# The distribution of pitch patterns and communicative types in speech-chunks preceding pauses and gaps

*Irena Yanushevskaya, John Kane, Céline De Looze, Ailbhe Ní Chasaide*

Phonetics and Speech Laboratory, School of Linguistic, Speech and Communication Sciences
Trinity College Dublin, Ireland

yanushei@tcd.ie, kanejo@tcd.ie, deloozec@tcd.ie, anichsid@tcd.ie

## Abstract

As part of a broader study of voice prosody in speech communication, this paper looks at intonation in turn-taking. It examines the distribution of pitch patterns and communicative types in the interpausal units (IPUs) preceding pause or gap silences extracted from a corpus of spontaneous speech of Irish English. IPUs preceding speaker change ('Gaps') and IPUs preceding silence where the same speaker continues talking ('Pauses') were selected in the course of automatic extraction of pause/gap silences in dyadic dialogue interactions. A listening test was conducted to establish 'human predictable' pause/gap data sets which were subsequently manually annotated in terms of pitch patterns and communicative types. Overall, the Gaps and Pauses subsets show differentiation in terms of both their communicative types and pitch tunes. Declaratives and Questions are mainly found in Gaps, whereas in Pauses we mainly find Hesitations and Incomplete Declaratives. Gaps are generally characterised by falling or rising pitch patterns, whereas in Pauses a large proportion of speech samples are realised with level pitch. Classification experiments reveal discrimination of pauses and gaps for both prosodic and functional annotation labels. Follow-up work aims to relate intonational characteristics of turn taking with voice quality and temporal dynamics, to provide a holistic view of the processes involved.

**Index Terms**: dialogue speech, pause, gap, intonation, communicative type

## 1. Introduction

This paper is part of a broader study of the interaction of intonation and voice source parameters in prosody at the Phonetics and Speech Lab, Trinity College Dublin. The present study complements parallel research exploring voice source [1], $f_0$ and temporal features (see, for example [2] on the role of prosodic features as well as [3] on $f_0$ range declination trends in turn-taking organisation). In this paper we describe intonational (pitch patterns) and functional (communicative types) annotation of the interpausal units (IPUs) preceding pause or gap silences extracted from a corpus of spontaneous speech of Irish English.

Robust prediction of turn-taking is crucial for dialogue systems. To date, prediction is largely based on the duration of pause or gap silent intervals, and on the speech interval immediately preceding them. The present detailed analysis of both functional and intonational characteristics of pre-silence chunks for a corpus of Irish English, aims to establish their linkage to turn-taking, and their potential for discriminating speaker changes vs. holds. It further aims to provide the intonational baseline with which we can later correlate voice source and temporal features.

The decision on whether and when we begin to speak in a conversation depends on numerous factors, e.g., lexical and syntactic [4], prosodic [5], vocal effort and audible respiratory cues [6], as well as gestural signals (e.g., head movement, gaze [7] etc.). The importance of lexical-syntactic features for turn-taking management has been emphasised since the early scientific work in this field [4, 8] as well as in more recently reported work [9]. In fact, the perceptual experiment carried out in [9] showed that artificially flattening intonation contours had less of an impact on the predictability of 'end-of-turn' compared to artificially removing the intelligibility of the utterance (by low-pass filtering).

Nonetheless, many researchers focus entirely on prosodic features, an approach which is somewhat justified given that previous studies (e.g., [2, 10, 11]) have found significant discriminative power of prosody-related features, and that this has direct relevance for prosody-only turn prediction (e.g., [12]) in dialogue systems. The role of prosodic patterns in turn-taking has been discussed in [5, 9, 13], see also references therein. For a number of languages (English, German, Dutch, Japanese and Mandarin Chinese), it has been reported that level pitch accents or flat contours at the end of an utterance are indicative of a pause (silent interval within the speech of the same speaker) while any other terminal contours such as rises and falls are indicative of a gap (silent interval between the speech of different speakers) [14-19].

