Extraction of paralinguistic information carried by mono-syllabic interjections in Japanese

Carlos Toshinori Ishi, Hiroaki Hatano, & Norihiro Hagita

Intelligent Robotics and Communication Laboratories
ATR, Kyoto, Japan
{carlos; hatano; hagita}@atr.jp

Abstract

Mono-syllabic interjections are often used to express a reaction in conversational speech. It is known that there is relationship between the speaking style, given by intonation and voice quality-related prosodic features, and the paralinguistic information carried by an interjection. However, it is also known that this relationship is dependent on the interjection type. In the present work, we analyzed the relationship between speaking style and the conveyed paralinguistic information item for several mono-syllabic interjection types in Japanese. Evaluation results show that acoustic parameters related to intonation and voice quality features in conjunction with the identity of the interjection are effective for disambiguating 71% of the paralinguistic information items.

1. Introduction

Besides the linguistic information, the understanding of paralinguistic information (including intentions, attitudes and emotions) is also important in spoken dialog systems, especially in non-verbal communication using grunt-like utterances such as “eh”, “ah”, and “un”. Such utterances are frequently used to express a reaction to the interlocutor’s utterance in a dialogue scenario, for expressing an 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 [7], 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.

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 [8,9]. 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).

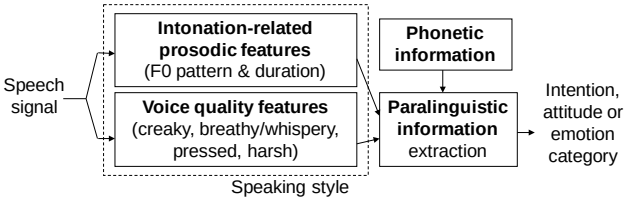

Figure 1: *Our framework for paralinguistic information extraction considering prosodic and voice quality features.*

In [9], we analyzed the discriminability of paralinguistic information by acoustic features for the interjections “o”, “on”, “ya” and “wa”, appearing in spontaneous speech. We found similarities and differences in the paralinguistic information conveyed by different interjections according to their speaking styles. In particular, a large overlap in the acoustic space was found among different items for “wa” and “ya”.

Regarding acoustic feature extraction, we also have proposed several acoustic parameters for representing the features of syllable tones and specific voice qualities [10-13]. In the present work, we make use of these acoustic parameters and evaluate how well they can discriminate between different paralinguistic information items, for several types of monosyllabic interjections.

2. Description of the speech data for analysis

In the present work we focused on the monosyllabic interjections “a”, “o”, “ha”, “ho”, and “he” (including variations such as “ah”, “oh”, “aa”, “aaaa”, and so on), which in conjunction with the previously studied “e”, “un, “wa” and “ya” cover almost all monosyllabic interjections appearing in Japanese. As stated in the introduction, these interjections carry a large variety of paralinguistic information depending on the speaking style. Possible paralinguistic information (speech acts, attitudes or emotions) conveyed by varying the speaking styles of the above interjections are listed below. These items were obtained from our past works concerning analyses of paralinguistic information conveyed by these interjections on conversational speech databases [8,9].

- backchannel (agreeable responses) (***backch***),
- understand, consent (***underst***)
- ask for a repetition (***askrep***),
- surprise, amazed or astonished (***surp***), unexpected (***unexp***)
- admired or be impressed (***adm***),
- notice (***notice***),
- embarrassed, hesitated (***emb***),
- blame, dissatisfaction (***blm***),
- sympathy, compassion (***symp***),
- tired (***tired***),
- disappointed (***disap***),
- suffer, moan (***suffer***).

As the items of the list are difficult to be clearly separated in terms of intentions, attitudes, or emotions, the term "paralinguistic information" (PI) is used in this paper to refer to all items of the list.

In the present work, speech data is recorded in order to get a balance in terms of the paralinguistic information carried by each interjection. For that purpose, sentences were elaborated in such a way to induce the subject to produce a specific PI.

The "inducing" sentences are spoken by one native speaker. Then, subjects are asked to produce a target PI, i.e., utter in a way to express a determined PI, through the specified interjections. Short sentences are also elaborated to be spoken after the interjections, in order to obtain a reaction as natural as possible. However, a short pause is requested between the interjection and the following short utterance.

Utterances spoken by 7 subjects (3 male and 4 female speakers between 20s and 30s) were recorded. In addition to the PI list, speakers are also asked to utter the interjections in a pressed voice quality, which frequently appears in natural expressive speech [6], but is more difficult to naturally occur in an acted scenario.

All interjection intervals are manually segmented for subsequent analysis and evaluation.

