ANNOTATING AND CATEGORIZING COMPETITION IN OVERLAP SPEECH

Shammur Absar Chowdhury, Morena Danieli, Giuseppe Riccardi

Dept. of Information Engineering & Computer Science, Univ. of Trento, Trento, Italy

ABSTRACT

Overlapping speech is a common and relevant phenomenon in human conversations, reflecting many aspects of discourse dynamics. In this paper, we focus on the pragmatic role of overlaps in turn-in-progress, where it can be categorized as *competitive* or *non-competitive*. Previous studies on these two categories have mostly relied on controlled scenarios and small datasets. In our study, we focus on call center data, with customers and operators engaged in problem-solving tasks. We propose and evaluate an annotation scheme for these two overlap categories in the context of spontaneous and *in-vivo* human conversations. We analyze the distinctive predictive characteristics of a very large set of high-dimensional acoustic feature. We obtained a significant improvement in classification results as well as significant reduction in the feature set size.

Index Terms— Spoken Conversation, Automatic Classification, Overlapping Speech, Discourse

1. INTRODUCTION

Speech community has been investigating acoustic and temporal properties of overlapping speech for many years. The recent interest in understanding more about this phenomenon is shown with the aim of improving the quality and naturalness of spoken dialog systems. However the original, and still very relevant, interest was motivated by understanding the dynamics of human-human conversations (e.g., turn-taking). One of the first studies on speech overlap in [1] suggested that turn changes with overlap is a very rare case and occurs as a result of self-selection, projecting turn endings. Whereas, a recent study [2] suggests that overlap is in fact a frequent phenomenon and is much more than just a turn-taking signal. In everyday conversation, speech overlap phenomena are discourse resources that speakers use for accomplishing their communicative intentions. As noted in [3], *non-competitive* overlaps indicate a support for the current speaker to continue speaking, whereas *competitive* overlaps indicate an intention to break the flow of the conversation or to compete for the turns [4].

While most of the previous studies have focused on meeting corpora [5] or other small datasets, we concentrate on spoken conversations collected from a call center, with engaged users and real tasks. The aim of this study is to classify the *non-competitive* or *competitive* overlaps by analyzing the distinctive characteristics of different acoustic feature groups. Moreover, to achieve this, we need an operational model for the annotation of overlaps.

Our contribution in this study includes the design of a speech overlap annotation scheme and the automatic classification of competitive *vs* non-competitive overlapping segments from spoken conversation. While doing so, we analyzed different acoustic feature groups, their combination and an optimal subset by using feature selection.

This paper is organized as follows. An overview of previous studies of overlaps is given in Section 2, followed by the description of the obtained dataset and the annotation scheme in Section 3. In Section 4, we discuss the details of the classification experiments, results and analysis of our findings. Conclusions are provided in Section 6.

2. RELATED WORK

There have been very few studies on the speakers' competitive and non-competitive turns compared to the other research area dealing with overlaps and turn-taking in spoken conversations. Typical usage of features includes fundamental-frequency (f0) [6], energy [7], or their combination [4] to discriminate competitive and non-competitive overlaps.

In [5, 8], the authors suggest that f0 is the most relevant feature. Other features such as speech rate, cut-offs and repetition are also analyzed by researchers in [9] and observed that it is used by speakers to show competitiveness.

Findings in [10] show the onset-position of the overlap is an important feature along with some temporal features related to the position of overlaps. Whereas in [4], to describe a competitive overlap, author argued that the phonetic design plays an important role rather than its precise location. This claim is later supported by [3, 11]. It is also observed in [11] that competitive overlaps include high pitch and amplitude to grab the attention from the current speaker.

In [12], the authors found that duration is the most distinguishing feature while classifying competitive and non-competitive overlaps using decision tree. The authors in [12] and [13] state that non-competitive overlaps tend to be shorter and resolved soon after the second speaker has recognized the overlap, whereas competitive overlaps are persistent because speakers keep on speaking despite overlapping. In [14], the

Table 1. Data set description. Dur: duration. No. of Inst: number of instances

Set	No. of dialog	Dur	No. of Inst		Class Dist	
			Cmp	Ncm	Cmp	Ncm
Train	233	3 hrs 27 min	2467	6356	28.03%	71.97%
Test	20	29 mins	238	788	23.20%	76.80%

