Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language xxx (2018) xxx-xxx

www.elsevier.com/locate/csl

Automatic classification of speech overlaps: Feature representation and algorithms☆

Q2 Shammur Absar Chowdhury*, Evgeny A. Stepanov, Morena Danieli, Giuseppe Riccardi

Department of Information Engineering and Computer Science, University of Trento, 38123, Italy

Received 24 March 2018; received in revised form 31 October 2018; accepted 2 December 2018
Available online xxx

Abstract

Overlapping speech is a natural and frequently occurring phenomenon in human−human conversations with an underlying purpose. Speech overlap events may be categorized as competitive and non-competitive. While the former is an attempt to grab the floor, the latter is an attempt to assist the speaker to continue the turn. The presence and distribution of these categories are indicative of the speakers' states during the conversation. Therefore, understanding these manifestations is crucial for conversational analysis and for modeling human−machine dialogs. The goal of this study is to design computational models to classify overlapping speech segments of dyadic conversations into competitive *vs.* non-competitive acts using lexical and acoustic cues, as well as their surrounding context. The designed overlap representations are evaluated in both linear − Support Vector Machines (SVM) − and non-linear − feed-forward (FFNN), convolutional (CNN) and long short-term memory (LSTM) neural network − models. We experiment with lexical and acoustic representations and their combinations from both speaker channels in feature and hidden space. We observe that lexical word-embedding features significantly increase the overall F_1-measure compared to both acoustic and bag-of-ngrams lexical representations, suggesting that lexical information can be utilized as a powerful cue for overlap classification. Our comparative study shows that the best computational architecture is an FFNN along with a combination of word embeddings and acoustic features.

Keywords: Overlap; Acoustic; Lexical; Deep learning; Spoken conversation

1. Introduction

The naturalness and the complexity of human interactions are manifested in the context of ordinary daily conversations. In such context, there is no prearranged format for taking the conversational floor, and turn-taking behavior depends on the local management between the speakers. Earlier studies on human conversation would suggest that one speaker speaks at the time, and the exchange of the floor (turn) between the speakers occurs with no to minimum gap and no overlapping turns (Sacks et al., 1974). Consequently, overlaps were considered as a violation of the

☆ This paper has been recommended for acceptance by Roger K. Moore.

* Corresponding author.
E-mail addresses: shammur.chowdhury@unitn.it (S.A. Chowdhury), evgeny.stepanov@unitn.it (E.A. Stepanov), morena.danieli@unitn.it (M. Danieli), giuseppe.riccardi@unitn.it (G. Riccardi).

https://doi.org/10.1016/j.csl.2018.12.001

fundamental rule of turn-taking (Sacks et al., 1974; Duncan, 1972). However, over the years, researchers from different fields studying different aspects of spoken conversations found out that overlapping speech is a frequently occurring phenomenon in the course of human−human interaction (Heldner and Edlund, 2010; Shriberg et al., 2001b). In particular, in Schegloff (2000, 2001), the author provided empirically grounded evidence of the occurrence of overlapping speech and hypothesized an 'overlap resolution device' at work in a conversation. According to the author, the 'overlap resolution device' exploits the resource of turn production, and uses such resources on the basis of an 'interactional logic' that drives the 'moves' of conversants in what, the author name "a competitive sequential topography" (Schegloff, 2000, p. 50).

Over the years, researchers from different fields studied the phenomenon (Shriberg et al., 2001b). The attention of researchers focused also on investigating overlap models that could be viable in speech technology (Heldner and Edlund, 2010). The research in those different, yet related, fields suggests that the tendency to overlap may reveal speakers' attitudes − in particular, their intentions with respect to the control of the turn-taking structure of a conversation. Additionally, it has been observed that overlaps are predictive of the conversation success (Chowdhury et al., 2016a). In West (1979), it has been proposed that speech overlaps are related to dominance or aggression towards the other speaker. However, not all of the overlaps are related to competitiveness − they also support cooperativeness in a conversation by providing the other speaker with the cues about the mutual understanding (Goldberg, 1990) and can reflect an empathic behavior of the speaker (Alam et al., 2016). An accepted categorization of overlaps, in the literature, is into *competitive (C)* and *non-competitive (NC)*. The former is *an attempt to grab the floor* and the latter is *an attempt to assist the speaker for the continuation of the current turn* (Chowdhury et al., 2015a; Schegloff, 2000; French and Local, 1983).

The automatic classification of the overlaps in terms of the overlapper's intention is a key component of systems designed to interpret humans' behavior from the language signals. In Alam et al. (2016), the authors suggest that incorporating the overlap discourse may lead to a better prediction of both basic and complex emotions. Additionally, the classification of the overlaps into competitive and non-competitive is also crucial for improving the quality and the naturalness of different speech technologies such as spoken dialog systems (SDS), virtual agents (VA), and automatic speech recognition (ASR). To improve the quality and the naturalness of SDS and to understand human−human interaction, the speech community has been investigating the acoustic and temporal properties of overlaps.