3. Paralinguistic information data

Perceptual experiments were conducted on the recorded data, to verify if the intended (induced) PI can be correctly recognized in two conditions. One is by listening only to the interjection, i.e., in a context-free situation, while the second is by listening also the utterance following the interjection, i.e., by considering some context information.

Two annotators (native speakers of Japanese) were asked to choose one or multiple items, from the PI list, that could be expressed by each of the 286 stimuli (i.e., the segmented "a", "o", "ha", "he" and "ho" utterances).

Firstly, regarding the "with-context" vs. "context-free" conditions, different PI items were attributed in 21% of the tokens for one of the annotators, while 34% for the other annotator. After grouping some of the PI items with close meanings (backchannel+understand and surprise+admiration), the above numbers reduce to 16% and 19%. Joining the intra-rater disagreement of the two annotators (between with-context and context-free conditions), it resulted on 25%. This roughly implies that in about 25% of these interjection stimuli, the discrimination of the PI item will be ambiguous based only on the speaking style of the interjection. On the other hand, we could expect that in 75% of the utterances, PI items could be correctly discriminated by using the interjection identity and speaking style information.

Regarding inter-rater agreement, 27% disagreement rate was obtained between the two annotators. The disagreement rate reduced to 22% after PI grouping (backchannel+understand and surprise+admiration).

The 78% of the stimuli, where agreement was achieved between the two raters were then used for subsequent analysis.

4. Extraction of paralinguistic information from speaking style and interjection identity

In this section, we analyzed the relationship between the speaking style (i.e., the prosodic features related with intonation and voice quality) and the paralinguistic information conveyed by each interjection type. The purpose of this analysis is to verify how good the differences in speaking style can distinguish between different PI items.

In the following sub-sections, we describe acoustic parameters that potentially represent the perception of features related with the different speaking styles, for discrimination of the paralinguistic information.

4.1. Acoustic parameters related with intonation

In [10], a set of parameters was proposed for describing the intonation patterns of phrase final syllables, based on F0 and duration information.

For the pitch-related parameters, F0 is first estimated based on a classical method of peak picking in the normalized autocorrelation function of the LPC inverse-filtered residue of the pre-emphasized speech signal. All F0 values are converted to the musical (log) scale before any subsequent processing.

The (monosyllabic) utterance is broken in two segments of equal length, and representative F0 values are extracted for each segment. In [10] several candidates for the representative F0 values were tested, and here, we use the ones that best matched with perceptual scores of the F0 movements. For the first segment, an average value is estimated using F0 values within the segment (***F0avg2a***). And for the second segment, a target value is estimated as the extrapolated F0 value at the end of the segment of a first order regression line of F0 values within the segment (***F0tgt2b***). A variable called ***F0move*** is defined as the difference between *F0tgt2b* and *F0avg2a*, quantifying the amount and direction of F0 movement within the syllable. *F0move* is positive for rising F0 movements, and negative for falling movements.

Fig. 2 shows the distributions of *F0move* and duration for each paralinguistic information item. The PI items are separated in two panels for better visualization of the distributions. The items where F0 could not be estimated, due to irregularities in periodicity, are not shown in the figure.

From the distributions of the parameters shown in the figure, we can observe that the PI items can be partly discriminated by using intonation-related prosodic features. For example, "backchannel" tokens are concentrated on short fall tones (duration < 0.4 seconds, *F0move* < -3 semitones), "notice" tokens are concentrated on short flat and slightly rising tones (duration < 0.3 seconds, -2 < *F0move* < 3 semitones), "asking for repetition" tokens are concentrated on short rising tones (duration < 0.3 seconds, *F0move* > 2

semitones), "blame" tokens are concentrated on rising tones (*F0move* > 6 semitones) , "tired" tokens are concentrated on very long falling tones (duration > 0.4 seconds, *F0move* < 0). The "understand" token distributions show mainly two types of tones. One is similar to "backchannel", with short falling tones, while the other is a flat or slightly rising tones (-2 < *F0move* < 3 semitones). In the bottom panel of Fig. 2, it can be observed that "sympathy" tokens are concentrated on long flat or long slightly falling tones (duration > 0.4 seconds, -3 < *F0move* < 1 semitones). And finally, "surprise" and "admiration" tokens are spread over a large area, being overlapped with other PI items.

