9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

Analysis of relations between hand gestures and dialogue act categories

Carlos Ishi [1], *Ryusuke Mikata* [1], *Hiroshi Ishiguro* [1]

[1] ATR Hiroshi Ishiguro Labs.

carlos@atr.jp, mikata.ryusuke@irl.sys.es.osaka-u.ac.jp, ishiguro@sys.es.osaka-u.ac.jp

Abstract

Hand gestures commonly occur in daily dialogue interactions, and have important functions in communication. In this study, we analyzed a multimodal database of three-party conversations, and investigated the relations between the occurrence of hand gestures and speech, with special focus on dialogue act categories. Analysis results revealed that hand gestures occur with highest frequency in turn-keeping phrases, and seldom occur in backchannel-type utterances. On the other hand, self-touch hand motions (adapters) occur more often in backchannel utterances and in laughter intervals, in comparison to other dialogue act categories.

Index Terms: hand gestures, dialogue acts, visual prosody, non-verbal communication, natural conversation

1. Introduction

The background of this work is the generation of natural motions in humanoid robots, matched with the speech utterances. So far, we have investigated the relations between speech and several modalities including facial, head and torso movements, accounting for dialogue act functions, laughing speech and surprise utterances [1-3], and proposed several methods for generating natural motions in humanoid robots from the speech signal of a tele-operator [4-6]. Besides facial and head movements, hand gestures also commonly occur in dialogue interactions, having important functions in human-human communication [7-10]. Although there are controversies on whether hand gestures are speaker-directed or listener-directed, we consider that in either of the cases they are important for expressing human-likeness in human-robot interactions. Thus, in the present study, we focus on the analysis of hand gestures from a production perspective.

Several studies have been conducted on text or speech-driven gesture synthesis in CG animated agents [11-15]. For example, lexicon-based approaches have been proposed for generating iconic gestures in [11]. An imagistic description tree was proposed for representing the semantics of shape-related expressions in [12]. A framework that combines data-driven with model-based techniques to model the generation of iconic gestures with Bayesian decision networks was proposed in [13]. In [14], gesture and speech features are associated and modeled by HMMs, and directional (pointing) gestures are generated. In [15], a system that converts text into an animated agent by synchronizing gestures and speech was implemented. It is reported that lexical and syntactic information are strongly correlated with gesture occurrences, suggesting that syntactic structures are more useful for judging gesture occurrences than local syntactic cues.

Prosodic information has also been exploited when generating hand gestures, mainly by considering relations between prosodic focus (emphasis) and beat gestures. For example, relationship between gestures and intonation has been investigated for English conversational data in [16]. It is reported that apexes of gestural strokes and pitch accents aligned consistently, and gestural phrases and intermediate phrases aligned quite often. In [17], a prosody-based approach has been proposed for synthesis of body language, by associating motion and speech streams. It is reported that realistic and compelling body language could be produced, for English. However, we consider that it is not guaranteed that these studies can be straightly applied for any language, since prosody is language-dependent. For example, pitch-accent languages like Japanese and tonal languages like Chinese have lexicon-dependent prosodic variations. Indeed, it is reported that the relationship between head motions and speech prosody differs from languages [18]. In our previous research we also have pointed out that head motion occurrences are more related to the dialogue act functions of the utterances, rather than to prosodic features in Japanese [1].

From the above cited past studies, we can say that iconic gestures can be associated with the lexical contents, while beat gestures can be associated with the prosodic features of the speech utterances. However, the past studies remain unclear when and to what extent gestures should be generated. Hand gestures do not occur in every utterance, so that some criterion is necessary to decide when to generate or not a gesture.

In the present study, we analyzed hand gesture events in face-to-face human interactions in a multimodal three-party dialogue database, and investigated how the occurrences of hand gestures are affected by functional properties of the speech utterances.

2. Analysis data

2.1. Description of the data

For analysis, we use a dataset of the multimodal three-party conversational speech database collected at our research institute (ATR). The database contains multiple sessions of face-to-face conversations among three speakers. The data includes audio, video and motion information of each speaker. Fig. 1 shows a picture of the data collection environment setup.