The research questions we are addressing in this study are: (1) How to represent the overlapper's (initiator of the overlapping speech) information along with the content of the overlappee's (current speaker) turn for the classification of overlaps into competitive and non-competitive? (2) How to represent and combine different sources of information − i.e. what is the best representation for the acoustic and lexical features in isolation and jointly? (3) Do non-linear modeling techniques add any benefit for the classification of overlaps into competitive and non-competitive? Our final objective is to automatically classify overlaps as competitive *vs.* non-competitive using acoustic and lexical information from both speakers in isolation and in combination.

For the purpose of modeling competitiveness, we use a large amount of ecological conversational data, i.e. the data has been collected using real users solving real problems that require understanding the speakers' situations, have real consequences, and lead to rising of different emotions. The data has been used in previous studies to (a) design the guidelines for the annotation of speech overlaps (Chowdhury et al., 2015a); (b) explore the potential of low-level acoustic features in the task of predicting overlap categories (Chowdhury et al., 2014; 2015a); (c) study the role of each speaker and the surrounding information (context), individually and in combination, using basic bag-of-ngrams lexical feature, psycholinguistic features along with part-of-speech tags and acoustic features (Chowdhury et al., 2015b); and (d) investigate if a deep learning approach gives an edge for overlap classification using acoustic and lexical (bag-of-ngram) features (Chowdhury and Riccardi, 2017).

In the context of previously published work, as well as seminal works such as Schegloff (2000), Beňuš et al. (2011) among others, the novel contributions of this paper include: (a) an investigation of different lexical feature representations (ngrams, word-embeddings, etc.), which, to the best of our knowledge, is a first study to utilize different lexical representations of information for the overlap classification task; (b) a study of the effects of acoustic and lexical feature combination techniques: feature space and hidden space combination; (c) a comparative analysis of different computational architectures (e.g. SVM, FFNN, LSTM) for overlap classification; and (d) modeling the concurrent speech information from both speakers to classify the competitiveness in overlapping speech segments.

Please cite this article as: S. Chowdhury et al., Automatic classification of speech overlaps: Feature representation and algorithms, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.12.001

The rest of the paper is structured as follows. In Section 2, we discuss previous studies on overlapping speech. In Section 3 we describe the data set used throughout the paper and its annotation. Section 4 provides details on the experimental methodology − including feature extraction, classification algorithms, and the evaluation methodology. Section 5 presents the results of the overlap classification experiments; followed by the discussion of the results in Section 6 and a presentation of an application of overlap classification to another task − prediction of user satisfaction − in Section 7. In Section 8 we provide a conclusion of the study and the future directions.

2. Related work

Most of the studies on overlaps indicate the importance of prosodic features, recognizing fundamental frequency (f0) and intensity as the dominant features (French and Local, 1983; Kurtić et al., 2013; Truong, 2013). In Shriberg et al. (2001a), the authors suggested that speakers raise their energy and voice when they attempt to interrupt the current speaker. In Hammarberg et al. (1980), similar observations were made for pitch and amplitude.

Features such as speech rate, cut-offs, and repetitions were also analyzed by conversational analysts. In Schegloff (2000), it was observed that speakers use variations in prosodic profiles and repetitions to indicate competitiveness. The phenomenon was also observed in languages other than English: in Danieli et al. (2002), Bazzanella (1996), it was observed that in Italian human−machine dialogs repetitions and overlaps are not necessarily competitive, but play an important pragmatic role with respects to discourse cohesion.

In Jefferson (1982), it was observed that temporal features related to the position of overlaps and the onset position of an overlap are important for their classification. Whereas in French and Local (1983), the authors argued that the phonetic design plays an important role in the representation of competitive overlaps, rather than its precise location. The observation was also supported by Wells and Macfarlane (1998) and Hammarberg et al. (1980). The overlap duration was observed to be the most distinguishing feature for the classification of overlaps into competitive and non-competitive using decision trees (Kurtic et al., 2010; Jefferson, 2004). The authors stated that non-competitive overlaps tend to be shorter and are resolved shortly after the second speaker recognizes the overlap; and competitive overlaps are longer since interlocutors keep on speaking regardless of overlapping.

Automatic classification of overlaps has been addressed using different types of features, such as disfluencies and hand gestures (Lee et al., 2008), body movements from both speakers, contextual prosodic features from the overlapper (Oertel et al., 2012), gaze, voice quality and contextual features from preceding and overlapping segments (Truong, 2013). In Chowdhury et al. (2015a), the authors used different high-dimensional acoustic features for classifying overlaps and suggested that prosodic and spectral features play an important role.

Similar to many other studies, competitive overlaps were also studied under the term *interruption* in Lee and Narayanan (2010). The authors observed that interruptions are not random, and context can be used to predict their occurrences. A similar conclusion was reported in Gravano and Hirschberg (2012), suggesting that interruptions are more likely to occur during/after intonational phrase units (IPUs), which have fewer turn-yielding cues than the IPUs signaling more turn-yielding cues (Gravano and Hirschberg, 2011). In Chowdhury et al. (2015b), the authors observed that context (surroundings of the overlapping events, including the turns preceding and following the overlap along with the content of the overlappee) plays an important role in distinguishing the competitiveness. Among all the studies on overlapping speech, very few have addressed the utility of lexical content of overlaps (Chowdhury et al., 2015b; Chowdhury and Riccardi, 2017). Moreover, the representations of lexical content that have been used in these studies are very basic.