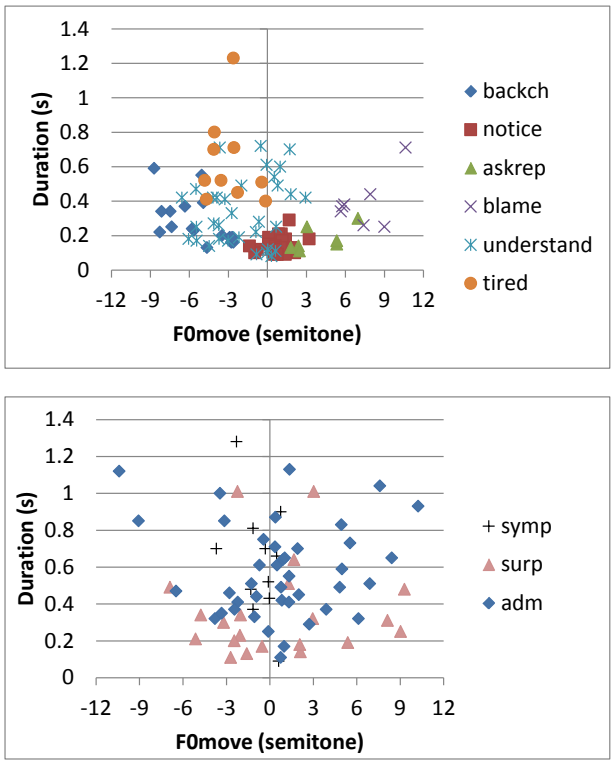

Figure 2: *Distributions of the prosodic parameters for each perceived paralinguistic information item.*

In the following section, voice quality related features are used for disambiguating between some of the paralinguistic information items which are overlapped in the intonation space.

4.2. Acoustic parameters related with voice quality features

In this section, the use of parameters related with voice quality (non-modal phonations) is evaluated for a better discrimination between paralinguistic information items that cannot be discriminated by the only use of prosodic features. In the present work, we focus on the use of parameters related to breathy and pressed voice qualities, which have been shown to convey paralinguistic information in past works [12,13].

4.2.1. Detection of breathy segments

Breathy segments refer to turbulent noise due to an air escape at the glottis, occurring in vowel intervals. Thus, breathy segments include breathy, whispery and aspirated sounds.

The breathy segment detection algorithm is based on our recently proposed parameter in [13]. A normalized breathiness power (***NBP***) measure is estimated as the power in the mid-frequency band weighted by a parameter called *F1F3syn*.

NBP is then given by the following expression:

$$NBP = P_{MF} + w_{voicing} \cdot 10\log_{10}(1-|F1F3syn|), \quad (1)$$

where P_{MF} is the power of the mid-frequency band in dB, normalized (subtracted) by the maximum value of the whole-band power in the utterance, and $w_{voicing}$ is a weighting factor for the contribution of *F1F3syn* measure.

F1F3syn is a measure of synchronization (using a cross-correlation measure) between the amplitude envelopes of the signals obtained by filtering the input speech signal in two frequency bands, one around the first formant (F1) and another around the third formant (F3). The boundary between F1 and F3 bands is adaptively set to 500, 1500 or 2500 Hz, according to presence/absence of periodicity in the F3 band. If breathiness is absent, *F1F3syn* has values close to 1, while if it is present, *F1F3syn* has values closer to 0. Therefore, the *NBP* measure in expression (1) will be the mid-frequency power biased by larger negative amounts, as the *F1F3syn* value becomes closer to 1, i.e., when breathiness is absent. More details about the algorithm and evaluation of the parameter can be found in [13].

Here, breathy segments are detected in the frames where *NBP* (normalized breathiness power) is larger than -20 dB, using a weighting factor of 4 for the effects of *F1F3syn* parameter ($w_{voicing}$). A syllable is detected as breathy if breathiness is detected over 5 consecutive frames (i.e., 50 ms).

12 of the 15 tokens (80%) in "tired" were detected as breathy in "a" and "ha". In "sympathy" 7 of the 9 tokens (78%) with falling tones were detected as breathy. On the other hand, 9 of the 33 tokens (27%) in "understand" were detected as breathy, which means that part of the above items can be discriminated by using breathiness information.

2 of the 21 tokens (9%) in "backchannel", 8 of the 25 tokens (32%) in "noticed", and 19 in 26 tokens (73%) in "surprise" were also detected as breathy, so that, breathiness is effective for discriminating part of these PI items with overlap in the intonation pattern.

4.2.2. Detection of pressed segments

Pressed segments refer to a phonation type produced by pressing/straining the vocal folds in voiced intervals.

We use a spectral tilt measure proposed in [12], called H1'-A1', which deals with the problem in pressed segments where the harmonic structure is disturbed or sometimes inexistent due to irregularities in periodicity. In such cases, in place of H1 and A1, the use of the maximum peak amplitude in the range of 100 to 200 Hz (H1'), and the maximum peak amplitude in the range of 200 to 1200 Hz (A1'), where the first formant is likely to be present, are employed. For periodic signals, H1' = H1, and A1' = A1.