However, the picture emerging is not always as clear cut. In some studies, similar intonation contours have been found for both turn-taking and turn-holding. In [20], 51% of the rising intonation patterns co-occurred with speaker changes while 49% of rises were associated with speaker holds. Furthermore, most studies report a high level of inter-speaker variability.

Other researchers have suggested that turn-taking is likely to be positively affected by the number of prosodic cues present [21, 22]. In addition to pitch contours, prosodic features reported as contributing to turn-management include voice quality (e.g., creaky voice [23]), speech rate and final lengthening [24]. At the level of prosody, we feel it is the dynamic patterning of the voice as a whole (the combination of intonation, voice quality and temporal aspects) that effectively cues speaker changes and holds. While the focus of this paper is on the formal and functional aspects of pitch contours as relating to turn-taking, the bigger picture is a longer term objective.

## 2. Speech data

### 2.1. Recordings

The speech data for annotation is taken from the Dublin Institute of Technology Emotional Speech Corpus [26] which consists of seven 10-minute dyadic (male-male and female-female, Irish English) interactions. Six dyadic interactions involving six male and six female speakers were selected from



10.21437/SpeechProsody.2014-180

the original corpus to ensure gender balance. The interactions were elicited in a shipwreck scenario game where participants were presented with 15 items and were given 10 minutes to jointly rank them in order of usefulness for their survival. Recordings were made with participants in separate booths using a professional Neumann microphone connected to an Apple Mac-based Digidesign Pro-Tools Mbox2 recording system. The audio signal was recorded using Pro-Tools software as two separate audio streams and digitised at 96 kHz/24 Bit. Audio was then downsampled to 16 kHz/8 Bit.

### 2.2. Extraction of the IPUs preceding pause and gap silent intervals

Automatic identification of pauses and gaps was carried out on the speech data using an approach similar to that described in [27]. Binary voice activity detection (VAD) using the VAD algorithm proposed in [28] was carried out on both speaker channels for each dyadic interaction. The threshold for silence interval duration was set to 200 ms to avoid false detection of pauses for speech events like plosives. Silent intervals below this threshold were bridged. Fig. 1 illustrates schematically the output of the VAD process. Overall, 460 gaps and 410 pauses were identified automatically.

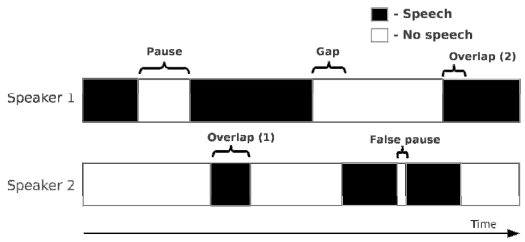



Figure 1. *Schematic representation of a dialogue interaction illustrating pauses, gaps and overlaps. The 'false pause' indicates a silence which is below the threshold (here set to 200 ms).*

As one of our goals is to establish whether and to what extent prosodic characteristics of the speech-chunks immediately preceding pause or gap silent intervals allow automatic prediction of turn-taking in human-machine interaction, we (in a previous study [2]) selected a subset of data where pauses and gaps were clearly predictable by human listeners. A listening test was conducted in which the IPUs (n=870) preceding automatically identified pause and gap silent intervals were presented individually to three raters in random order. Each rater was to indicate, on a 5-point scale, whether in their opinion a pause (same speaker continues) or a gap (speaker change) follows. The rating scale was defined as follows: (1) Very certain the CURRENT speaker continues, (2) Quite certain the CURRENT speaker continues, (3) Don't know! (4) Quite certain the OTHER speaker begins, (5) Very certain the OTHER speaker begins. The raters had an option to indicate that there was an error in the automatic extraction of stimuli, e.g., due to premature truncation of utterances. In total, 6% of the stimuli were marked as an error by the raters. The inter-rater agreement was measured using Krippendorff's α [29]. Analysis revealed fairly high inter-rater agreement (α = 0.74). Only the samples which all three raters identified as being followed by a pause or a gap were retained to form the ultimate 'human predictable' dataset. In total, 302 IPUs preceding gaps and 288 IPUs preceding pauses were retained, which amounts to 70% of the original dataset. This 'human predictable' data set was subsequently manually annotated to explore the distribution of pitch patterns and communicative types of the utterance in the speech-chunks immediately preceding pause and gap silent intervals. For simplicity, we will refer to the IPUs immediately followed by gaps (speaker change) as 'Gaps' and the IPUs immediately followed by pauses (same speaker continues talking) as 'Pauses'. The terms 'Pauses' and 'Gaps' are therefore used to refer to speech-chunks immediately before silent intervals rather than the silent intervals themselves. The data samples in the 'human predictable' data set represent male and female speakers in fairly equal proportion, however the amount of data selected from individual speakers is not necessarily balanced.