3. Data set

In this research we have used the *SISL Human−Human Conversational Discourse Corpus* (Alam et al., 2018; Chowdhury, 2017). The corpus is a subset of a larger Italian call-center corpus that consists of recordings of conversations between customers and call-center agents. The customers were calling to solve specific problems or seeking information.

The phone conversations, allowing duplex communication, were recorded on two separate channels with 16 bits sample and 8 kHz sampling rate. The average duration of 10 K conversations is 396.6 ± 197.9 s. Out of the 10 K conversations, 565 (approximately 62 h in total, with an average duration of 395 s) have been annotated for speech overlaps by two expert annotators (native Italian speakers) following the guidelines described in Section 3.1.

Please cite this article as: S. Chowdhury et al., Automatic classification of speech overlaps: Feature representation and algorithms, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.12.001

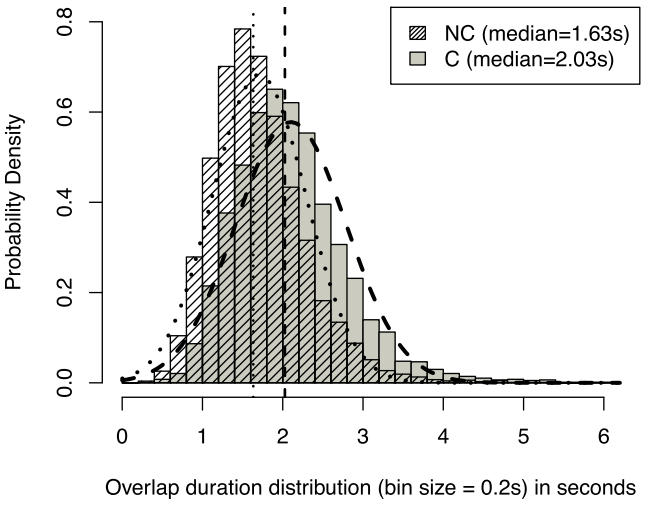

Fig. 1. Distribution of durations of competitive (**C**) *vs.* non-competitive (**NC**) intervals in the corpus (Chowdhury et al., 2015b). The density of the distribution is presented using the dotted (for NC) and dashed (for C) curves.

The annotation task consisted of manual segmentation of speech for overlaps and labeling the overlapping segments as competitive (C) or non-competitive (NC). For the task, the annotators are provided with recordings of conversations, with no transcription, to aid the manual annotation of the overlap boundaries. The quality of annotation is evaluated as inter-annotator agreement, which is discussed in Section 3.2.

The distribution of types of turn-transitions, such as gaps, between and within speakers lapses, pauses and overlaps present in the dialogs is shown in Fig. 2 along with the duration distribution of both overlap classes in Fig. 1, indicating competitive (C, median duration of 2.03 s) overlap instances are usually longer with respect to non-competitive (NC, median duration of 1.63 s).

3.1. Annotation guidelines

The annotation guideline has been used to train annotators to segment and label all speech overlaps in the dialogue corpus. The binary labeling into competitive (C) *vs.* non-competitive (NC) has been motivated by the studies in Schegloff (2000, 2001) and French and Local (1983). In our study, the competitive and non-competitive overlap categories can be seen as 'meta' categories for the labels used in Schegloff (2001). However, unlike Schegloff (2000) – where the authors excluded the overlapping speech segments that do not suggest the activation of the proposed 'overlap resolution device' – the annotation, in this study, consider all the speech overlap occurrences[1] for the binary classification task of **C** *vs.* **NC**. In the following descriptions, we use the terms defined by Schegloff (2001) to explain some observed phenomena.

Prior to the annotation, a sample of conversations has been analyzed by a psycholinguist, who listened to each recorded call by applying a systematic direct observation protocol (Ericsson and Simon, 1984), focusing only on overlapping speech segments. The observations allowed the psycholinguist to identify different kinds of overlapping speech segments, varying with respect to their pragmatic functions, speaker's intentions and linguistic structures. The annotation guidelines for segmentation and categorization of the speech overlaps into competitive and non-competitive categories have been designed on the basis of this observational analysis. The salient parts of the annotation guidelines are the following:

1. An overlapping segment may contain more than one overlap instance of the same category.[2]

[1] With the exception of human sounds, for example, laughs and voice clearing, that although may contribute to manifest speaker intentions and mental states, are not considered in this study.

[2] In the annotated data it has been observed that the overlap instances may be separated from each other with a gap of less than 40 ms.