A threshold of 15dB (H1'–A1' < 15dB) was used for identifying pressed segments. A syllable is detected as pressed if pressed voice is detected over 5 consecutive frames (50 ms).

Pressed segments were detected in 36 tokens (12 for the interjection "a" and 24 for the interjections "ha", "he" and "ho"). For the 12 tokens of "a", 8 (67%) were "suffer", while the remaining were "hesitation". For the 24 tokens of "ha", "he" and "ho", 20 (84%) were "admiration". These results indicate that the above PI items can be discriminated from the others, resolving the ambiguity in the intonation-related feature space, as was shown in Fig. 2.

4.3. Discriminability of paralinguistic information by using acoustic-prosodic parameters and interjection type

Fig. 3 shows the detection rates of the perceived paralinguistic information items by using the acoustic features related to intonation and voice quality, described in the Sections 4.1 and 4.2, and the interjection identity. A classification tree was constructed for discriminating between the PI items.

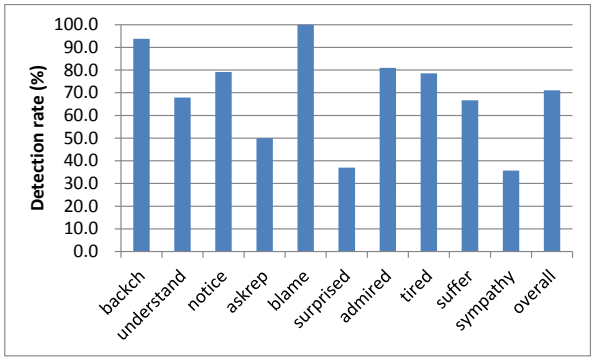

Figure 3: *Detection rates of the perceived paralinguistic information items by using acoustic features related to intonation and voice quality, and interjection identity.*

An overall detection rate of 71% could be achieved. The results in Fig. 3 shows detection rates lower than 50% in "ask for repetition", "surprise" and "sympathy". In the case of "ask for repetition", confusion was found with "surprise", due to breathiness detection in these tokens. "Surprise" was confused with several other types of PI (such as "backchannel", "understand", "notice", "ask for repetition", and "blame"). "Sympathy" was confused with "tired" and "suffer", which share similar breathy flat and falling tones. For disambiguating most of the detections errors, context information is unavoidable. This is subject for future work.

Regarding the effects of the acoustic features, intonation related features only (F0move and duration) lead to 56%, the addition of breathiness parameter improves to 59%, the addition of pressed parameter improves to 65% and finally, the addition of the interjection identity increases the overall detection rate to 71%.

5. Conclusions

The relationship between the speaking style and the paralinguistic information conveyed by the interjections "a", "o", "ha", "he" and "ho" was analyzed, and the discrimination of paralinguistic information items by using acoustic prosodic features and the interjection identity was evaluated.

Analysis results indicated that prosodic features are effective for discriminating part of the paralinguistic information items with specific functions (backchannels/ understand, notice, ask for a repetition, and blame), while voice quality features are effective for identifying items expressing some emotion or attitude. The detection of pressed voice is effective for identifying "suffer" in the interjection "a", and "admiration" for the interjections "ha", "he" and "ho". The detection of breathy voice is effective for discriminating "tired" and "sympathy" from "understand" in long flat and long fall tones, and "surprise" from "notice" in short flat and short rise tones. In conjunction with the identity of the interjection, an overall detection of 71% could be achieved for discriminating the paralinguistic information.

Future works include evaluation of a full automatic detection by running speech recognition for providing the interjection type, and allowing an appropriate mapping between speaking styles and the conveyed paralinguistic information. Another remaining and important topic is how to deal with context information for disambiguating between items sharing the same speaking style.

6. Acknowledgements

This research is supported by the Ministry of Education, Culture, Sports, Science and Technology.

7. References

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

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

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

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

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

[6] Sadanobu, T., 2004. A Natural History of Japanese Pressed Voice, *J. of Phonetic Society of Japan*, Vol. 8 (1): 29-44.

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

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

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

[10] 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.

[11] Ishi, C.T., Sakakibara, K-I., Ishiguro, H., Hagita, N. (2008). A method for automatic detection of vocal fry. *IEEE Trans. on Audio, Speech and Language Processing*, Vol. 16, No. 1, 47-56, Jan. 2008.

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

[13] Ishi, C.T., Ishiguro, H., and Hagita, N. (2011). "Improved acoustic characterization of breathy and whispery voices," Proc. of *Interspeech' 2011*.