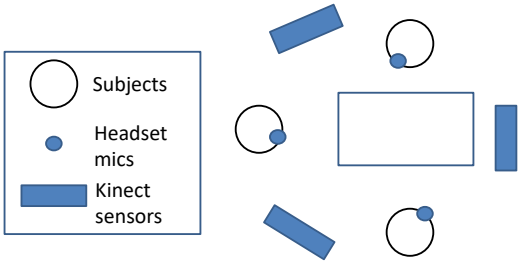


Figure 1. *Environment setup for the three-party dialogue collection.*

Three headset microphones (DPA4060) and three RGB-D sensors (Microsoft Kinect-V2) were used to capture data of each speaker. The dialogue participants had a seat in chairs around a table, with a distance of about 2 meters between each other. The table is 60cm height, so that the Kinect sensors can

10.21437/SpeechProsody.2018-96

detect hand motions around the participants' knee area. The video data from the Kinect sensors were used in the analysis of the present work.

Each dialogue session comprises about 30 minutes of random topic conversations. The speech utterances were segmented in phrase units (accentual phrases) and text-transcribed by a native speaker. For the present analysis, data of 2 Japanese dialogue sessions by a total of 6 native speakers (5 female and 1 male speakers) were used. All female participants are all in their 30s, while the male speaker was 22 years old. The speaker IDs are F01 ~ F05 for the female speakers, and M01 for the male speaker. All speakers within a session knew each other. The speakers F01, F02 and F03 participated in one session, and talked about topics like lunch, their experiences abroad, and their future plans. The speakers F04, F05 and M01 participated in the other session, and talked about dating, relationships and personality.

2.2. Annotation data

Dialogue act categories were annotated for each phrase, according to the following label set, which is based on previous studies [1].

- Interjectional backchannels (**bc**): feedback responses like "*un*", "*hai*" (equivalent to "*uhm*", "*uh-huh*", "*yes*" in English)
- Non-interjectional backchannels (**bc2**): "*hontodesuka*", "*sugoine*" ("*really?*", "*great!*"); also includes feedback utterances by repeating the words or phrases of the dialogue partner utterances.
- Turn-giving statements (**g**)
- Turn-giving questions (**q**): utterances requesting an answer from the dialogue partner.
- Turn-giving/keeping (**gk**): utterances ambiguous between turn-giving and turn-keeping
- Turn-keeping in strong boundaries (**k**): boundaries between intonational phrases; a short pause or a clear pitch reset is accompanied.
- Turn-keeping in weak boundaries (**k2**): boundaries between accentual phrases.
- Fillers (**f**): utterances like "*eetoo*", "*anooo*" (equivalent to "*uhmmm*", "*I mean…*" in English)

In comparison to the previous dialogue act label set, the category "*bc2*" was introduced in the present study, since non-interjectional backchannel utterances were partly mixed with the turn-giving category "*g*" in the previous studies.

A research assistant (native speaker of Japanese) annotated the labels above, by listening to the speech utterances.

Hand gestures are also categorized according to their functions [9,10]. The following categorization was adopted in the present research.

- Iconic: Gestures presenting images of concrete entities and/or actions. The gesture, as a referential symbol, functioning via its formal and structural resemblance to event or objects.
- Metaphoric: Gestures not limited to depictions of concrete events. They also picture abstract content, in effect, imagining the unimageable. In a metaphoric gesture, an abstract meaning is presented as if it had form and/or occupied space.
- Deictic (pointing): The prototypical deictic gesture is an extended 'index' finger, but almost any extensible body part or held object can be used.
- Beats: so called because the hand appears to beating time. Beats are mere flicks of the hand(s) up and down or back and forth, zeroing in rhythmically on the prosodic peaks of speech.
- Emblems: conventionalized signs, such as thumbs-up or the ring (first finger and thumb tips touching, other fingers extended) for "OK". Emblems are culturally specific, have standard forms and significances, and vary from place to place.

Besides the above categories, we also take the following category into account.

- Adapters: movements that often involve self-touch, such as scratching or touching the hairs. Adapters happen almost entirely unaware, and may or not be accompanied by speech. Adapters can reflect the perceived emotional stability [19].

In the present work, we will refer as "hand motion" to all items described above, and "hand gestures" to all items excluding adapters.

