*INTERSPEECH 2016*
September 8–12, 2016, San Francisco, USA

# Predicting User Satisfaction from Turn-Taking in Spoken Conversations

*Shammur Absar Chowdhury, Evgeny A. Stepanov, Giuseppe Riccardi*

Signals and Interactive Systems Lab
Department of Information Engineering and Computer Science
University of Trento, Italy
sachowdhury@disi.unitn.it, {evgeny.stepanov, giuseppe.riccardi}@unitn.it

## Abstract

User satisfaction is an important aspect of the user experience while interacting with objects, systems or people. Traditionally user satisfaction is evaluated a-posteriori via spoken or written questionnaires or interviews. In automatic behavioral analysis we aim at measuring the user emotional states and its descriptions as they unfold during the interaction. In our approach, *user satisfaction* is modeled as the final state of a sequence of emotional states and given ternary values `positive, negative, neutral`. In this paper, we investigate the discriminating power of turn-taking in predicting user satisfaction in spoken conversations. Turn-taking is used for discourse organization of a conversation by means of explicit phrasing, intonation, and pausing. In this paper, we train different characterization of turn-taking, such as competitiveness of the speech overlaps. To extract turn-taking features we design a turn segmentation and labeling system that incorporates lexical and acoustic information. Given a human-human spoken dialog, our system automatically infers any of the three values of the state of the user satisfaction. We evaluate the classification system on real-life call-center human-human dialogs. The comparative performance analysis shows that the contribution of the turn-taking features outperforms both prosodic and lexical features.

**Index Terms**: Spoken Conversation, Human-Human Interaction, Turn-Taking Structure, Overlap Discourse

## 1. Introduction

A satisfying communication plays an important role in social interaction such as multiparty and dyadic conversations in call-center, doctor-patient, and student-teacher scenarios. Over the years, user satisfaction has been evaluated using spoken or written questionnaires and interviews. In such an evaluation, users are usually asked to fill up questionnaires or rate certain aspects of a conversation that address users' satisfaction, as reported in [1]. User satisfaction has been addressed in other research fields as well – consumer satisfaction with products [2] and Spoken Dialog Systems (SDS) such as problem-solving [3] and tutoring [4]. In SDS, user satisfaction is used as one of the metrics to assess the quality of a dialog system [5, 6]. Thus, the increasing importance of user experience as a quality assessment demands a computational model for observed user satisfaction rather than self-reported measure.

In a natural conversation, parallel to the exchange of information, there is also a flow of speakers' emotional states, unfolding with or without any intent. A sequence of emotional states manifested during a conversation is a strong cue for predicting user experience. The goal of this paper is to exploit these sequences of emotional states, specifically the final state, to model user satisfaction. For the automatic prediction of the user satisfaction, the final emotional states are categorized into three labels as **Positive** (Pos), **Negative** (Neg), and **Neutral** (Neu). We investigate how the organizational structure of a conversation, such as turn-taking, contributes to the prediction of user satisfaction along with other more common levels of conversation description such as lexical and prosodic.

Turn-taking is a remarkable phenomenon that is fundamental for human communication [7]. Over decades the intriguing cues of turn-taking attracted researchers from conversational analysis, linguistics, psycholinguistics, and speech. One of the first studies on turn-taking was conducted by [8], where turn-taking is defined as a way to signal and perceive cues for Transition Relevance Place (TRP). The authors also suggest that the transition from the current speaker to the next should occur very frequently with minimum gap or overlap in speech. In [8, 9], overlaps have been considered as a violation of the fundamental rule, but the authors in [10] suggest that about 40% of all between-speaker intervals are overlaps. It has been proposed that speech overlaps relate to the dominance, aggression, competitiveness or cooperativeness towards the other speaker [11, 12, 13]. Other relevant studies include overlap detection [14, 15] (including word-level as overlap vs. clean-speech [16]), interruption detection [17], and studies on types of turn-taking and their correlation with speakers' turn-taking behavior [7].

Considering the literature on overlaps and turn-taking in spoken conversations, competitiveness and non-competitiveness of the speaker turns did not receive enough attention. Among the few, [18] demonstrate the importance of the onset position of the overlap along with the temporal features. On the other hand, in [19], the author argue that overlap is better described by the phonetic design rather than its precise location; which is later supported by [20, 21].

