-type" backchannel, as in short-fall tones, and "admiration" expression.

Several patterns appear in "surprised/admired" item, as shown in the top right panel in Fig. 3. Samples can be observed in short-flat or short-rise tones (shorter than 200 ms), long rise and long fall tones. "hesitated/filler" are concentrated on long flat, while "short filler" are concentrated in short flat. However, we can observe overlaps between paralinguistic groups, e.g., between "notice" and "short filler", so that further context information would be necessary for disambiguation.

### 3.2. on/oon

The interjections "on" and "oon" are not defined in the dictionary, being conversational variations of the interjection "un". They were transcribed as "on" and "oon", but their pronunciation sounds between /on/, /an/ and /un/, as part of the tokens in "o/oo". They appeared mostly in casual situations, for expressing backchannels or affirm/agree, as shown in Table 2.

It can be observed in the mid-left tone map of Fig. 3 that short-fall tones are the most frequent pattern in "backchannel, affirm" group, as in "on" and "un", followed by long backchannels. Comparing with "o/oo", it can be noted that long-flat tones are predominant in "oo", while long-fall tones are predominant in "oon".

### 3.3. ya/yaa

"ya/yaa" includes variations like "iya", "iyaa" and "iiyaa". The word "iya" means "I don't like" or "I don't want", but its use as an interjection is much more common. The interjection "iya" is defined in the dictionary as "1) deny, "no"; 2) word used occasionally, without specifically agreeing or disagreeing;", while the interjections "ya" and "yaa" are defined as "1) voice sound occurred when surprised or unexpected; 2) word used to call someone or to respond to someone's calling."

In the present database, the most frequent item was "deny", as shown in Table 2. However, although "ya" has a negation meaning, a large amount of tokens were labeled as "short filler" or an opener, without specific negative meanings.

We can observe in the tone map of Fig. 3 that most of "ya" tokens are short in duration (shorter than 200 ms). The tones are flat, slightly falling or slightly rising. However, from

the prosodic features of short "ya" utterances, there is no clear distinction between "deny", "noticing" or "short filler", so that contextual information would be necessary for their discrimination.

For the long "yaa" tokens (longer than 400 ms), almost all tokens were accompanied by a change in voice quality (a,p,h,w), when some emotion or attitude (surprise, negative reaction, or hesitation) is expressed. Pressed voice quality often appears to express a deep surprise or embarrassment. A high-pitched voice or a harsh/whispery voice quality also appears to express a strong surprise.

### 3.4. wa/waa

"wa/waa" includes variations like "uwaa" and "uuwa". The interjection "wa" is defined in the dictionary as: "1. Voice sound occurring when surprised; 2. Voice sound occurring when loudly crying or laughing."

In the present database, most of the tokens were labeled as surprised or admired. However, many tokens were also labeled as "negative reaction" (including disgust), "hesitated" and "sympathy", as shown in Table 2.

From the tone map in Fig. 3, overlap between these two groups of paralinguistic items can be observed.

Regarding voice quality, 56% of the "wa" tokens were accompanied by a change in voice quality (pressed, whispery and harsh whispery), as was shown in Fig. 2. However, the use of these voice qualities was not discriminative of paralinguistic items. The effects of pressed voice quality in "waa" utterances were of increasing the degree of surprise, admiration or disgust, while the effects of harsh/whispery voice qualities in "wa" utterances were of increasing the degree of surprise.

Regarding intonation, 56% were flat tones (including tokens where F0 could not be computed due to irregular phonation). In the 42% of the tokens, where pressed voice was accompanying, intonation patterns like fall-flat-rise and fall-flat-flat, where an F0 dip occurs in the middle of the token, were frequent. This phenomenon of F0 reduction in pressed segments occurs due to physiological constraints, rather than due to a conscious production of intonation curve, and the perception of voice quality change is more prominent than a change in the tone shape [15]. Therefore, in utterances including pressed segments, one should be careful in the interpretation/use of F0 curves.

In other 7% of the tokens, attitudes and emotions were expressed by high pitch. However, for the use of absolute pitch heights, an F0 normalization will be necessary for different speakers.