## 3. Annotation of the selected data

The selected Gaps and Pauses data were annotated separately. The aim of this preliminary annotation was to explore the patterns in the distribution of communicative types and pitch tunes in these two data subsets. The manual annotations involved auditory analyses of the extracted data and were initially done by one annotator. Pitch patterns were independently analysed by a second annotator, and the two annotators agreed on 71% of the data. The labels used for the annotation are described in the sections below.

**COMMUNICATIVE TYPES:** A variety of approaches exist in dialogue speech annotations, e.g., [30-32] and annotation schemes usually include both communicative types of the utterance and functional analyses of dialogue acts. Here we report only the results of communicative type annotation. Since the IPUs were obtained by automatic extraction using a predefined minimum pause threshold, they may contain more than one sentence. An example of such IPU would be *I still like the life jacket. You could drown, like.* In such cases, only the sentence closest to the pause/gap silence (in this example, *You could drown, like*) was analysed. The following communicative type labels were used:

**Declarative** - grammatically complete/well-formed declaratives, e.g., *Yeah, but I mean we've already lost so many points*.

**Incomplete Declarative -** grammatically incomplete fragmented declaratives, e.g., *…starting from the most important. Seven…er… a knife.*

**Yes/No-Q** - grammatically complete/well-formed Yes/No questions, e.g., *Does that mean we picked right every other one?*

**WH-Q** - grammatically complete/well-formed WH-questions, e.g., *What's the knife gonna do?*

**Incomplete Q** - grammatically incomplete questions, e.g*., I'd say radio next? …the survival guide and the knife, does it?*

**Alt Q** - alternative questions, e.g., *Did you say compass or map again ?*

**Dec Q -** declarative question, e.g., *Yep, flare sounds good to me, so flare's for four?*

**Tag Q -** tag questions, e.g., *We did do, didn't we?*

**Exclamation -** [largely based on the intonation with which these utterances were produced], e.g., *Binoculars! Binoculars! Oh god! We did so well!*

**Imperative -** e.g., *So, do you wanna rank them? Mac, please!*

**Hesitation -** filled pauses, e.g., *erm, ahm, er,* repetitions and self-corrections, or markedly prolonged words, e.g., *And th[e:]n….*

**Backchannel** - short acknowledgements such as *sure, yeah, uhum, yep.*

Other labels included '**?**' for uncertain cases and '**n/a**' for the IPUs with no propositional content (e.g., only laughter).

**PITCH PATTERNS:** In the annotation of pitch patterns in Pauses and Gaps subsets we described the nuclear tunes (each containing the intonation-phrase final pitch accent and its associated boundary tone) using the IViE system [33]. The following tune labels were used:

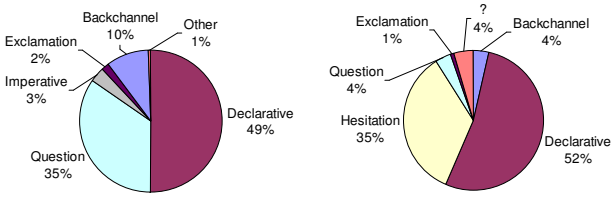
| | |
|---|---|
| **H*+L 0%** | fall |
| **H*+L H%** | fall-rise |
| **L*+H 0%** | low rise |
| **H* H%** | high rise |
| **L*+H L%** | rise-fall |
| **!H* 0%** | downstep (e.g., in item lists) |
| **H* 0%** | level (no change/movement of pitch)**.** |

Analysis of the pitch patterns reported below was conducted on the final intonational phrase in the sentence closest to the pause/gap silent intervals rather than on the whole IPU.

