A DEEP LEARNING APPROACH TO MODELING COMPETITIVENESS IN SPOKEN CONVERSATIONS

Shammur Absar Chowdhury, Giuseppe Riccardi

Signals and Interactive Systems Lab
Department of Information Engineering and Computer Science
University of Trento, Trento, Italy

ABSTRACT

The motivation behind the research on overlapping speech has always been dominated by the need to model human-machine interaction for dialog systems and conversation analysis. To have more complex insights of the interlocutors' intentions behind the interaction, we need to understand the type of overlaps. Overlapping speech signals the interlocutor's intention to grab the floor. This act could be a competitive *or* non-competitive act, which either signals a problem or indicates assistance in communication. In this paper, we present a Deep Learning approach to modeling competitiveness in overlapping speech using acoustic and lexical features and their combination. We compare a fully-connected feed-forward neural network to the Support Vector Machine (SVM) models on real call center human-human conversations. We have observed that feature combination with DNN (significantly) outperforms SVM models, both the individual feature baselines and the feature combination model by 4% and 2% respectively.

***Index Terms*—** Spoken Conversation, Overlapping Speech, Discourse, Context, Automatic Classification, DNN, SVM

1. INTRODUCTION

Traditionally, in the field of conversational analysis, overlaps have been considered as a violation of the fundamental rule of turn-taking, which suggests that one person speaks at a time [1, 2]. But in our daily social interaction, specifically in spontaneous conversations, overlapping speech is one of the most frequently occurring natural phenomena. It has been suggested that about 40% of all between-speaker intervals are overlapping speech [3]. Overlapping speech not only influences the organizational flow of a conversation, but also may reveal speakers' attitudes, dominance or aggression towards each other [4]. Further studies suggest that speakers' intentions, motivations or states of user-experience [5] behind the conversations can also reflect the use of overlaps in the interaction. However, not all the overlapping speech represent competitiveness intension. They are also utilized to indicate cooperativeness; for example, providing the other speaker the cues about the mutual understanding [6].

Over the years speech scientist studied overlaps to improve the quality of human-machine dialog systems and the systems for analysis of the human-human conversations. Overlaps are generally categorized as **Competitive (Cmp)** – *an attempt to grab the floor* – and **Non-Competitive (Ncm)** – *an attempt to assist the speaker to continue the current turn*. The act of a cooperative (non-competitive) overlap is one of the most frequent phenomena, which makes the task of classifying the overlap a challenging problem due to its unbalanced natural distribution. Another challenge is the modeling of the perception of the discourse of these overlaps from the overlappee side.

Previous studies have been conducted on discriminating the speakers' competitive and non-competitive turns using temporal cues, phonetic organization, and position of the overlaps [7, 8, 9]. In [10], the authors suggest that variations in prosodic profiles and repetitions used by speakers are a strong indication of the turn's competitiveness. Various features have been explored for the automatic categorization of the overlaps as competitive and non-competitive, such as hand motions and disfluencies [11], body movements from both speakers and contextual prosodic features from the overlapper [12], gaze, voice quality and contextual features –preceding and within overlaps [13]. In [14, 15], the authors have used higher-dimensional acoustic features and the context for characterizing competitive and non-competitive overlaps.

Until recently, most signal and information processing studies have focused on 'shallow' supervised machine learning algorithms such as Support Vector Machines (SVM), which use a shallow linear pattern separation model. Use of such architecture has been proved effective in solving many classification problems. However, in a case of a natural speech, such a shallow representation can be problematic. Natural way of understanding human conversation suggest the need for a deep architecture. Due to the advancement of high-performance computing over the last years, such as modern graphics processing unit (GPU) [16], neural networks, containing several hierarchical layers, have been widely applied to all sorts of problems in Speech and Natural Language Processing (NLP) and Computer Vision with the huge success [17, 18, 19]. The approach is termed as “deep learning or deep neural networks (DNN)”.