The hand gestures were further segmented in gesture phases according to [8, 9].

- Preparation (optional): The hands from a rest position moving into a gesture space where it can begin the stroke.
- Stroke (obligatory): The stroke is the gesture phase with meaning.
- Retraction (optional): the hands returning to the rest position (not always the same position as at the start). There may not be a retraction phase if the speaker immediately moves into a new stroke.
- Pre- and post-stroke hold phases (optional): temporary cessations of motion either before or after the stroke motion. Holds ensure that the meaningful part of the gesture – the stroke – remains semantically active during the co-expressive speech.

The gesture phases were segmented by the second author, by looking at the videos and listening to the speech contents. In the present research, a separate segmentation layer was conducted for beat gestures, since those can co-occur with the hold phases of other gestures. For example, beats can occur during a deictic gesture or a metaphoric gesture.

For each gesture stroke segment, the gesture categories were annotated by a research assistant, also by looking at the video and listening to the speech contents.

3. Analysis results

3.1. Distributions of hand motions

The overall distributions of the hand motions were first analyzed. Fig. 2 shows the distributions of hand motion types and Fig. 3 shows the distributions of the total durations for utterance, hand gesture and adapters, of each speaker.

Firstly, it can be seen that among the gesture categories, beat gestures occur with the highest frequency. Part of this gesture occurs along with other gesture categories, while the other part occurs individually. Some of the speakers (F02, F05 and M01) have more adapters than others. The speaker F04 employed less hand gestures in comparison to the other

speakers. However, it can also be observed in Fig. 3 that F04 talked much less than the others. The percentages of the time when the hand gestures (excluding adapters) occur relative to the summed time of the utterance intervals are: 23%, 35%, 41%, 26%, 67% and 31% for speakers F01, F02, F03, F04, F05 and M01 respectively.

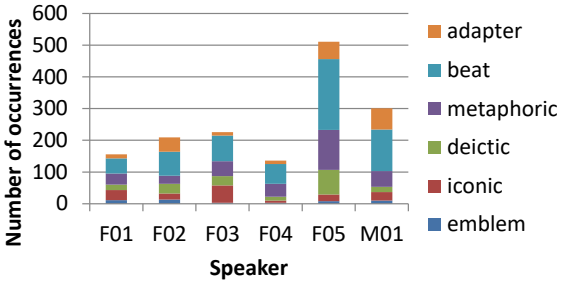


Figure 2. *Distributions of hand motion types for each speaker.*

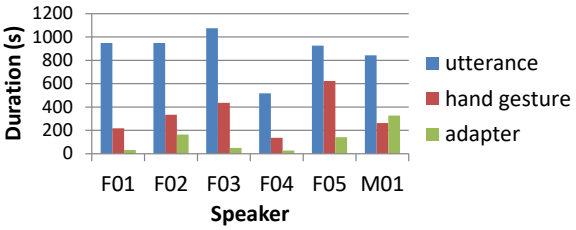
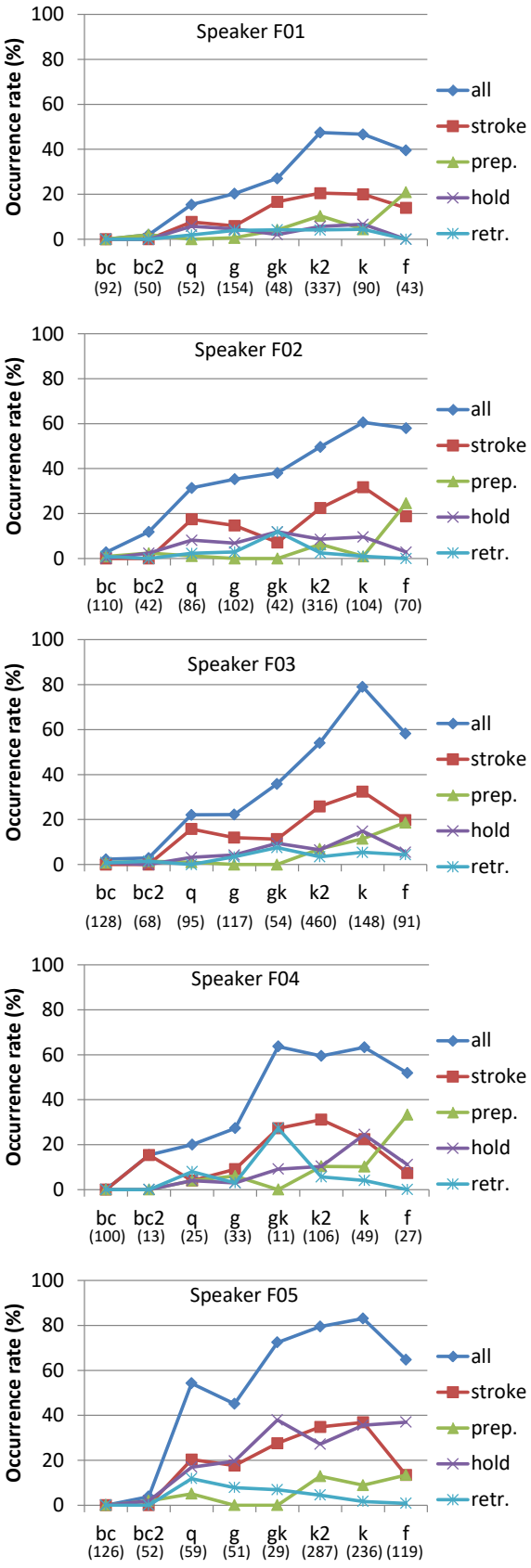