# 4. Results and discussion

## 4.1. The distribution of communicative types

The distribution of communicative types in the Gaps and Pauses subsets is shown in Figure 2. Due to space limitations, the different types of questions are pooled into one category and the same is done for well-formed and incomplete declaratives.




(a) Gaps (b) Pauses

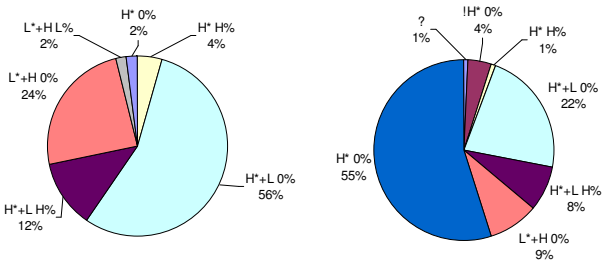
Figure 2. *The distribution of communicative types in the Gaps (a) and Pauses (b) subsets.*

As is clear from Figure 2, almost half of the IPUs in the Gaps subset are Declaratives (of which about a quarter are incomplete). Questions comprise 35% of the Gaps subset, with Incomplete Questions and Yes/No Questions making up the majority of the question types. Among the remaining communicative types, Backchannels are the most frequent (10%), with only 5% of IPUs classified as either Imperatives or Exclamations.

Similar to Gaps, a large part of the Pauses subset is made up by Declaratives (55%). However, the majority of the declaratives in Pauses are incomplete (74% of all declaratives). A fairly large proportion of the IPUs from the Pauses subset are classified as hesitations, a communicative type that does not appear in the Gaps set. The proportion of questions in the Pauses subset is 4% which is substantially lower than in the Gaps set. The proportion of Backchannels is also lower in Pauses, only 4%. The least common communicative type occurring in Pauses is Exclamation (1%). In 4% of the cases the communicative types of the extracted IPUs in the Pauses set were ambiguous and were not annotated.

## 4.2. The distribution of pitch patterns

The overall distribution of pitch patterns in the Gaps and Pauses data subsets is given in Figure 3 (the data is pooled across all communicative types). More than half of the intonation phrases (IPs) preceding gaps (56%) are realised with a falling pitch, H*+L 0%. Rises L*+H 0% (24%) and fall-rises H*+L H% (12%) are the next most frequent tune types in Gaps. The most frequently occurring pitch pattern in the IPs preceding pauses is level tone H* 0% (55%). The second most common tune type here is fall H*+L 0%, although its proportion in the Pauses subset is substantially lower (22%) than in the Gaps subset (56%). Generally speaking, Gaps are characterised overwhelmingly by pitch movement, whereas Pauses have level tone in the majority of cases. These findings corroborate what has been described in the literature, e.g., [5, 17]. A more detailed analysis of pitch patterns that characterise communicative types most frequently found in Gaps and Pauses is given in the sections below (although not all are illustrated due to space limitations).




(a) Gaps (b) Pauses

Figure 3. *The distribution of pitch patterns in the Gaps (a) and Pauses (b) subsets (across all communicative types).*

A closer look at the pitch tunes in different communicative types found in the Gaps set reveals similar pattern for Declaratives, Incomplete Declaratives, WH Questions and Backchannels: in about 60-70% of the cases, the H*+L 0% tune (fall) is used, followed by L*+H 0% (low rise) as the next most frequent tune type. The proportion of rises is the lowest in Declaratives (10%), it is higher in Incomplete Declaratives and WH questions (17%), and is the highest in Backchannels (24%). In Incomplete Questions, the pitch is predominantly rising (only 12% of samples here have falling pitch). In Yes/No Questions, both falling and rising pitch pattern is used, with a slight preference for rises (about 54% in total).