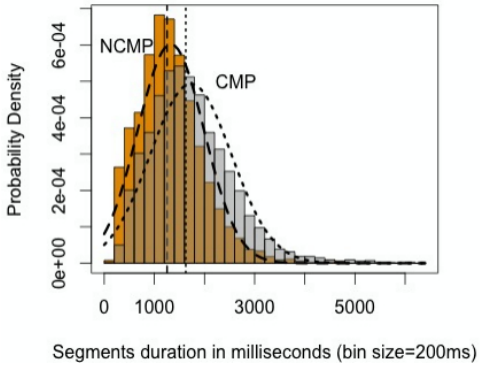

Fig. 1. Overlap Segments duration distributions; NCM - non-competitive overlap segments; CMP - competitive overlap segments

authors used higher-dimensional acoustic feature sets for categorizing overlaps using an unsupervised technique.

Other relevant research includes automatically classifying a word as clean-speech *or* overlap [15], detecting overlaps [16, 17, 18], interruptions [19], understanding the types of turn-taking and their correlation with speakers' turn-taking behavior [20] among others.

3. CORPUS AND ANNOTATION SCHEME

3.1. Corpus

The Italian human-human spoken conversations were sampled from large scale call centers conversations providing customer care support. The conversations were recorded over two separate channels at a sample rate of 8 kHz, 16bits and have an average duration of 395 seconds. It consists of 253 conversations with approximately 27 hours, from which we obtained 9858 overlaps segments, for a total duration of 3 hours and 56 minutes. For the experiments, we split our data into training and test sets. Details of the dataset is shown in Table 1. In Figure 1, the distribution of the segments containing overlaps is presented where the median is shown using the dotted and dashed vertical lines.

3.2. Annotation Scheme

The analysis of speech overlap was done by an expert psycholinguist who listened to a set of recorded calls, by applying a systematic direct observation protocol [21], and focused on overlapping speech segments. The observations allowed the psycholinguist to identify different kinds of overlapping speech segments, differing with respects to their pragmatic functions, speaker intentions and linguistic structure. For instance, most of the analyzed conversations showed that overlapping speech segments are co-occurring with greetings at the end of the phone conversation. The occurrences of speech overlap were characterized by significant variations of their prosodic profiles where some of them showed the intention of the intervening speaker to "grab the floor" of the conversation, i.e., to compete with the other speaker in view of controlling the turn taking structure of the dialog. One such case is the tendency of agents to interrupt the customer when they believe to have understood the customer's question while the latter insists on providing more information. Sometimes, however, the intention to "grab the floor" did not show a competitive attitude of the speaker. For example, several overlapping speech segments sound as being collaborative completions by the intervening speaker. Those occurrences could be classified as one out of several forms of back-channeling phenomena.

On the basis of this observational analysis, we designed the annotation guidelines for segmenting and annotating the speech overlaps with the competitive and non-competitive labels. The annotation guidelines include the following:

1. Each overlapping segment may contain more than one overlap instance of the same category. Instances may be separated from each other with a gap less than 40ms.
2. If a speaker thinks aloud during another speaker's turn that is considered an overlap instance.
3. Co-occurrences of "false start" by both the speakers are considered instances of speech overlap if and only if the segments contain complete words and the annotator can infer the speaker's intention on the basis of the perceived intonation of speech.
4. Annotators are asked to reject a conversation or ignore segments if they contain poor quality audio, unintelligible speech, background noise, human sounds like cough, sneezes and laughs.
5. The annotator's judgment includes the appraisal of the speakers' intention on the basis of supra-segmental variations including speech rhythm, accent and intonation along with peculiarities of the semantic content of the portion.

Using the above guidelines, the annotators were asked to classify the segments into one of the following two categories:

Competitive (Cmp): Scenarios where 1) the intervening speaker starts prior to the completion of the current speaker, 2) both the speakers display interest in the turn for themselves, and 3) speakers perceive the overlap as problematic.

Non-Competitive (Ncm): Scenarios where 1) another speaker starts in the middle of an ongoing turn, 2) both parties do not show any evidence for grabbing the turn for themselves, 3) speakers perceive the overlap as non-problematic and 4) speakers use it to signal the support for the current speaker's continuation of speech.

Table 2. Dialog excerpts from the annotated corpus. Speech overlaps: bold form between [and], Hesitations: (.), Rising intonation: ↗, Falling intonation: ↘.