Figure 3. *Distributions of total durations for utterances, hand gestures and self-adaptors, for each speaker.*

3.2. Relations between hand gestures and dialogue act categories

The number of overlapping incidents between hand gesture intervals and speech intervals were counted in each dialogue act categories. The adapter (self-touch) hand motion events are removed from the present analysis and discussed in the next sub-section. Fig. 4 shows the distributions of hand gesture occurrences in different dialogue act categories, for each of the six speakers. The number of occurrences in each dialogue act category is shown within brackets. The distributions are shown for all hand motion intervals ("*all*"), and for each gesture phase intervals (stroke, preparation, hold and retraction).

From the results in Fig. 4, it can firstly be observed that for all speakers, the occurrence rates of hand gestures during both interjectional and non-interjectional backchannel utterances ("*bc*" and "*bc2*") show very low occurrence of hand gestures. Higher occurrence rates can be observed for turn-keeping ("*k*" and "*k2*"), fillers ("*f*") and "*gk*" (ambiguous category between turn-giving and turn-keeping), in almost all speakers (excluding speaker M01). This suggests that when the speaker is in a listening mode, where backchannels are predominant, the occurrence rate of gestures is low, while when the speaker is in speaking mode, the occurrence rate of gestures becomes higher.



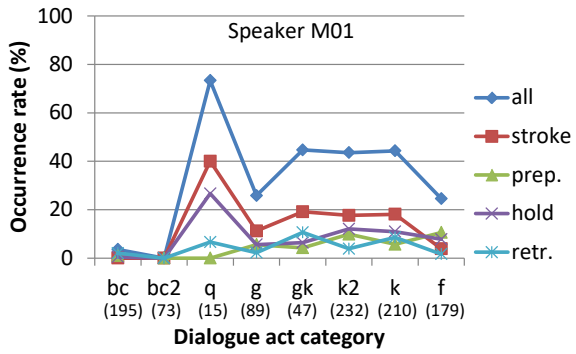


Figure 4. *Occurrence rates of hand gestures in different dialogue act categories, on the whole gesture intervals ("all") and on each gesture phase. The total number of utterances in each dialogue act is shown within brackets.*

In questions and turn-giving phrases ("*q*" and "*g*"), the occurrence rates of gesture were intermediate between backchannels ("*bc*") and turn-keeping ("*k*") categories, in almost all speakers. This indicates that gestures occur with higher frequency in the middle of long sentences (where phrases with turn-keeping occur) rather than at the end of the sentences (where "*g*" and "*q*" phrases occur).

Regarding the fillers ("*f*"), relatively high occurrence rates are observed on the whole gesture intervals ("*all*"), but relatively lower occurrence rates are observed for the stroke intervals only ("*stroke*"), in comparison to the others. It is also observed that preparation and hold phases are predominant during fillers ("*prep.*" and "*hold*").