Please cite this article as: S. Chowdhury et al., Automatic classification of speech overlaps: Feature representation and algorithms, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.12.001

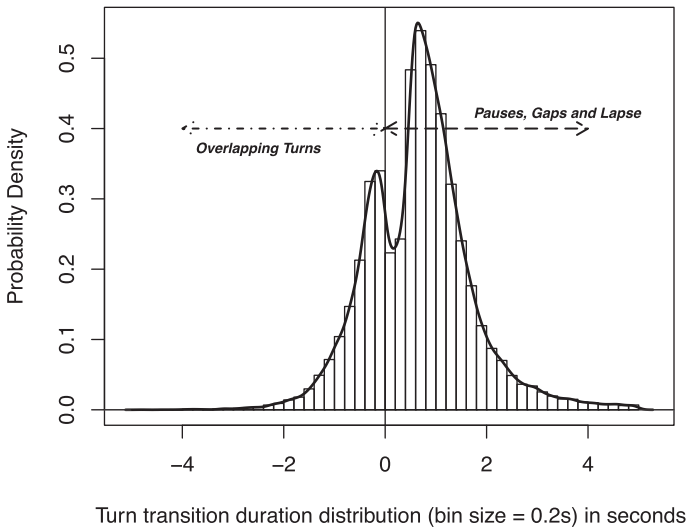

Fig. 2. Turn transition duration distribution in the 565 dialogs. The positive transitions include gaps (and smooth-switches), pauses, and lapses (inter and intra speaker silences with duration longer than 2 s); and negative transitions are overlaps.

2. A speaker's thinking aloud during another speaker's turn is considered an overlap and categorized as non-competitive.
3. Co-occurrences of "simultaneous starts" by both of the speakers are considered instances of speech overlaps, if and only if the segments contain complete words[3] 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, and other sounds like cough, sneezes and laughter.
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 utterance.
6. Inferring the overlap category on the basis of the annotator's knowledge of what will occur later in the conversation, i.e. outside of the turn being considered, is to be strictly avoided.

In the following section, we provide annotated examples of the speech overlaps and their dialog context.

3.1.1. Competitive overlaps

Competitive (C) overlaps comprise scenarios where the intervening speaker starts prior to the completion of the current speaker's turn; and both the speakers display interest in holding the turn for themselves. The annotators perceive the overlap event in these scenarios as problematic for both speakers.

Examples 1−3 illustrate different instances of competitive overlaps with their English translations. In the examples, the overlap segments are represented within square brackets ([...]) and silences greater and equal to 0.5 s are indicated as (**.**). The dialog excerpt in Example 1 illustrates the use of overlap to clarify the provided information. The excerpt in Example 2 illustrates the agent (A) predicting what the customer (C) is going to say and asking the customer a question with a competitive act. Whereas Example 3 illustrates an instance of competitive overlap where the speaker who initiates the overlapping sequence failed to interrupt the other speaker. It is observed that competitive overlap instances are more diverse in their functionality and lack a proper model for their further categorization.

[3] To avoid instances of a speech onset lacking the complete cognitive and motor planning of the communicative act.

Please cite this article as: S. Chowdhury et al., Automatic classification of speech overlaps: Feature representation and algorithms, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.12.001

Example 1.

A: allora sì (.) vedo che c'è una riduzione della potenza in seguito a una serie di fatture non pagate che
yes (.) (I) see that there is a reduction of the power due to some unpaid bills that ...
A: [fanno un importo ah]
[amount to ...]
C: [stamattina è stata disattivata] non è stata ridotta la potenza
[this morning it was suspended] it was not a power reduction

Example 2.

C: la vostra raccomandata di risposta alla mia alla [mia richiesta]
your recorded-delivery letter in reply to my to [my request]
A: [ah e che c'era scritto la raccomandata]
[ah and what was the content of the recorded-delivery letter?]
C: so cosa c'era scritto voi dovete averla io (.) ce l'ho davanti
I know the content it is you that need to have it I (.) have it in front of
A: cioè non posso aiutarla in questo le posso dire solo che qui come saldo fornitura mi risultano 3,652 euro
that is I can not help you with this I only can say you that here in terms of balance of supply I see 3,652 euros

Example 3.

C: perfe allora [lei deve fare una cosina lei ha un delle] belle schermate a disposizione mi deve aprire la mia ehe il mio fax inviato il ventitrè zero otto duemiladodici cortesemente
perfe. then [you have to do a small thing you have some] beautiful screens available you have to open my own and you will find my fax sent on 23rd of August 2016
A: [però e se]
[but and if]
A: vediamo subito
let us see immediately

3.1.2. Non-competitive overlaps

Non-competitive (NC) overlaps comprise scenarios where another speaker starts in the middle of an ongoing turn; both parties do not show any evidence for grabbing the turn for themselves. The annotators do not perceive this overlap as problematic for both the speakers. The intervening speaker uses such overlaps to signal the support for the current speaker's continuation of the turn.