The distribution of pitch patterns in the Pauses subset is examined mainly for Declaratives and Hesitations which comprise 98% of this data set. Overall, Hesitations are realised predominantly with level pitch (H* 0%), with only a small proportion (14%) having a falling pitch. Incomplete Declaratives are realised with either level or rising pitch (in 80% of the cases), with falls occurring in only 20% of the cases. The proportion of falling pitch is higher in [well-formed] Declaratives (41%), however, in the majority of cases the pitch is either rising or stays level.

#### *4.2.1. The distribution of tunes: a case of Declaratives*

We compare here in some detail the distribution of tunes in declaratives which make up about 50% of communicative types in each of these two data sets (see Figure 2). The tunes found in Declaratives and Incomplete Declaratives are shown in Figure 4 separately for Gaps (left panel, a) and Pauses (right panel, b). Note that the proportion of incomplete and complete declaratives is reversed in the Pauses set compared to the Gaps set: well formed/complete declaratives constitute 76% of all declaratives in Gaps, but their proportion is reduced to only 26% in Pauses.

There is relatively little difference in the distribution of tunes between complete and incomplete declaratives in the Gaps set. A striking feature of incomplete declaratives in Pauses is a high number H* 0% (level) tunes.

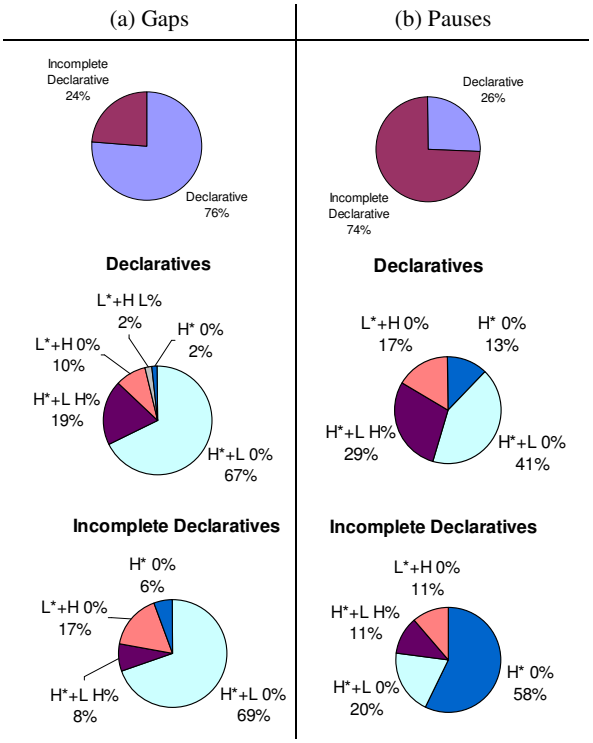



Figure 4. *The distribution of pitch patterns in Declaratives and Incomplete declaratives in the Gaps (a) and Pauses (b) subsets.*

### 4.3. Intonation and communicative type annotation in classification experiments

In order to investigate the combined discriminative power of functional and intonation labels (derived from speech-chunks immediately preceding pause and gap silent intervals) for differentiating pauses and gaps we carry out a speaker independent classification experiment. For this we utilise a support vector machine (SVM) based classifier with a radial basis function kernel. As input features we use the manually obtained annotation labels, separated into multiple binary features (e.g., the feature for H*+L 0% would have all samples with this annotation label assigned the value of 1 and all others 0). Classification is carried out using a leave-one-speaker-out procedure where the data of a single speaker is held out solely for testing, with the remainder of the data used for training the SVM classifier. This procedure is repeated for all 12 speakers with classification error (%) retained each time.