The results in Fig. 4 showed similar trends in the hand gesture occurrences among all speakers, but also showed some differences for specific speakers. For example, speaker F04 was the only speaker in which hand gesture strokes were observed during non-interjectional backchannel utterances ("*bc2*"). A close look in the data revealed that these utterances were expressing negation/denial ("*iyaiya sonna koto nai*" meaning "*not really*", with a modest attitude), and the accompanying hand gestures were also expressing negation/denial (by shaking the hands).

Regarding the speaker M01, higher occurrence rates are observed for question-type utterances ("*q*"), in comparison to the other speakers. A close look in the data revealed that most of the question utterances were clarification-type questions (e.g., "*…toyuu kotodesuka?*" meaning "*you mean that …?*"). A more detailed classification of dialogue act categories may lead to a clearer relationship between hand gestures and dialogue acts. This raises another point and needs further investigation.

3.3. Relation between adapters and dialogue act categories

We also investigated if there are any relationships between hand motion types and dialogue act categories. Although no clear evidences were found for specific hand gesture categories (iconic, metaphoric and deictic), some trends were found for adapters.

Fig. 5 shows distributions of adapter-related hand motions for each dialogue act category, and each speaker. Individual laughter intervals (i.e., excluding the intervals of laughing while speaking) were also considered in the present analysis, since self-touch motions of bringing one of the hands in front of the mouth were often observed during laughter.

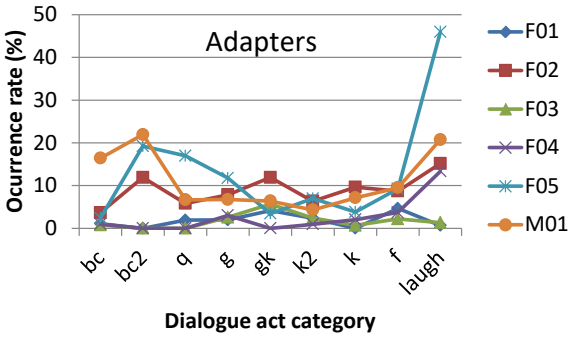


Figure 5. *Distributions of adaptor-related hand motion in each dialogue act category, for each speaker.*

The results in Fig. 5 indicate that two speakers (F05 and M01) present higher occurrence rates of adapters during laughter and backchannel utterances ("*bc2*"). Although hand gestures (excluding adapters) occur with less frequency during backchannel utterances, as shown in the results of Fig. 4, adapters may occur with higher frequency.

4. Discussions

The analysis results in Section 3 showed relations between hand gesture occurrence and dialogue act categories. We also conducted intonation analysis on the gesture-accompanied phrases. Although relations between intonation and gesture are reported for English [16], a straight relationship has not been found for our Japanese dialogue data. However, hand gestures and intonation features could be indirectly associated through dialogue acts, since previous studies on Japanese dialogue speech have indicated relationship between tones and dialogue acts [20,21].

5. Conclusions

In this study, we analyzed the relationship between hand motion events and dialogue act categories, in three-party face-to-face dialogue interactions.

We found that the both interjectional and non-interjectional backchannel-type utterances are seldom accompanied by hand gestures, whilst hand gestures are regularly employed with turn-keeping phrases rather than with turn-giving phrases. On the other hand, it was found that the occurrence rates of adapters (self-touch movements) were higher in backchannel utterances and in laughter, in comparison to other dialogue act categories.

Future works should cover analysis of linguistic information linked to the hand gestures, and application of the analysis results to hand gesture motion generation in humanoid robots.

6. Acknowledgements

This study was supported by JST, ERATO, Grant Number JPMJER1401. We thank Taeko Murase, Miki Okuno, Megumi Taniguchi and Kyoko Nakanishi for contributions in the annotations and data analyses.

7. References

[1] Ishi, C., Ishiguro, H., Hagita, N. (2013). Analysis of relationship between head motion events and speech in dialogue conversations. *Speech Communication 57*, No.2014, 233–243, June 2013.