Ncm
S1: è una piccola [**cosa però**] ↘ se (.)
S2: [**no signora ↘ ha**] fatto bene ↘
S1: it is a [***little thing***] ↘ *if (.)*
S2: [***no madam ↘ have***] *done well* ↘
Cmp
S1: perché questa [**è la vostra ultima**] che ho ↗↘
S2: [**no signora ↗ dal**] 31 marzo non è con noi ↗
S1: because this [***is the your latest***] *that have* ↗↘
S2: [***no madam ↗ from***] *march 31 you are not with us* ↗

Two expert annotators, Italian native speakers, performed the annotation task. As specified in the guidelines, they manually segmented the speech overlap occurrences and labeled each segment as competitive or non-competitive.

In Table 2, we report two examples of overlap segments with their English translation. The overlap segments are represented in bold form between square brackets and reported tone direction, based on IPA notation [22]. In the first example, the overlap speech segments of speaker S1 and S2 have a falling intonation: S1 hesitates and S2 intervenes for reassuring her. The opposite occurs in the second example: S1 speech has a rising-fall intonation, whereas the tone of S2 speech is constantly rising. S1 is surprised and overwhelmed by the sharp tone of S2.

3.3. Evaluation of the Annotation

To assess the reliability of the annotations we calculated inter-annotator agreement by using the kappa statistics [23, 24]. For calculating the agreement two annotators worked independently over a set of 28 spoken conversations randomly extracted from the call center corpus. The amount of spontaneous speech annotated for the inter-annotator agreement test was around 3 hours 17 minutes. The Kappa statistics is frequently used to assess the degree of agreement among any number of annotators by excluding the hypothetical probability that they agree by chance. By evaluating our data we reported kappa = 0.7033. Additionally, to quantify the inter-annotator agreement as human-performance in categorization of overlaps, a Positive (Specific) Agreement [25], identical to the widely used F-measure [26], was also used to obtain pairwise F-measure as an evaluation to the annotator agreement. In this case we obtained F1 = 85. The cases of disagreement were discussed in a consensus meeting by the annotators and the author of the guidelines. The most relevant disagreement between annotators concerned speech disfluencies, including false starts, repairs, and filled pauses. In most of the cases consensus was reached between the two annotators.

4. CLASSIFICATION EXPERIMENTS

4.1. Feature Extraction and Selection

One of the main focuses of this study is to understand the discriminative characteristics of acoustic features in categorizing competitive *vs* non-competitive overlaps. For this, we have extracted different groups of low-level features using openSMILE [27], motivated by their successful utilization in several paralinguistic tasks discussed in [28]. These sets of acoustic features were extracted with approximately 100 overlapping frames per second and with 25 milliseconds of window. The low-level features are extracted as a group-wise, presented in Table 3. For example, the *Prosody (P)* group includes pitch, loudness and voice-probability features.

These low-level features are then projected on 24 statistical functionals, which include range, absolute position of max and min, linear and quadratic regression coefficients and their corresponding approximation errors, moments-centroid, variance, standard deviation, skewness, kurtosis, zero crossing rate, peaks, mean peak distance, mean peak, geometric mean of non-zero values and number of non-zeros.

As mentioned in Section 3.1, the overlap segment components appear on the agent and the customer channel recordings. Therefore, we extracted the same number of features from channel-1, $CH1 = \{a_1, a_2, ..., a_m\}$ and channel-2, $CH2 = \{b_1, b_2, ..., b_m\}$. Then, merged the features from both channels to form a new feature vector, $X = \{a_1, a_2, ..., a_m, b_1, b_2, ..., b_m\}$. This procedure is applied for each feature group such as P, V, M, E, and S. Hence, the representation of each group is same as X.

Moreover, we formed a new feature set, $\boldsymbol{All} = P \cup V \cup M \cup E \cup S$, by merging the feature group, to understand their combined contribution. In addition, we applied automatic feature selection technique to each group, which leads to subset of features for each group, such as $P^{'}, V^{'}, M^{'}, E^{'}$ and $S^{'}$. After that, we designed an optimal feature subset, $\boldsymbol{FeatSub} = P^{'} \cup V^{'} \cup M^{'} \cup E^{'} \cup S^{'}$.

For the automatic feature selection we used Correlation-based Feature Selection (CFS) [29] and its implementation in Weka [30]. It ranks subsets containing features, which are highly correlated with the class labels and yet uncorrelated with each other using a best-first-search heuristic.