Previous work on incorporating turn-taking with social signals have mainly focused on group dynamics or task-oriented dialogs, like modeling participant's affects from turn-taking with post-meeting ratings [22] or studies about participant's involvement or interest [23, 24].

To the best of our knowledge, turn-taking has not been utilized for predicting user satisfaction as emotional manifestation. Hence, in this paper, we focus on turn-taking features for predicting user satisfaction; to achieve this goal we are:

- modeling the state of the user satisfaction in terms of the final emotional manifestation of the customer.
- automatically predicting the state of the user satisfaction as



http://dx.doi.org/10.21437/Interspeech.2016-859

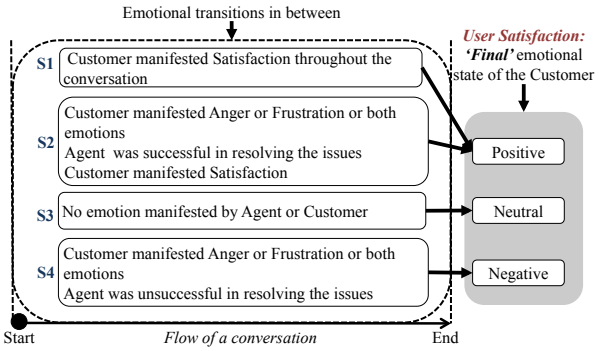


Figure 1: *Different scenerios of emotional manifestation with associated class labels representing user satisfaction.*

Pos, Neg, Neu, using the lexical, prosodic and turn-taking feature sets.

- designing a turn segmentation and labeling system by utilizing automatic transcriptions and acoustic features, to extract turn-taking features.
- comparatively evaluating and analyzing the turn-taking features to understand their discriminative power.

For the study, we analyzed a large dataset of Italian call-center spoken conversations where customers and agents are engaged in problem-solving tasks.

The paper is organized as follows. An overview of corpus along with dataset preparation is given in Section 2. Followed by details of the system framework, extracted features and classification experiments in Section 3. Section 4 presents the results and analysis of the observations. Conclusions are provided in Section 5.

## 2. Data Description

In this paper, we consider a corpus of 1894 call-center conversations [25], collected over the course of six-months (210 hours of speech, with an average length of 406 seconds per conversation). The conversations were recorded on two separate channels with 16 bits and 8kHz sampling rate.

The corpus was annotated for basic and complex emotions following the *modal model* of emotions developed by Gross [26, 27]. The model emphasizes the attentional and appraisal acts underlying the emotion-arousing process. For the annotation, the considered basic emotion was *anger*; and the complex social emotions were *satisfaction, dissatisfaction, frustration* and *empathy*. Empathy was annotated for the agent channel only; the rest of emotions for the customer channel. The inter-annotator agreement of the annotation has kappa = 0.74 (additional details of the annotation process can be found in [28]).

A subset of 739 conversations ($\approx$ 86 hours) was selected such that conversations annotated with customer emotion has also been annotated with empathy in the agent channel.

With respect to the annotation, the final manifested emotional state can be satisfaction, anger or frustration, or there might be no emotional manifestation. As shown in Figure 1, we define three labels for modeling *user satisfaction* concerning the final emotional state in the conversations. *Positive*, **Pos** is used for the conversations where the final emotional manifestation of the customer is satisfaction. Satisfaction may be the only manifested emotion in the customer channel ($S1$) or it may come as a results of a change from anger or frustration

Table 1: *Train, Dev and Test set split and their distribution.*

| Sets | Pos (%) | Neg (%) | Neu (%) | Total(%) |
|---|---|---|---|---|
| **Train** | 205 (34.0%) | 198 (32.84%) | 200 (33.17%) | 603 (100%) |
| **Dev** | 21 (30.43%) | 22 (31.88%) | 26 (37.68%) | 69 (100%) |
| **Test** | 19 (28.36%) | 25 (37.31%) | 23(34.33%) | 67 (100%) |

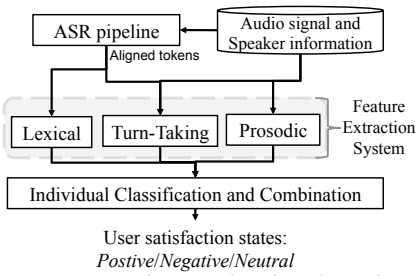


Figure 2: *Computational system for classifying the state of user satisfaction.*

due to agent's manifestation of empathy ($S2$); thus, yielding a sequence Customer: Anger/Frustration $\rightarrow$ Agent: Empathy $\rightarrow$ Customer: Satisfaction. *Negative*, **Neg** is used for the conversations where the final emotional manifestation of the customer is either anger, frustration or both ($S4$). The conversations that do not have any emotional manifestations are labeled as *Neutral*, **Neu** ($S3$). The split of the data into training, development and test sets are given in Table 1.