[2] Ishi, C., Hatano, H., Ishiguro, H. (2016). "Audiovisual analysis of relations between laughter types and laughter motions," Proc. of the 8th international conference on Speech Prosody (*Speech Prosody 2016*), pp. 806-810, May, 2016.

[3] Ishi, C., Minato, T., Ishiguro, H. (2017). "Motion analysis in vocalized surprise expressions," Proc. Interspeech 2017, pp. 874-878, Aug. 2017.

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

[5] Ishi, C., Funayama, T., Minato, T., Ishiguro, H. (2016). "Motion generation in android robots during laughing speech," IEEE/RSJ International Conference on Intelligent Robots and Systems (*IROS 2016*), pp. 3327-3332, Oct., 2016.

[6] Ishi, C., Minato, T., Ishiguro, H. (2017). "Motion analysis in vocalized surprise expressions and motion generation in android robots," *IEEE Robotics and Automation Letters*, Vol.2, No.3, 1748 - 1754, July 2017.

[7] Kendon, A. "Gesticulation and speech: two aspects of the process of utterance," In M. R. Key (ed), The Relationship of Verbal and Nonverbal Communication, The Hague: Mouton and Co, pp.207-227, 1980.

[8] McNeill, D. (1992). Hand and Mind: What Gestures Reveal about Thought, Chicago and London: The University of Chicago Press, 1992.

[9] Kita, S. (2000). How representational gestures help speaking. In D. McNeill (ed.), Gesture and Language, pp. 162-185. Cambridge: Cambridge University Press.

[10] McNeill D. (2006). Gesture: a psycholinguistic approach, in The Encyclopedia of Language and Linguistics, eds Brown E., Anderson A. (Amsterdam; Boston: Elsevier;), 58–66

[11] J. Cassell, M. Stone, and H. Yan. (2000) "Coordination and context-dependence in the generation of embodied conversation," In Proc. of the *First International Conference on Natural Language Generation*, 2000.

[12] T. Sowa and I. Wachsmuth (2005). "A model for the representation and processing of shape in coverbal iconic gestures." In Proc. *KogWis05*, pages 183–188, 2005.

[13] Bergmann, K., and Kopp, S. (2009) "GNetIc - Using Bayesian Decision Networks for Iconic Gesture Generation" in Proc. of *the 9th International Conference on Intelligent Virtual Agents*, Ruttkay, Z., Kipp, M., Nijholt, A., and Vilhjálmsson, H. eds. LNAI, vol. 5773, (Berlin/Heidelberg, Germany: Springer), pp.76-89, 2009.

[14] Sargin, M.E., Aran, O., Karpov, A., Ofli, F., Yasinnik, Y., Wilson, S., Erzin, E., Yemez, Y., Tekalp, A.M. (2006). Combined gesture–speech analysis and speech driven gesture synthesis. In: Proc. of *IEEE International Conference on Multimedia*.

[15] Y.I. Nakano, M. Okamoto, D. Kawahara, Q. Li, T. Nishida. (2004). "Converting Text into Agent Animations: Assigning Gestures to Text," In Proc. *Human Language Technology Conference of the North American Association for Computational Linguistics* (HLT-NAACL 2004), pp. 153–156.

[16] Loehr, D. 2004. Gesture and Intonation. Washington DC: Georgetown University, PhD Dissertation.

[17] S. Levine, C. Theobalt, V. Koltun (2009). "Real-Time Prosody-Driven Synthesis of Body Language," In *SIGGRAPH* Asia 2009.

[18] Yehia, H.C., Kuratate, T., Vatikiotis-Bateson, E., 2002. Linking facial animation, head motion and speech acoustics. *J. Phonetics* 30, 555-568.

[19] M. Neff, N. Toothman, R. Bowmani, J.E. Fox Tree, and M. Walker (2011). "Don't Scratch! Self-adaptors Reflect Emotional Stability," in Proc. *10th International Conference on Intelligent Virtual Agents* (IVA 2011), Reykjavik, Iceland, September 15-17, 2011, pp.398-411.

[20] 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.

[21] Ishi, C.T., Ishiguro, H., and Hagita, N. (2006). "Analysis of prosodic and linguistic cues of phrase finals for turn-taking and dialog acts," Proc. *Interspeech 2006*, pp. 2006-2009.