Examples 4 and 5 illustrate instances of non-competitive overlaps along with their English translations. Similar to previous examples, the overlap segments are represented within ([...]) and silences greater or equal to 0.5 s are indicated as (.). The dialog excerpt in Example 4 illustrates the use of overlap to provide support and encourage the current speaker to continue his/her turn. Example 5, on the other hand, illustrates a simple feedback overlap. Other observed non-competitive overlap instances, in the annotated data, includes collaborative utterance constructions, some choral phenomena (e.g. greetings and salutations) among others.

Example 4.

C: un blocco che avete un problema voi tra uffici
a block (due to) a problem you have within your (administrative) departments
A: ah
ah ...
C: allora mi faccia una cortesia [ehe perché se] non riesco a parl devo parlarle ho parlato con cinque suoi colleghi e mi hanno chiamato due consulenti
then please [eh because if] I cannot tal ... I need to talk ... I talked with five colleagues of you and two consultants called me
A: [mi dica]
[(please) tell me]

Please cite this article as: S. Chowdhury et al., Automatic classification of speech overlaps: Feature representation and algorithms, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.12.001

Example 5.

A: ah ascolti qui ci sono una serie di fatture malgrado
Listen (please) we have here a number of unpaid bill in spite of
A: [ci]
[(in spite of) there is]
C: [mh beni]
[mhm well]
A: sia il blocco per sisma vedo che c è in LOCATION per
the block due to the earthquake I see that there is in LOCATION

3.2. Inter-annotator agreement

To assess the reliability of the annotations, two annotators have independently labeled a subset of 28 conversations (approximately 3 h 17 min) randomly extracted from the call center corpus. The subset is used to compute Cohen's κ (Cohen, 1960; Carletta, 1996) and positive (specific) agreement (Fleiss, 1975).

The Kappa statistics (κ, Eq. (1)) is frequently used to assess the degree of agreement among any number of annotators by adjusting the observed agreement (P_o, Eq. (2)) for the hypothetical probability that the annotators agree by chance (P_e, Eq. (3)). In the equations, observed (P_o) and chance (P_e) agreements are expressed in terms of *true positives* (TP), *true negatives* (TN), *false positives* (FP) and *false negatives* (FN); and N = TP+TN+FP+FN.

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

$$P_o = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{N}} \tag{2}$$

$$P_e = \frac{\frac{(\mathrm{TP} + \mathrm{FP}) * (\mathrm{TP} + \mathrm{FN})}{\mathrm{N}} + \frac{(\mathrm{TN} + \mathrm{FP}) * (\mathrm{TN} + \mathrm{FN})}{\mathrm{N}}}{\mathrm{N}} \tag{3}$$

Positive (Specific) Agreement (P_{pos}, Eq. (4)) (Fleiss, 1975), which is identical to the widely used F_1-measure (Eqs. (5)–(7)) (Hripcsak and Rothschild, 2005), is computed to quantify the inter-annotator agreement as human-performance in categorization of overlaps.

$$P_{pos} = \frac{2 * \mathrm{TP}}{2 * \mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{4}$$

$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{5}$$

$$\mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{6}$$

$$F_1 = 2 * \frac{\mathrm{precision} * \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{2 * \mathrm{TP}}{2 * \mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{7}$$

The inter-annotator agreement on the subset of 28 conversations is $\kappa = 0.70$ and $P_{pos} = 0.85$. The cases of disagreement have been discussed in a consensus meeting by the annotators and the author of the guidelines. The most relevant disagreement between annotators concerns speech disfluencies, including simultaneous starts, repairs, and filled pauses. In most of the cases consensus between two annotators has been reached. The kappa statistic was computed at the end of the consensus meeting.

3.3. Lexical characteristics of overlaps

The lexical content of speech overlaps might be indicative of their category. The content of the overlapper (initiator of the overlap) turn has been analyzed in terms of the frequency of unigrams (Fig. 3), bigrams and trigrams (Table 1) in each overlap category.

Please cite this article as: S. Chowdhury et al., Automatic classification of speech overlaps: Feature representation and algorithms, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.12.001

(a) Non-competitive (NC) (b) Competitive (C)

Fig. 3. Word-clouds for competitive (C) and non-competitive (NC) overlap categories.

Table 1
List of the most frequent bigrams and trigrams for competitive (**C**) and non-competitive (**NC**) overlaps with their English translations (in parentheses).

NC	C
sí sí (yes yes), *sí sí sí* , *va bene* (well),	*no no, no no no*,
no no, no no no,	*non é* (it is not),
ho capito (I have understood),	*c é* (there is),
lo so (I know), *eh sí* , *grazie a* (thanks to),	*ho capito* (understood),
la ringrazio (thank you),	*un attimo* (one moment),
grazie a lei (thanks to you),	*no non* (not),
no non, ah okay,	*io non* (I do not),
ah ho capito, un attimo (just a moment),	*ma non* (but not),
si figuri (never mind), *mh mh*,	*eh ma* (yeah but),
bene va (goes well), *va bene va* (alright),	*sí sí* (yes yes),
mi dica (tell me), *non si*,	*ma io* (but I),
sí perché (yes why), *sí no*,	*no ma* (no but),
non lo (not), *eh eh*,	*mi dá* (he gives me),
é stata (it was), *sí infatti* (yes indeed),	*sí ma* (yes but),
okay allora (ok then),	*mi scusi* (excuse me),
va benissimo (thats great),	*non mi* (I do not),
mi conferma (I confirmed),	*no signora* (no madam/lady),
di nulla (nothing),	*no perch* (no because),
non lo so (I don not know),	*mi dá il* (gives me),
conferma che (confirms that),	*io ho* (I have),
sí esatto (yes right),	*però io* (but I),
ci mancherebbe (God forbid).	*non é possibile* (it is not possible).