## 3. System Framework

In Figure 2, we present a pipeline for predicting the state of the user satisfaction, which takes an audio and speaker information of a conversation as an input. The audio signals are then passed through Automatic Speech Recognition (ASR) pipeline, which consists of a speech vs. non-speech detector and domain-specific ASR. Each detected speech segment is passed to the ASR [29]. The time aligned output of the ASR along with audio signal is then used to extract turn-taking, lexical and prosodic features.

The individual feature sets – lexical, prosodic, and turn-taking – are then used to train and evaluate classifiers. Additionally, we perform feature-level and decision-level fusion. For decision-level fusion, we are using weighted majority voting, where the weight of each classifier is the overall F1 of the system on dev set. Moreover, to understand the discriminative characteristics of the turn-taking features, they are analyzed using logistic regression model.

### 3.1. Feature Extraction

#### *3.1.1. Turn-Taking Features*

The Turn-Taking Feature Extraction System, described in Figure 3, consists of a *turn segmentation and labeling system* and the *feature generation* step. The system uses lexical and acoustic information to extract the features. The pipeline uses the time aligned ASR output as tokens to create Inter-Pausal Units (IPUs) for each input channel. IPUs are defined as the consecutive tokens with no less that 50 ms gaps in between. Using the time information of inter-IPUs and intra-IPUs, we defined **steady conversation segments** where each segments maintain a steady timeline in both interlocuters channel. The labels of each segment are then defined by a set of rules. Labels of the segments are as follows:

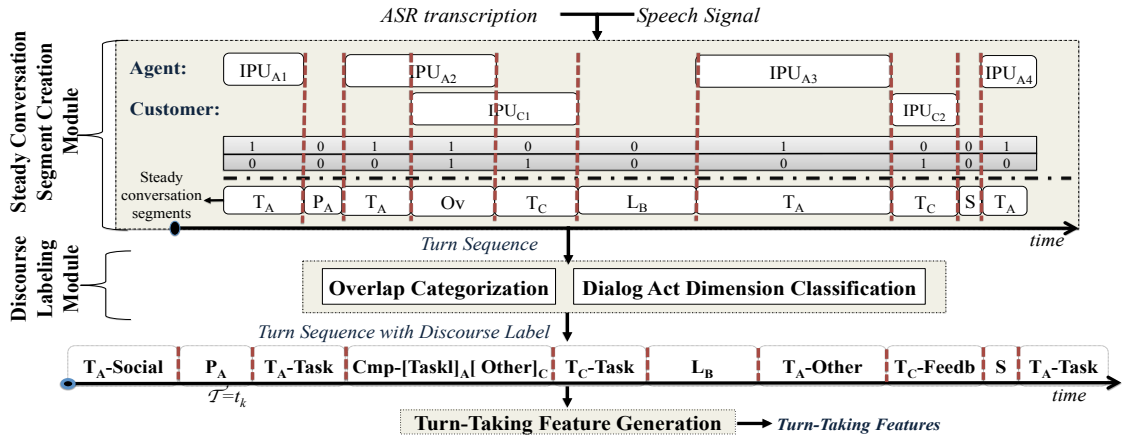


Figure 3: *Schematic diagram of automated Turn-Taking Feature Extraction System with speech signal and asr transcription as input.* $\boldsymbol{T_A}$, $\boldsymbol{T_C}$, $\boldsymbol{P_A}$*: agent and customer's turn and Pause,* $\boldsymbol{Ov}$*:overlaps,* $\boldsymbol{L_B}$*:Lapse between speakers,* $\boldsymbol{S}$*:Smooth switch,* $\boldsymbol{T_{A/C} - DA}$*:* $T_{A/C}$ *with* $DA$*, Dialog Act dimension, where* $DA \in \{Social, Task, Feedb, Other\}$*,* ***Cmp****: Competitive overlap.*