The goal of this study is to automatically categorize competitive *vs* non-competitive overlaps by exploiting many layers of the non-linear information processing for high-dimensional features. For modeling competitiveness, we focus on a Deep Learning approach where we use high-dimensional acoustic and lexical features along with their combination. We compare the performance of the ‘deep’ system to the SVM systems for individual feature sets and their combination.

The paper is organized as follows. In Section 2, we provide a brief overview of the data set. We discuss the experimental methodology in Section 3. Section 4 presents the results and the analysis of the experimental observations. Conclusions are provided in Section 5.

2. CORPUS DESCRIPTION

The data set used in this study consists of Italian human-human spoken conversations in the domain of customer care support. The corpus includes a total of 565 conversations with 62 hours of data with an average duration of 395 seconds. These conversations were recorded in two separate channels (for the agent and a customer) at a sample rate of $8\ kHz$, $16\ bits$.

The conversations were manually segmented for overlap boundaries using the audio signals and annotated for competitiveness and non-competitiveness. The complete annotation process is described in [14]; and the inter-annotator agreement (kappa) is 0.70.

The annotation guidelines define competitive (Cmp) scenarios in overlaps such that the intervening speaker (overlapper) starts speaking prior to the completion of the turn of the current speaker (overlappee) where both the interlocutors are interested in holding the turn for themselves, and the speakers perceive the overlap as problematic. In Non-Competitive (Ncm) scenarios, the overlapper starts in the middle of an ongoing turn with no evidence shown by both the speaker for the intention of grabbing the turn for themselves. In the latter case the overlapper uses the overlap to support the current speaker to continue the turn, and both the speakers perceive the overlap as non-problematic.

3. EXPERIMENTAL METHODOLOGY

Figure 1 depicts the pipeline for modeling and predicting the competitiveness in the overlapping speech. The pipeline shows the flow of both the training and the evaluation processes. In the data preparation phase (see the details in Section 3.1), audio signal, manual transcription and the manual overlap discourse annotation are used to prepare the training, development and test sets. In the feature extraction phase, acoustic and lexical features are extracted. Next, SVM and DNN models are trained and evaluated for each of the individual features sets and their combinations.

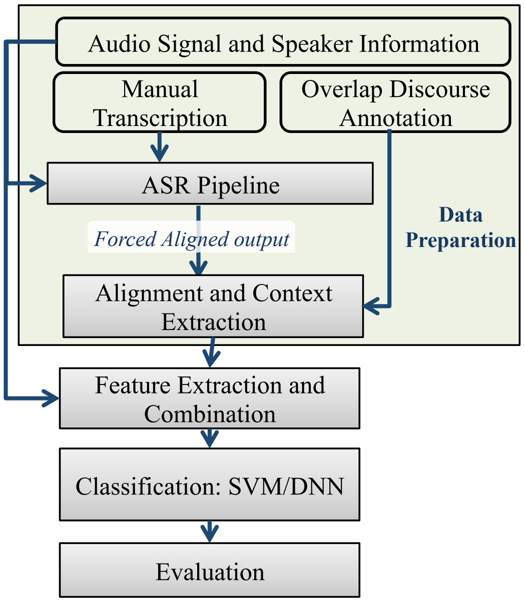
Audio Signal and Speaker Information
Manual Transcription
Overlap Discourse Annotation
ASR Pipeline
Forced Aligned output
Alignment and Context Extraction
Data Preparation
Feature Extraction and Combination
Classification: SVM/DNN
Evaluation

Fig. 1: System architecture for modeling competitiveness in overlapping speech.

3.1. Data Preparation

As shown in the Figure 1, the data preparation phase consists of several steps. The first step uses the audio signal and the speaker information along with the manual transcription and the manual overlap boundaries. The audio and manual transcriptions are passed to the Automatic Speech Recognition (ASR) [20] to obtain forced aligned tokens for each channel. In the *Alignment and Context Extraction* step, the tokens from the ASR output are aligned with the manual overlap boundaries, and the context of overlaps is extracted.