From the unigrams, it is observed that *sí* (yes) is the most frequently used word to start a non-competitive overlap. For the competitive overlaps, on the other hand, the most frequently used word is *no*. From the analysis of bigrams, it is observed that adversative conjunctions like *ma* (but) are frequently used with *no* to start a competitive overlap.

The Italian phrases that are used to indicate support or agreement with the interlocutor, such as *bene* (well), *ho capito* (I have understood), *certo* (sure), are more frequent in non-competitive overlap segments. On the other hand, the phrases that are used to indicate disagreement (and function as discourse markers), such as *no, ma* (but), *mi scusi* (excuse me), *no perché* (no because), are more frequent in competitive overlap segments. Due to this difference in the distribution of words, lexical features have a potential for capturing intervening speaker's intentions.

4. Methodology

In this section we describe our experimental methodology, which consists of (1) data preparation, (2) feature extraction, (3) machine learning algorithms applied to the automatic classification of speech overlaps, and (4) the evaluation methodology. The overall pipeline for the experiments on overlap classification is presented in Fig. 4.

Please cite this article as: S. Chowdhury et al., Automatic classification of speech overlaps: Feature representation and algorithms, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.12.001

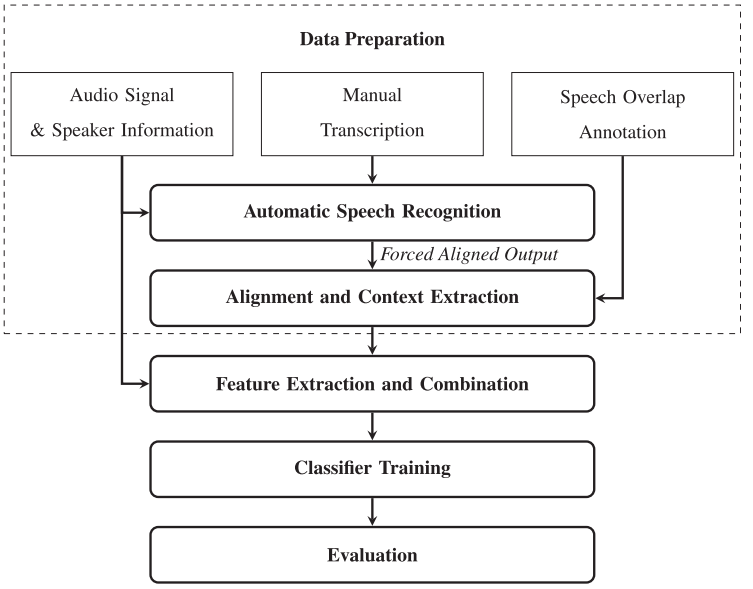

Fig. 4. The pipeline for modeling competitiveness in overlapping speech.

We first prepare the data by selecting overlap instances and extracting their context. Then, we perform feature extraction and combination. Then we train and evaluate classification algorithms. We experiment with different types of features: acoustic (AC) and lexical (Bag-of-Ngrams (BoN) and word embeddings (WE)); and classification algorithms: Support Vector Machines (SVM), feed-forward (FFNN), convolutional (CNN) and long short-term memory (LSTM) neural networks.

4.1. Speech overlap segment normalization

Manually annotated overlap segments frequently include pre- and post-overlap silences and/or non-overlapping portions of the turn. Moreover, overlaps do not necessarily happen at word boundaries, which also complicated the extraction of lexical content of overlaps. Thus, the manually annotated boundaries are adjusted performing the forced alignment using domain-specific Automatic Speech Recognition (ASR) model (Chowdhury et al., 2014) and manual word-level transcriptions. It is important to note that in the forced alignment step only overlap boundaries are adjusted, no additional segments are introduced.

As the result of the forced alignment, we have 15,899 overlap segments with a total duration of 5 h and 8 min (Fig. 5). For the experiments, we split our data into training, development and test sets. Details of the data set are summarized in Table 2.