- Turn $(T)$: Maximal sequences of IPUs where one single speaker has the floor, and none of IPUs from the interlocutor are present [30]. $T_A$ and $T_C$ represent agent and customer's turns respectively.
- Pause $(P)$: Gaps between the turns of the same speaker with no less than 0.5 sec. $P_A$ and $P_C$ represent agent and customer's pauses respectively.
- Overlaps $(Ov)$: Overlapping turns between the two interlocutors.
- Lapse between speakers $(L_B)$: Floor Switches between the speakers with a silence duration of 2 sec or more.
- Lapse within speaker $(L_W)$: Gaps between a speakers' turns with a silence duration of 2 sec or more.
- Switch $(S)$: Floor Switches between the speakers with silence less than 2 secs or with overlapping frames not more than 20 ms.

The generated turn sequences along with the audio signals are then passed to Discourse Labeling Module (DLM) followed by the Turn-Taking Feature Generation module for extracting turn-taking features.

**Discourse Labeling Module:** The DLM module includes Overlap Categorization and Dialog Act Dimension Classification systems as described below.

*Overlap Categorization*: The automatic overlap labeling includes **Competitive (Cmp)** and **Non-Competitive (Ncm)** categories. In Cmp scenario, the intervening speaker starts prior to the completion of the current speaker and both the speakers perceive the overlap as problematic and display interest in the turn for themselves. In Ncm scenario, the intervening speaker starts at the middle of an ongoing turn with no evidence for the intent to grab the turn.

To automatically label these two categories of overlaps we use an in-domain overlap categorization model [12]. The model was trained using acoustic features with the left and right context of 0.2 and 0.3 seconds of speech. The overall F-measure of the system using acoustic features is 64.36% on the test set as reported in [12].

*Dialog Act Dimension Classification*: To get an overview of the function of each turn in the conversation, we use an in-house developed *dialog act segmenter* and *dialog act dimension classifier* [31]. The labels of output turns are the dimensions of the dialog acts from DiaML ISO specification [32] including dimensions such as: Task (e.g., question, instruct, suggest), Social (e.g., greeting, apology), TimeManagement and Feedback (e.g., stalling, positive-negative feedback), Others or None. The overall F-measure of the system, using bag-of-word features, is 72% (in-domain test set) and 60% (out-of-domain test set).

**Turn-Taking Feature Generation:** The turn-taking features are generated using the turn sequence output from the DLM module (see Figure 3). To understand the impact of overlaps – Cmp vs. Ncm, silence and other predictability factors of turn-taking structure are extracted as turn-taking features at both conversation and individual speaker levels. A brief description of extracted features are as follows:

- Participation equality [33], $P_{eq}$:

$$P_{eq} = 1 - (\frac{\sum_i^N (T_i - T)^2 / T}{E}) \quad (1)$$

  where $T$ is the average speech duration of the speakers. $T_i$ is the total speech duration for each speaker. $E$ represents the total speech duration. $N = 2$, represents two speakers as agent and customer.
- Turn-taking Freedom, as defined in [22], $F_{cond}$:

$$F_{cond} = 1 - \frac{H_{max}(Y|X) - H(Y|X)}{H_{max}(Y|X)} \quad (2)$$

  where we calculate $H(Y|X)$, the conditional entropy of speaker $Y$ being the next speaker after $X$ begins the turn, $H_{max}(Y|X)$ being the maximal possible value for this. $W = \{agent, customer\}$, $X \in W$, $Y \in W$ and $X \neq Y$. The value of $F_{cond}$ is between 0 and 1, where 0 represents a strict turn-taking.
- Percentage of overlaps.
- Percentage of Cmp and Ncm on total overlap duration.
- Percentage of agent's and customer's speech
- Median duration of $T_A$, $T_C$, $P_A$, $P_C$, Cmp, Ncm, $L_W$ and $L_B$.
- Probability of speaker $X$'s turn after a Cmp: $P(X|Cmp)$ or Ncm: $P(X|Ncm)$.
- Probability of Cmp after speaker X's turn: $P(Cmp|X)$ or Ncm after speaker X's turn : $P(Ncm|X)$.
- Rates of each dialog act dimension with respect to speaker's speech duration.