4.2. Classification and Evaluation

For the classification task, we used Sequential Minimal Optimization (SMO), a support vector machine implementation with its linear kernel and default parameters. To understand the relevance of each feature set for competitiveness and non-competitiveness binary classification task, we designed per-category classifier using SMO.

For the evaluation, there has not been any well-agreed metric for the task. Studies [12] used accuracy as an evaluation measure. It is evident that accuracy is not a good mea-

Table 3. Description of low-level acoustic features and grouped by type.

Prosody (P): 288 *2 = 576 features
Pitch (Fundamental frequency f0, f0-envelope), loudness, voice-probability.
Voice-Quality (V): 288 *2 = 576 features
Jitter, shimmer, logarithmic harmonics-to-noise ratio (logHNR)
MFCC (M): 936*2 = 1872 features
Mel-frequency cepstral coefficients (MFCC 0-12)
Energy (E): 72*2 = 144 features
Logarithmic signal energy from pcm frames
Spectral (S): 864*2 = 1728 features
Energy in spectral bands (0-250Hz, 0-650Hz, 250-650Hz, 1-4kHz), roll-off points (25%, 50%, 70%, 90%), centroid, flux, max-position and min-position.

sure for imbalanced class distribution [31], therefore, we considered to measure Precision (**P**), Recall (**R**) and (**F1**). As we want to evaluate our system considering both the classes, we computed macro-averaged P, R and F1. Statistical significance has been reported using McNemar's test, for the full feature set "All" with other results we present in Table 4.

4.3. Results and Discussion

The performances of different feature groups are reported in Table 4. A Naive-Bayes classifier has been designed using segment duration of overlap as a feature for the baseline results. The performance of the system using the feature set "All" gives a significant improvement of F1, 65.6 over the baseline result F1, 62.1. Following, each feature group results are reported to give an insight of how the group of features contributes to differentiate between the class labels.

The results indicating that spectral and prosody are the key distinguishing feature groups, giving a score of F1 68.8 and 67.8. It is also worth noticing that some feature groups contribute more on a specific class decision rather than overall. For instance, voice quality does not perform well in terms of overall system results, but when categorizing Ncm class, this feature group outperforms other feature groups, with F1 for Ncm of 86.9.

Most of the feature-category based classifiers outperforms the system trained on "All" features. Such data sparseness problem has been addressed by the feature selection. Using the optimal subset "FeatSub", with a reduction of 89.4% features, a F1 of 69.5 performance has been achieved, which is significantly better than the result with "All" set. The result of "FeatSub" is not statistically different from feature group. However, in terms of features and the dimension, "FeatSub" contains reduced and most distinctive features and further investigations is needed to understand its usefulness. The "FeatSub" contains approximately 79.34% spectral, mfcc and energy features, and 20.66% contains prosody and voice quality features. It is also noticed that the presence of Spectral feature group, containing features such as spectral flux, overshadows all other feature groups.

Table 4. Classification results on test set. Precision, Recall and F1 are macro-averaged. Dim. : feature dimension. *: significant change over full feature set with p<0.05. °: significant change over baseline result with p<0.05. All: Full feature set. VQ: Voice Quality. FeatSub: optimal feature subset, selected features from each group and then merged.

Features	Dim.	P(Avg)	R(Avg)	F1(Avg)
Baseline	1	64.4	59.9	62.1
Prosody (P)	576	67.7	68.1	67.8*
VQ (V)	576	67.8	60.2	63.8*
MFCC (M)	1872	66.5	68.4	67.4*
Energy (E)	144	67.4	67.5	67.5*
Spectral (S)	1728	68. 4	69.3	68.8*
All	4896	64.4	66.9	65.6°
FeatSub	518	69.1	70.0	**69.5***

The experimental results of our study reveals that high-dimensional low-level features projected onto statistical functionals followed by feature selection provides better insights to discriminate the type of overlaps.

5. CONCLUSIONS AND FURTHER WORK

This study illustrated automatic classification of competitiveness and non-competitiveness of overlap segments in real-world call-center data. It also introduced and evaluated an annotation scheme for the overlap categories. Different high-dimensional acoustic feature groups, their combination and an optimal subset by using feature selection were also discussed. We obtained a significant improvement in results using the model designed with the optimal feature subset compared to the full feature set. It is observed that spectral and prosody features play an important role for differentiating both classes, whereas, voice quality feature performs very well for defining non-competitive overlaps. Given the performances obtained, there is more room for improvement, which include the analysis of contextual and lexical features.