4.2. Overlap context representation

In order to utilize the contextual information to differentiate between different overlap types, the overlapping segments are extracted along with their left (0.2 s of speech) and right (0.3 s of speech) contexts. The overlap context durations are selected according to Chowdhury et al. (2015b), where the authors experiment with different combinations of overlap and context segments as a representation of overlap instance for automatic classification. Since the calls were recorded on two channels (speaker-per-channel), the feature vectors for interlocutors (O_{s1} and O_{s2}) are extracted according to Eqs. (8) and (9) and then concatenated to form an overlap feature vector (OPC, see Eq. (10)) as depicted in Fig. 6.

$$O_{s1} = \left\{o_{s1}^1, o_{s1}^2, \ldots, o_{s1}^m\right\} \tag{8}$$

Please cite this article as: S. Chowdhury et al., Automatic classification of speech overlaps: Feature representation and algorithms, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.12.001

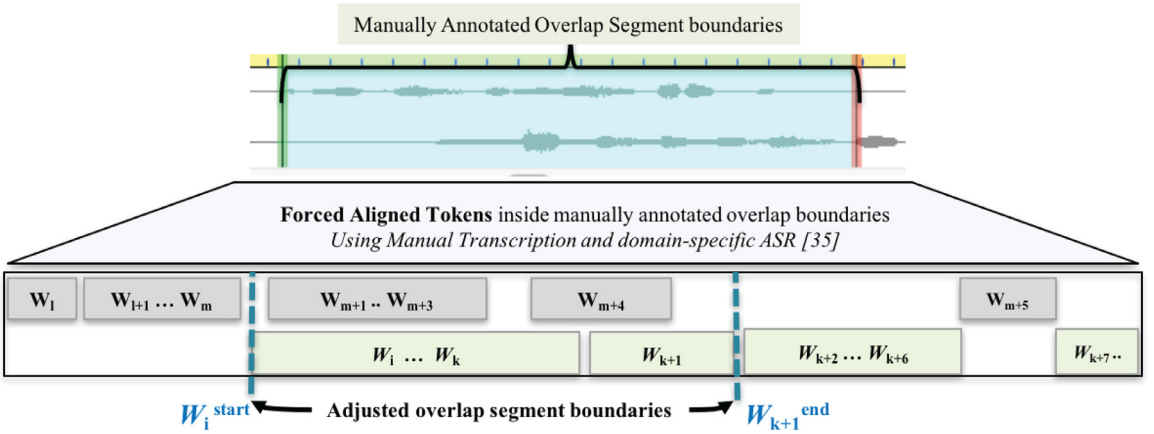

Fig. 5. Normalization of overlapping segment boundaries using the forced alignment. The *adjusted overlap segment boundaries* are later used to extract the overlap and its surrounding (preceding and following) contexts for the automatic classification.

Table 2
Distribution of overlap instances in the whole corpus and training, development and test sets in terms of their duration, raw counts (#) and percentages (%), as well as their distribution into competitive (**C**) and non-competitive (**NC**) categories.

	Dialogs		Overlaps		C		NC	
	#	%	#	**Duration**	#	%	#	%
Train	341	(60.35)	9537	(2 h 55 m)	2379	(24.95)	7158	(75.06)
Dev	109	(19.29)	3019	(1 h 15 m)	724	(23.98)	2295	(76.02)
Test	115	(20.35)	3343	(0 h 58 m)	763	(22.82)	2580	(77.18)
Total	565	(100.0)	15,899	(5 h 08 m)	3866	(24.32)	12,033	(75.68)

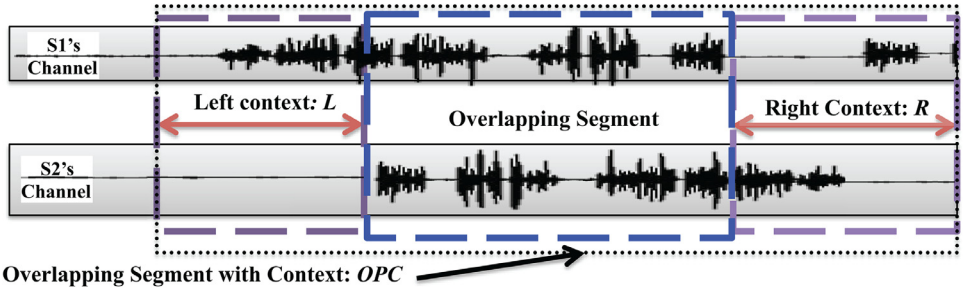

Fig. 6. An example of overlapping speech segment from the speaker $S1$ in channel 1 and speaker $S2$ in channel 2. The concatenation of O_{s1} and O_{s2} forms OPC vector – a representation of the speech overlap segment along with the left and right context (outer box).

$$O_{s2} = \{o^1_{s2}, o^2_{s2}, \ldots, o^m_{s2}\} \tag{9}$$

In the equations, m is the dimension of the extracted features, O_{s1} and O_{s2} represent the feature vectors for speakers $s1$ and $s1$ respectively, o^i_{s*} is the feature i in each vector. The speech overlap (OPC) is then modeled as the concatenation of the two vectors (see Eq. (10)).

$$OPC = \{o^1_{s1}, o^2_{s1}, \ldots, o^m_{s1}, o^1_{s2}, o^2_{s2}, \ldots, o^m_{s2}\} \tag{10}$$

Please cite this article as: S. Chowdhury et al., Automatic classification of speech overlaps: Feature representation and algorithms, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.12.001

Table 3
Low-level acoustic features (LLD) and their derivatives used for automatic overlap classification.