During the manual annotation, annotators first selected the overlap segment and then assigned a discourse label to it. In the manual annotation overlapping segments may consist of a combination of several short overlapping and non-overlapping segments. Moreover, an overlap boundary does not necessarily occur at token boundary. For our experiments, it is important to align the turn tokens to the overlap boundaries. This alignment takes place during the *Alignment and Context Extraction* step utilizing heuristics. For example, if the majority of a token falls into the overlap segment, it is assigned to it; and the start/end time of the segment is adjusted to the start/end time of the token. Then, for each overlap instance, the left and the right contexts are extracted using a window of $0.2s$ and $0.3s$, respectively. The threshold are chossen based on previous experimental results and motivated by [15].

Overall, the process yields $15,899$ overlap segments with a total duration of 5 hours and 8 minutes. Table 1 presents the distribution of these into the training, development and test sets.

Table 1: The distribution of the overlaps in the data set.

Set	No. of Dialogs (%of Dialogs)	Duration	No. of Instances		Class Distribution	
			Cmp	*Ncm*	*Cmp*	*Ncm*
Train	341 (60.4)	2 hrs 55 mins	2379	7158	24.94	75.06
Dev	109 (19.3)	1 hrs 15 mins	724	2295	23.98	76.02
Test	115 (20.4)	58 mins	763	2580	22.82	77.18

3.2. Feature Extraction

In this section we present the feature extraction and combination processes.

3.2.1. Lexical Features (Lex):

The lexical features are from both of the interlocutor's channels. To capture the context, trigrams of tokens are extracted. Since this yields a very large feature vector, thus increases computational complexity, top $5,000$ features are selected. The features are then transformed into a bag-of-words (i.e., bag-of-trigrams) vector space model [21].

3.2.2. Acoustic Features (AC):

The recent success of the use of low level acoustic features and their projection onto statistical functionals has been applied to many paralinguistic tasks [22, 15, 23, 24]. The acoustic features are extracted using openSMILE [25] with frame size of 25 milliseconds and 100 frames per second. The low level acoustic features include prosodic, spectral, voice quality, mfcc and energy. These low-level features along with their derivatives are then projected onto 24 statistical functionals such as 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 [14].

The low level features are extracted from both agent and customer channels. As shown in Equation 1, $CH1$ and $CH2$ represents the feature vectors of channel-1, and channel-2, respectively. We merge these feature vectors to create a new feature vector S that is used for the categorization experiments.

$$\begin{aligned} CH1 &= \{a_1, a_2, ..., a_m\} \\ CH2 &= \{c_1, c_2, ..., c_m\} \\ S &= \{CH1, CH2\} \\ S &= \{a_1, a_2, ..., a_m, c_1, c_2, ..., c_m\} \end{aligned} \tag{1}$$

3.2.3. Feature Combination:

In addition to the individual feature set, we also evaluate the linear combination of acoustic and lexical features. Let $S = \{s_1, s_2, ..., s_m\}$ and $L = \{l_1, l_2, ..., l_n\}$ denote the acoustic and lexical feature vectors respectively. After the linear combination, the feature vector is represented by $Z = \{s_1, s_2, ..., s_m, l_1, l_2, ..., l_n\}$ with $Z \in R^{m+n}$.

3.3. Classification Algorithms

3.3.1. Support Vector Machines

For the classification we use Support Vector Machine (SVMs) implementation of Weka [26]. The models are trained using the linear kernel with the default parameters; and the feature values are normalized within $[0, 1]$ interval.

3.3.2. Deep Neural Networks

Figure 2 depicts the architecture of the fully-connected feed-forward neural network. In the architecture, the layers are densely connected, and each layer consists of a different number of units (u). The input to the DNN architecture is a vector x, which consists of individual feature sets or a linear combination of the acoustic and lexical features. The input is mapped to the output $y = f(x) = g(W.x)$, where the function $g(.)$ is some activation function, and $W \in R^2$ is a matrix of parameters. For the input, the feature values are scaled with zero mean and unit variance.