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] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, pp. 696–735, 1974.

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

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

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

[5] Emina Kurtić, Guy J Brown, and Bill Wells, "Resources for turn competition in overlapping talk," *Speech Communication*, vol. 55, no. 5, pp. 721–743, 2013.

[6] Emina Kurtic, Guy J Brown, Bill Wells, D Barth-Weingarten, D Dehé, and A Wichmann, "Fundamental frequency height as a resource for the management of overlap in talk-in-interaction," *Where Prosody Meets Pragmatics,. In: Studies in Pragmatics*, vol. 8, pp. 183–205, 2009.

[7] Chi-Chun Lee, Sungbok Lee, and Shrikanth S Narayanan, "An analysis of multimodal cues of interruption in dyadic spoken interactions.," in *Proc. of INTERSPEECH*, 2008, pp. 1678–1681.

[8] Khiet P Truong, "Classification of cooperative and competitive overlaps in speech using cues from the context, overlapper, and overlappee," in *Proc. of INTERSPEECH*, 2013, pp. 1404–1408.

[9] Emanuel A Schegloff, "Overlapping talk and the organization of turn-taking for conversation," *Language in society*, vol. 29, no. 01, pp. 1–63, 2000.

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

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

[12] Emina Kurtic, Guy J Brown, and Bill Wells, "Resources for turn competition in overlap in multi-party conversations: speech rate, pausing and duration.," in *Proc. of INTERSPEECH*, 2010, pp. 2550–2553.

[13] Gail Jefferson, "A sketch of some orderly aspects of overlap in natural conversation," *PRAGMATICS AND BEYOND NEW SERIES*, vol. 125, pp. 43–62, 2004.

[14] Shammur A. Chowdhury, Giuseppe Riccardi, and Firoj Alam, "Unsupervised recognition and clustering of speech overlaps in spoken conversations," in *Proc. of Workshop on Speech, Language and Audio in Multimedia - SLAM2014*, 2014.

[15] Elizabeth Shriberg, Andreas Stolcke, and Don 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.

[16] Stuart N Wrigley, Guy J Brown, Vincent Wan, and Steve Renals, "Speech and crosstalk detection in multichannel audio," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 84–91, 2005.

[17] Kofi Boakye, Beatriz Trueba-Hornero, Oriol Vinyals, and Gerald Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings," in *Proc. of ICASSP*. IEEE, 2008, pp. 4353–4356.

[18] Martin Zelenák and Javier Hernando, "The detection of overlapping speech with prosodic features for speaker diarization.," in *Proc. of INTERSPEECH*, 2011, pp. 1041–1044.

[19] Chi-Chun Lee and Shrikanth Narayanan, "Predicting interruptions in dyadic spoken interactions," in *Proc. of ICASSP*. IEEE, 2010, pp. 5250–5253.

[20] Štefan Beňuš, Agustín Gravano, and Julia Hirschberg, "Pragmatic aspects of temporal accommodation in turn-taking," *Journal of Pragmatics*, vol. 43, no. 12, pp. 3001–3027, 2011.

[21] K Anders Ericsson and Herbert Alexander Simon, *Protocol analysis*, MIT-press, 1984.

[22] International Phonetic Association, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*, Cambridge University Press, 1999.

[23] Cohen J., "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[24] Jean Carletta, "Assessing agreement on classification tasks: the kappa statistic," *Computational linguistics*, vol. 22, no. 2, pp. 249–254, 1996.

[25] Joseph L Fleiss, "Measuring agreement between two judges on the presence or absence of a trait," *Biometrics*, pp. 651–659, 1975.

[26] George Hripcsak and Adam S Rothschild, "Agreement, the f-measure, and reliability in information retrieval," *Journal of the American Medical Informatics Association*, vol. 12, no. 3, pp. 296–298, 2005.

[27] Florian Eyben, Felix Weninger, Florian Gross, and Björn 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.

[28] Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011.

[29] Mark A Hall, *Correlation-based feature selection for machine learning*, Ph.D. thesis, The University of Waikato, 1999.

[30] Ian H Witten and Eibe Frank, *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, 2005.

[31] Nathalie Japkowicz and Mohak Shah, *Evaluating Learning Algorithms*, Cambridge University Press, 2011.