Table 2: *Classification results for predicting user satisfaction state. Feat.Comb: Feature-level combination, D.Fuse: Decision level fusion, Oracle-D.Fuse: Oracle of D.Fuse. Reported value is F1 measure on the test set.*

| Experiments | Pos | Neg | Neu | Overall |
|---|---|---|---|---|
| **Random-Baseline** | 0.24 | 0.30 | 0.27 | 0.27 |
| **Lexical** | 0.44 | 0.58 | 0.35 | 0.48 |
| **Prosodic** | 0.33 | 0.32 | 0.52 | 0.40 |
| **Turn-Taking** | 0.61 | 0.57 | 0.62 | 0.61 |
| **Feat.Comb** | 0.49 | 0.57 | 0.55 | 0.54 |
| **D.Fuse** | 0.57 | 0.57 | 0.60 | 0.59 |
| **Oracle-D.Fuse** | 0.90 | 0.86 | 0.80 | 0.85 |

#### *3.1.2. Prosodic Features*

We extracted prosodic features using openSMILE [34] with the frame size of 25 ms and a frame step of 10 ms. These low-level features such as pitch, loudness, and voice-probability together with their derivatives are then projected onto 24 statistical functionals such as mean and range among others. More details of these features are in [35].

We extract the prosodic features for agent and customer channels separately, then linearly merge them to design an equal sized vector for each conversation. Let $\boldsymbol{A_{s1}} = \{\boldsymbol{A_1}, \boldsymbol{A_2}, ..., \boldsymbol{A_m}\}$ and $\boldsymbol{C_{s2}} = \{\boldsymbol{C_1}, \boldsymbol{C_2}, ..., \boldsymbol{C_m}\}$ denote agent and customer channels' feature vectors respectively. The combined feature vector is $\boldsymbol{P_c} = \{\boldsymbol{A_1}, \boldsymbol{A_2}, ..., \boldsymbol{A_m}, \boldsymbol{C_1}, \boldsymbol{C_2}, ..., \boldsymbol{C_m}\}$ with $P_c \in R^{m+m}$.

#### *3.1.3. Lexical Features*

Lexical features are extracted from automatic transcriptions for the whole conversation from the ASR pipeline. The features are then transformed into a bag-of-words (vector space model) [36], to represent the words as numeric features. For this study, we extracted trigram features, to use the contextual benefit of n-grams. The frequencies in the feature vectors were then transformed into tf-idf values - the product of the logarithmic term frequency (tf) and inverse document frequency (idf).

#### *3.1.4. Feature Combination*

For this study, we also analyze the combined contribution of the feature sets. As shown in Figure 2, after extracting turn-taking, prosodic and lexical features we merge the feature vectors into a single vector and then use that for classification.

### 3.2. Classification and Evaluation

A Sequential Minimal Optimization (SMO), a support vector machine implementation of weka [37], is used to train the classifiers with feature values normalized within $[0, 1]$ interval. Due to the difference between the dimensionality of the feature vectors, we experiment with different kernels such as linear and RBF of SVM on the dev set. As for the evaluation, we report F-measure ($F1$) for individual classes, along with macro-averaged F-measure.

## 4. Results and Discussion

In Table 2 we present the results for predicting the state of user satisfaction in terms of *Pos*, *Neg* and *Neu*, using individual feature sets and their combination and decision level fusion. For comparison, a random baseline is calculated by randomly generating class labels based on prior class distribution.

It is observed that all the systems have higher performance than the baseline. Regarding overall system **F1**, the turn-taking features outperform all other systems. As for individual classes, turn-taking is noticed to be the best discriminator for Pos and Neu classes and has 1% F1 less in Neg class compared to the lexical feature set. This indicates the potential of lexical features to predict for Neg state of user satisfaction.

It is important to note that we have used the linear kernel of SVM for all the experiments except for turn-taking feature set, for which we used the RBF kernel, tuned on the dev set. The F1 of turn-taking features with linear kernel ($Tt - L$) and an optimized penalty parameter $C = 0.4$ are: Pos: 0.55, Neg: 0.52, Neu: 0.63 and Overall: 0.58. Even with linear kernel the turn-taking feature set exceeds the lexical and prosodic features by 10% and 18%, respectively.

Using feature combination (*Feat.Comb*), we have 6% and 14% improvement over lexical and prosodic feature sets but not over turn-taking feature set. One possible reason could be the fact that these feature sets vary in terms of dimensionality and their representations (dense *vs* sparse). The vector size for turn-taking feature is 34, which is very small compare to prosodic and lexical feature sets. The performance of the individual system is reflected in decision fusion result and the upper bound of decision fusion is shown by Oracle results in Table 2.