In the hidden layers of the DNN, we use rectified linear unit (ReLU) [27] as an activation function. We experimented with ReLU function due to its linear, non-saturating form, which helps greatly to accelerates the convergence of stochastic gradient descent compared to the other functions, such as sigmoid or tanh. For the output layer we use the softmax function. The number of hidden units per layer is given in Figure 2. These optimal values are obtained empirically on the development set using Adagrad [28] optimization and a batch size of 100.

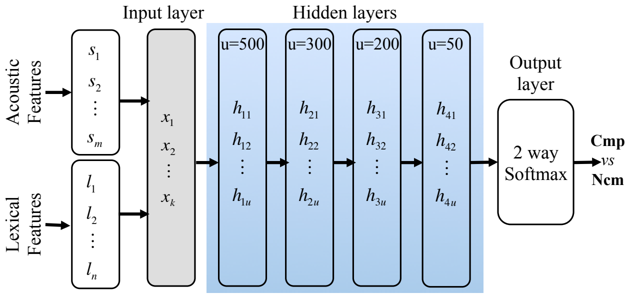

Fig. 2: The DNN architecture for the classification of competitiveness in overlapping speech. u represents number of units in each hidden layer. The input layer vector x can be acoustic feature vector S ($k = m$) or lexical feature vector L ($k = n$) or their linear combination ($k = m + n$)

.

3.4. Evaluation Methodology

The system performances are evaluated using standard Precision (**P**), Recall (**R**) and F-measure (**F1**). Due to the imbalanced distribution of labels, overall system performance is evaluated as macro-averaged precision, recall and F-measure. For the clarity of presentation, we report only the F_1 measure of each class and the overall system. The statistical significance of the results is evaluated using McNemar's test in Section 4.

4. RESULTS AND DISCUSSION

The results of the classification experiments are reported in Table 2. The SVM model performances for the acoustic and lexical features are considered as a baseline and are taken from [15].

For the SVM, we observe a significant $(p < 0.05)$ improvement in performance using linear feature combination, especially for competitive overlaps. A significant increase in F1 of 4.50% and 5.31% on the test set is observed compared to the individual SVM models using acoustic and lexical features only. For the non-competitive overlap class, on the other hand, the feature combination outperforms the lexical model only. The model trained on acoustic feature outperforms both the lexical and the linear combination models.

The DNNs use the same architecture for the individual features sets and their combination. DNN architecture for the acoustic feature set significantly outperforms both individual feature SVM models. An improvement of $\approx 2\%$ in F1 is observed for competitive overlap with respect to the acoustic feature set using SVM model. We do not observe a similar pattern for the non-competitive class where SVM with acoustic features yields F1 of 0.85 compared to F1 of 0.84 for the DNN with acoustic features only. The overall performance for the lexical features is poor with respects to the rest of the experimental results. The weak performance of lexical features has been observed especially for competitive overlaps. This can be due to the fact that lexical pattern describing non-competitive classes are closed set whereas for competitive they are very open i.e., any words can be used to express the competitiveness intension. Moreover, from experimental point of views, lexical feature design used here is very basic. So there are future scope for using more advanced feature extraction technique such as convolution based features.

For the combined feature set, DNN architecture not only improves the F-measure of the competitive overlap class, but also of the non-competitive class, and, consequently, the performance of the whole system. An improvement of 7.39% and 8.20% is observed for competitive overlaps when compared to individual feature SVM models. A similar pattern is observed for the non-competitive overlap class with DNNs using feature combination when compared to the individual feature SVM models.

Comparing SVM and DNN models using the feature combination, we observe an increase of 2.89% in F1 for competitive overlap class, 2.48% in non-competitive overlap class and 2.24% for the system overall.