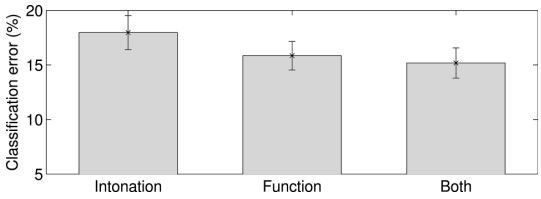



Figure 5. *The results of the support vector machine-based classification experiment. Shown are mean and standard error values.*

The results of this analysis are summarised in Figure 5 and are shown for intonation and functional label features separately and combined. The overall classification error is strikingly low for what is an extremely difficult discrimination problem. Using intonation labels alone one achieves a mean classification error of around 18%. The result provides a strong motivation to produce robust automatic characterisation of such intonation patterns.

Functional labels provide an even lower mean classification error (~15%). This also suggests that detection of functional labels would be beneficial for the prediction of pauses and gaps. However, deriving such information automatically would require the combination of an automatic speech recognition component as well as a subsequent text analytics procedure both of which are liable to introduce significant errors. The combination of the intonation and functional labels brings only a minor reduction in mean classification error and suggests a high level of redundancy between the two classes of labels.

## 5. Conclusions

In this paper we examined the distribution of communicative functions and pitch tunes in the 'human predictable' Pauses and Gaps subsets selected from the Dublin Institute of Technology Emotional Speech corpus. Overall, Gaps and Pauses subsets show differentiation both in terms of their communicative types and pitch patterns. Declaratives and Questions are commonly found in Gaps, whereas in Pauses it is mainly Hesitations and Incomplete declaratives. Gaps are mainly characterised by falling or rising pitch patterns (pitch movement), whereas in Pauses a large proportion of speech samples are realised with level pitch. Results suggest that including information on pitch patterns in the speech-chunks immediately preceding pause and gap silent intervals appears to enhance automatic discrimination of pauses and gaps. Our future work will exploit the findings of this study to examine other prosodic dimensions, voice quality and temporal characteristics, and their interaction with intonational features.

## 6. Acknowledgements

This research is supported by the Science Foundation Ireland Grant 09/IN.1/I2631 (FASTNET). We would like to thank Dr. Brian Vaughan (Dublin Institute of Technology) for providing the DIT Emotional Speech Corpus.

## 7. References


[1] J. Dalton, J. Kane, I. Yanushevskaya, A. Ní Chasaide, and C. Gobl, "GlóRí - the glotal research instrument," Speech prosody 2014, Dublin, Ireland, [accepted].

[2] J. Kane, I. Yanushevskaya, C. De Looze, B. Vaughan, and A. Ní Chasaide, "Analysing the prosodic characteristics of speech-chunks preceding silences in task-based interactions," Interspeech 2014, Singapore, [submitted].

[3] C. De Looze, I. Yanushevskaya, J. Kane, and A. Ní Chasaide, "Pitch range declination and reset in turn-taking organisation," Speech Prosody 2014, Dublin, Ireland, [accepted].

[4] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language,* vol. 50, 1974.

[5] M. Heldner, J. Edlund, K. Laskowski, and A. Pelcé, "Prosodic features in the vicinity of pauses, gaps and overlaps," in *Nordic Prosody. Proceedings of the Xth Conference*, M. Vainio, R. Aulanko, and O. Aaltonen, Eds., ed Berlin: Peter Lang, 2009, pp. 95-106.

[6] A. Rochet-Capellan and S. Fuchs, "The interplay of linguistic structure and breathing in German spontaneous speech," presented at the Interspeech 2013, Lyon, France, 2013.

[7] A. Kendon, "Some functions of gsze direction in social interaction," *Acta Psychologica,* vol. 26, pp. 22-63, 1967.

[8] S. Duncan, "Some signals and rules for taking speaking turns in conversations," *Journal of Personality and Social Psychology,* vol. 23, pp. 283-292, 1972.

[9] J. P. De Ruiter, H. Mitterer, and N. J. Enfield, "Projecting the end of a speaker's turn: a cognitive cornerstone of conversation," *Language,* vol. 82, pp. 515-535, 2006.