## 4. Discussion

Discrimination of part of the paralinguistic information presented in Section 3 can roughly be realized based on prosodic and voice quality features, so that a "dictionary" of speaking styles and corresponding paralinguistic functions can be created for each interjection group. However, there are ambiguities between many of the paralinguistic information items, which would need additional contextual information for disambiguation. For example, it was reported in a past work that the automatic extraction of paralinguistic information resulted in about 75% for the interjections "*e*" and "*un*" [6]. The errors were in part due to acoustic discrimination errors, but part due to ambiguities in the acoustic space. Improvements on acoustic discrimination and analysis of context dependency are subjects for future work.

Although there are several difficulties for recognizing the paralinguistic information carried by an interjection, the results of the present work could be more straightly applied to generation of the speaking styles of specific interjections for paralinguistic information expression, so that a more human-like reaction could be produced by robots in human-robot dialogue interactions.

## 5. Conclusions

We conducted acoustic-prosodic analyses for investigating the relationship between speaking styles and the paralinguistic information conveyed by interjections appearing in spontaneous speech. We found similarities and differences in the paralinguistic information carried by different interjections according to their speaking styles.

In particular for voice quality, among the common features found over several interjections, pressed and harsh/whispery voices often appeared to strength the expression of emotions like surprise, admiration and disgust.

However, it was found that part of the paralinguistic items cannot be discriminated only by their speaking styles, so that additional information on context is unavoidable.

Future works include investigation of contextual information, evaluation of automatic extraction of paralinguistic information by using acoustic-prosodic features, and the application of the dictionary of interjections for human-robot dialogue interactions.

## 6. Acknowledgements

This work is partly supported by the Ministry of Internal Affairs and Communication. We thank Jun Arai for his contributions in the paralinguistic annotation procedure.

## 7. References

[1] Laver, J., Phonatory settings. In *The phonetic description of voice quality*. Cambridge University Press, Ch. 3, pp. 93-135, 1980.

[2] Klasmeyer, G.; Sendlmeier, W. F., Voice and Emotional States. In *Voice Quality Measurement*, Singular Thomson Learning. 339-358, 2000.

[3] Gobl, C.; Ní Chasaide, A., The role of voice quality in communicating emotion, mood and attitude. *Speech Communication* 40, 189-212, 2003.

[4] Maekawa, K., "Production and perception of 'Paralinguistic' information," Proc. *Speech Prosody 2004*, 367-374, 2004.

[5] Erickson, D., Expressive speech: production, perception and application to speech synthesis. *Acoust. Sci. & Tech*., Vol. 26 (4), 317-325, 2005.

[6] Ishi, C.T., et al., Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality. *Speech Communication* 50(6), 531-543, June 2008.

[7] Ishi, C.T., et al. "The meanings of interjections in spontaneous speech," Proc. *Interspeech' 2008*, 1208-1211, 2008.

[8] Campbell, N., speech & expression; the value of a longitudinal corpus," Proc. *LREC2004*, 183-186, 2004.

[9] Maekawa, K., H. Koiso, S. Furui, and H. Isahara. "Spontaneous speech corpus of Japanese," Proc. *LREC2000*, 947-952, 2000.

[10] Goo online dictionary, http://dictionary.goo.ne.jp/

[11] Yahoo online dictionary, http://dic.yahoo.co.jp/

[12] Masuoka, T., Takubo, Y., *Kiso nihongo bunpoo (Basic Japanese Grammar)*, Kuroshio, 60-61, 1992. (in Japanese)

[13] Maynard, S.K., *An introduction to Japanese grammar and communication strategies*, The Japan Times, 115-117, 221-224, 264-265, 1990.

[14] Ishi, C.T. (2005) Perceptually-related F0 parameters for automatic classification of phrase final tones. *IEICE Trans. Inf. & Syst.*, Vol. E88-D, No. 3, 481-488

[15] Ishi, C.T., et al. "Acoustic, electroglottographic and paralinguistic analyses of "rikimi" in expressive speech," *Speech Prosody 2010*, ID 100139, 1-4, 2010.