Table 2: F1 measure for the individual classes and the macro-averaged F1 for the system as a whole on the development and test sets. AC – Acoustic, Lex – Lexical, AC + Lex – Feature combination of acoustic and lexical feature sets.

F1		**Dev-set**			**Test-set**		
Classifier	**Feat.Set**	**Cmp**	**Ncm**	**Overall**	**Cmp**	**Ncm**	**Overall**
SVM	AC	0.46	0.85	0.65	0.44	0.85	0.64
	Lex	0.46	0.83	0.64	0.43	0.82	0.63
	AC + Lex	0.54	0.84	0.69	0.48	0.83	0.66
DNN	AC	0.51	0.85	0.68	0.46	0.84	0.65
	Lex	0.37	0.84	0.61	0.32	0.84	0.58
	AC + Lex	*0.57*	*0.87*	*0.72*	**0.51**	**0.86**	**0.68**

5. CONCLUSIONS

In this paper we have investigated the power of acoustic and lexical features and their combination to automatically categorize overlapping speech as competitive or non-competitive using linear (SVM) and non-linear (DNN) algorithms. The combination of lexical and acoustic feature in a linear architecture significantly outperforms individual feature set models by 3% and 2% for acoustic and lexical features respectively. Exploiting many layers of a non-linear information processing for high-dimensional features in a Deep Learning approach, on the other hand, yields another significant improvement of 2% on top of a linear model with the feature combination. The unbalanced natural distribution presents a challenge for the DNN, however the representational power of the architecture also gives it an edge. To the best of our knowledge, this is the first study on competitive *vs* non-competitive overlaps using deep neural networks. Consequently, there is a space for further improvement considering different deep architectures.

6. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union - 7th Framework Programme (FP7/2007-2013) under grant agreement n° 610916 - SENSEI - http://www.sensei-conversation.eu/.

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

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

[4] Candace 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.

[5] Shammur Absar Chowdhury, Evgeny Stepanov, and Giuseppe Riccardi, "Predicting user satisfaction from turn-taking in spoken conversations," in *Proc. of INTERSPEECH*, 2016.

[6] Julia 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.

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

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

[9] 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.

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

[11] 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.

[12] Catharine Oertel, Marcin Wlodarczak, Alexey Tarasov, Nick Campbell, and Petra Wagner, "Context cues for classification of competitive and collaborative overlaps," *Proc. of Speech Prosody 2012*, 2012.

[13] 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.

[14] Shammur Absar Chowdhury, Morena Danieli, and Giuseppe Riccardi, "Annotating and categorizing competition in overlap speech," in *Proc. of ICASSP*. IEEE, 2015.

[15] Shammur Absar Chowdhury, Morena Danieli, and Giuseppe Riccardi, "The role of speakers and context in classifying competition in overlapping speech," in *Proc. of INTERSPEECH*, 2015.

[16] John Nickolls and William J Dally, "The gpu computing era," *Micro, IEEE*, vol. 30, no. 2, pp. 56–69, 2010.

[17] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. of ICASSP*. IEEE, 2013, pp. 6645–6649.

[18] Shujie Liu, Nan Yang, Mu Li, and Ming Zhou, "A recursive recurrent neural network for statistical machine translation.," in *Proc. of ACL*, 2014, pp. 1491–1500.

[19] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio, "Why does unsupervised pre-training help deep learning?," *JMLR*, vol. 11, no. Feb, pp. 625–660, 2010.

[20] 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*, 2014.

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

[22] 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.

[23] Firoj Alam and Giuseppe Riccardi, "Comparative study of speaker personality traits recognition in conversational and broadcast news speech," in *Proc. of Interspeech*. 2013, pp. 2851–2855, ISCA.

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

[25] 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 (ACMM)*. ACM, 2013, pp. 835–838.

[26] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[28] John Duchi, Elad Hazan, and Yoram Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011.