[10] A. Gravano and J. Hirshberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech and Language,* vol. 25, pp. 601-634, 2011.

[11] D. Schlangen, "From reaction to prediction: experiments with computational models of turn-taking," presented at the Interspeech 2006, Pittsburg, Pennsylvania, 2006.

[12] L. Ferrer, E. Shriberg, and A. Stolcke, "A prosody-based approach to end-of-utterance detection that does not require speech recognition," presented at the ICASSP, Hong Kong, China, 2003.

[13] J. Edlund and M. Heldner, "Exploring prosody in interaction control," *Phonetica,* vol. 62, pp. 215-226, 2005.

[14] B. Oreström, *Turn-Taking in English Conversation*: Krieger Publishing Company, 1983.

[15] J. Local and J. Kelly, "Projection and 'silences': notes on phonetic and conversational structure," *Human Studies,* vol. 9, pp. 185-204, 1986.

[16] K. Kohler, "Prosodic boundary signals in German," *Phonetica,* vol. 40, pp. 89-134, 1983.

[17] J. Caspers, "Local speech melody as a limiting factor in the turn-taking system in Dutch," *Journal of Phonetics,* vol. 31, pp. 251-276, 2003.

[18] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den, "An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs," *Language and Speech,* vol. 41, pp. 295-321, 1998.

[19] J. Fon, K. Johnson, and S. Chen, "Durational patterning at syntactic and discourse boundaries in Mandarin spontaneous speech," *Language and Speech,* vol. 54, pp. 5-32, 2011.

[20] J. Edlund, M. Heldner, and J. Gustafson, "Utterance segmentation and turn-taking in spoken dialogue systems," in *Computer Studies in Language and Speech*, ed Frankfurt am Main: Peter Lang, 2005, pp. 576–587.

[21] A. Gravano and J. Hirshberg, "Turn-yielding cues in task-oriented dialogue," presented at the SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue, London, UK, 2009.

[22] H. Friedberg, "Turn-taking cues in a human tutoring corpus," presented at the Annual meeting of the Association for Computational Linguistics, Portland, USA, 2011.

[23] R. Ogden, "Turn transition, creak and glottal stop in Finnish talk-in-interaction," *Journal of the International Phonetic Association,* vol. 31, pp. 139-152, 2001.

[24] M. Zellers, "Pitch and lengthening as cues to turn transition in Swedish," presented at the Interspeech 2013, Lyon, France, 2013.

[25] J. Local, J. Kelly, and W. H. G. Wells, "Towards a phonology of conversation: turn-taking in Tyneside English," *Journal of Linguistics,* vol. 22, pp. 411-437, 1986.

[26] B. Vaughan, "Naturalistic emotional speech corpora with large scale emotional dimension ratings. PhD thesis," Dublin Institute of Technology, 2011.

[27] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics,* vol. 38, pp. 555-568, 2010.

[28] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters,* vol. 6, pp. 1-3, 1999.

[29] A. F. Hayes and K. Krippendorff, "Answering the call for a standard reliability measure for coding data," *Communication Methods and Measures,* vol. 1, pp. 77-89, 2007.

[30] H. Bunt, J. Alexandersson, J.-W. Choe, A. C. Fang, K. Hasida, V. Petukhova*, et al.*, "ISO 24617-2: A semantically-based standard for dialogue annotation," presented at the LREC 2012, Istanbul, Turkey, 2012.

[31] H. Bunt, "Dimensions in dialogue annotation," presented at the LREC 2006, Genoa, Italy, 2006.

[32] C. Soria and V. Pirrelli, "A recognition-based meta-scheme for dialogue act annotation," presented at the Workshop Towards Standards and Tools for Discourse Tagging, Somerset, New Jersey, 1999.

[33] E. Grabe, B. Post, and F. Nolan, "Modelling intonational variation in English: the IViE system," presented at the Prosody 2000, Kraków, Poland, 2001.