We use multilevel logistic regression [38], to understand the impact of turn-taking feature for predicting each state of user satisfaction. The result shows a significant positive effect on the presence of non-competitive overlaps and use of social turns by customers in Pos class, while the median duration of $T_A$ has a negative effect. That is, the customer tends to be more satisfied when there is an increase of feedback and social turns flow rather than agent taking long turns. Similarly, the use of the time-management/feedback DA turns decrease the likelihood of the conversation to be Neg significantly, whereas the likelihood of Neg class increases when the percentage of competitive overlaps along with the use of DA-Other by agent increases. In [39], the authors reported that the automatic feature "BargeIns" were highly correlated with user satisfaction, which also supports our findings with Neg class.

## 5. Conclusions

In this study, we investigate the use of turn-taking in predicting user satisfaction in spoken conversations. We model user satisfaction as the final emotional manifestation of a conversation, which can be either positive, negative or neutral. We extract turn-taking features by designing a turn segmentation and labeling system. We compare turn-taking features with lexical, prosodic feature sets along with feature level combination and decision level fusion. We observe that turn-taking features outperform all other systems. The analysis of turn-taking features suggests that the use of non-competitive turns and social dialog acts increase the chance of a positive user experience, whereas competitive turns tend to decrease the chance of positive experience.

## 6. Acknowledgments

The research leading to these results has received funding from the European Union - Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 610916- SENSEI.

# 7. References


[1] J. R. Hackman and N. Vidmar, "Effects of size and task type on group performance and member reactions," *Sociometry*, pp. 37–54, 1970.

[2] R. L. Oliver, *Satisfaction: A behavioral perspective on the consumer*. Routledge, 2014.

[3] E. Shriberg, E. Wade, and P. Price, "Human-machine problem solving using spoken language systems (sls): Factors affecting performance and user satisfaction," in *Proc. of the workshop on Speech and Natural Language*. ACL, 1992, pp. 49–54.

[4] K. Forbes-Riley and D. J. Litman, "Modelling user satisfaction and student learning in a spoken dialogue tutoring system with generic, tutoring, and user affect parameters," in *Proc. of the main conference on Human Language Technology Conference of the NACL*. ACL, 2006, pp. 264–271.

[5] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella, "Paradise: A framework for evaluating spoken dialogue agents," in *Proc. of the eighth conference of EACL*. ACL, 1997, pp. 271–280.

[6] K.-P. Engelbrech, F. Gödde, F. Hartard, H. Ketabdar, and S. Möller, "Modeling user satisfaction with hidden markov model," in *Proc. of the SIGDIAL 2009 conference*. ACL, 2009, pp. 170–177.

[7] Š. Beňuš, A. Gravano, and J. Hirschberg, "Pragmatic aspects of temporal accommodation in turn-taking," *Journal of Pragmatics*, vol. 43, no. 12, pp. 3001–3027, 2011.

[8] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, pp. 696–735, 1974.

[9] S. Duncan, "Some signals and rules for taking speaking turns in conversations." *Journal of personality and social psychology*, vol. 23, no. 2, p. 283, 1972.

[10] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, 2010.

[11] C. West, "Against our will: Male interruptions of females in cross-sex conversation*," *Annals of the New York Academy of Sciences*, vol. 327, no. 1, pp. 81–96, 1979.

[12] S. A. Chowdhury, M. Danieli, and G. Riccardi, "The role of speakers and context in classifying competition in overlapping speech," in *Proc. of INTERSPEECH*, 2015.

[13] J. A. Goldberg, "Interrupting the discourse on interruptions: An analysis in terms of relationally neutral, power-and rapport-oriented acts," *Journal of Pragmatics*, vol. 14, no. 6, pp. 883–903, 1990.

[14] S. N. Wrigley, G. J. Brown, V. Wan, and S. Renals, "Speech and crosstalk detection in multichannel audio," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 84–91, 2005.

[15] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings," in *Proc. of ICASSP*. IEEE, 2008, pp. 4353–4356.

[16] E. Shriberg, A. Stolcke, and D. Baron, "Can prosody aid the automatic processing of multi-party meetings? evidence from predicting punctuation, dis uencies, and overlapping speech," in *ISCA Tutorial and Research Workshop (ITRW) on Prosody in Speech Recognition and Understanding*, 2001.

[17] C.-C. Lee and S. Narayanan, "Predicting interruptions in dyadic spoken interactions," in *Proc. of ICASSP*. IEEE, 2010, pp. 5250–5253.

[18] G. Jefferson, *Two explorations of the organization of overlapping talk in conversation*. Tilburg University, Department of Language and Literature, 1982.

[19] P. French and J. Local, "Turn-competitive incomings," *Journal of Pragmatics*, vol. 7, no. 1, pp. 17–38, 1983.

[20] B. Wells and S. Macfarlane, "Prosody as an interactional resource: Turn-projection and overlap," *Language and Speech*, vol. 41, no. 3-4, pp. 265–294, 1998.

[21] B. Hammarberg, B. Fritzell, J. Gaufin, J. Sundberg, and L. Wedin, "Perceptual and acoustic correlates of abnormal voice qualities," *Acta oto-laryngologica*, vol. 90, no. 1-6, pp. 441–451, 1980.

[22] C. Lai, J. Carletta, and S. Renals, "Modelling participant affect in meetings with turn-taking features," in *Proc. Workshop of Affective Social Speech Signals*, 2013.

[23] B. Wrede and E. Shriberg, "Spotting" hot spots" in meetings: human judgments and prosodic cues." in *Proc. of INTERSPEECH*, 2003.

[24] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being bored? recognising natural interest by extensive audiovisual integration for real-life application," *Image and Vision Computing*, vol. 27, no. 12, pp. 1760–1774, 2009.

[25] M. Danieli, G. Riccardi, and F. Alam, "Annotation of complex emotion in real-life dialogues," in *Proc. of 1st Italian Conf. on Computational Linguistics (CLiC-it) 2014*, R. Basili, A. Lenci, and B. Magnini, Eds., vol. 1, no. 122–127, 2014.

[26] J. J. Gross, "The emerging field of emotion regulation: An integrative review," *Review of General Psychology*, vol. 2, no. 3, p. 271, 1998.

[27] J. J. Gross and R. A. Thompson, "Emotion regulation: Conceptual foundations," *Handbook of Emotion Regulation*, vol. 3, p. 24, 2007.

[28] M. Danieli, G. Riccardi, and F. Alam, "Emotion unfolding and affective scenes: A case study in spoken conversations," in *Proc. of Emotion Representations and Modelling for Companion Systems (ERM4CT) 2015,*. ICMI, 2015.

[29] S. A. Chowdhury, G. Riccardi, and F. Alam, "Unsupervised recognition and clustering of speech overlaps in spoken conversations," in *Proc. of Workshop on Speech, Language and Audio in Multimedia*, 2014.

[30] Š. Beňuš, A. Gravano, and J. Hirschberg, "Pragmatic aspects of temporal accommodation in turn-taking," *Journal of Pragmatics*, vol. 43, no. 12, pp. 3001–3027, 2011.

[31] S. A. Chowdhury, E. A. Stepanov, and G. Riccardi, "Transfer of corpus-specific dialogue act annotation to iso standard: Is it worth it ?" in *Proc. of LREC*, 2016.

[32] H. Bunt, J. Alexandersson, J.-W. Choe, A. C. Fang, K. Hasida, V. Petukhova, A. Popescu-Belis, and D. R. Traum, "ISO 24617-2: A semantically-based standard for dialogue annotation." in *Proc. of LREC*, 2012, pp. 430–437.

[33] J. Carletta, S. Garrod, and H. Fraser-Krauss, "Placement of authority and communication patterns in workplace groups the consequences for innovation," *Small Group Research*, vol. 29, no. 5, pp. 531–559, 1998.

[34] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc. of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[35] S. A. Chowdhury, M. Danieli, and G. Riccardi, "Annotating and categorizing competition in overlap speech," in *Proc. of ICASSP*. IEEE, 2015.

[36] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Machine Learning: ECML-98*, ser. Lecture Notes in Computer Science, C. Nédellec and C. Rouveirol, Eds. Springer Berlin Heidelberg, 1998, vol. 1398, pp. 137–142.

[37] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[38] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," vol. 95, no. 1-2, pp. 161–205, 2005.

[39] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella, "Evaluating spoken dialogue agents with paradise: Two case studies," *Computer Speech & Language*, vol. 12, no. 4, pp. 317–347, 1998.