Low-level features
Pitch (fundamental frequency f0, f0-envelope), loudness, voice-probability, Jitter, shimmer, logarithmic harmonics-to-noise ratio (logHNR), Mel-frequency cepstral coefficients (MFCC 0-12), Logarithmic signal energy from pcm frames, Energy in spectral bands (0−250 Hz, 0−650 Hz, 250−650 Hz, 1−4 kHz), roll-off points (25%, 50%, 70%, 90%), centroid, flux, max-position and min-position.

4.3. Feature extraction

In this section we describe different feature classes that are utilized for the automatic classification of speech overlaps into competitive (C) and non-competitive (NC).

4.3.1. Acoustic features

Due to their successful application to many paralinguistic tasks (Schuller et al., 2011; Chowdhury et al., 2015b; Alam and Riccardi, 2013; Danieli et al., 2015; Alam et al., 2019) (including unsupervised overlap clustering Chowdhury et al., 2014), the low-level acoustic features (LLD) and their projections onto statistical functionals are used as the representation of acoustic properties of speech overlaps. The acoustic features are extracted using openSMILE (Eyben et al., 2013) with frame size of 25 ms and overlap of 10 ms (100 frames per second). Following Chowdhury et al. (2015a), the low-level acoustic features include prosodic, spectral, voice quality, mfcc, and energy (see Table 3). The low-level features and 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 (Chowdhury et al., 2015a).

4.3.2. Lexical features

To classify the competitiveness in overlapping speech, the lexical information extracted from overlap segments and the surrounding context after the forced alignment is represented as two types of vectors − Bag-of-Ngrams (BoN) and Word Embeddings (WE).

Bag-of-Ngrams (BoN). The most commonly used representation of text in Natural Language Processing is bag-of-words (vector space model) (Joachims, 1998), where text is represented as a 'bag' of its words ignoring their order. Bag-of-ngrams; consequently, represents a text as a 'bag' of ngrams. In our experiments, we use 5000 most frequent ngrams (unigrams, bigrams, and trigrams).

Word embeddings (WE). Word embedding is a distributional semantic models (DSMs) where words are mapped from a dimension-per-word space to a continuous vector space of much lower dimension. Unlike traditional DSMs that make use of co-occurrence counts (Baroni et al., 2014), embeddings are learned from large corpora by training neural networks. It has been demonstrated in the literature that such representation captures semantic and syntactic relations among words better (Bian et al., 2014). The word embeddings used in our experiments (Chowdhury, 2017) were trained using *gensim* (Řehůřek and Sojka, 2010) implementation of Mikolov et al. (2013), Mikolov and Dean (2013) word vector model. The model was trained using the CBOW approach with a size of the feature vector 500, a context window of 5, negative-sampling with a value of $k = 10$, and cut-off frequency 5. The resulting word-embedding model contains 6 billion words with a vocabulary size of 2.84 millions.

4.3.3. Feature combination

In addition to the experiments with acoustic and lexical features, we also experiment with their linear combination. The linear combination is performed following two approaches: (1) feature space combination and (2) hidden

Please cite this article as: S. Chowdhury et al., Automatic classification of speech overlaps: Feature representation and algorithms, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.12.001

space combination (Chowdhury, 2017). The linear combination of acoustic and lexical features in the feature space consists of concatenation of $S = \{s_1, s_2, \ldots, s_m\}$ and $L = \{l_1, l_2, \ldots, l_n\}$, which are the acoustic and lexical feature vectors respectively. After the linear combination, the feature vector is represented by $Z = \{s_1, s_2, \ldots, s_m, l_1, l_2, \ldots, l_n\}$ with $Z \in R^{m+n}$. Whereas for hidden space combination, we concatenate the hidden units of the dense layer of both acoustic and lexical deep neural network architectures. Details of the architectures are provided in Section 4.4.

4.4. Classification algorithms

Support Vector Machines (SVM), which use a linear pattern separation model, have been proven effective in solving many classification problems. However, in the case of natural conversational speech, such a shallow representation can be problematic. The natural way of understanding human conversation suggests 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) (Nickolls and Dally, 2010), neural networks, containing several hierarchical layers, have been widely applied to different problems in Speech and Natural Language Processing (NLP) and Computer Vision with a significant amount of success (Graves et al., 2013; Liu et al., 2014; Erhan et al., 2010). The approach is termed as "deep learning" or "deep neural networks (DNN)". In this study, we first apply Support Vector Machines (SVM) to classify competitive (C) *vs.* non-competitive (NC) overlaps. Then, we experiment with different Deep Neural Network (DNN) architectures and compare their performances to SVM.

The neural network architectures are also explored for the combination of the information from speaker channels (preserving the overlapper and overlappee distinction) to obtain a good representation for overlaps. The models trained on individual feature sets (acoustic and lexical) are compared to the models trained on the combinations of these features.

4.4.1. Support Vector Machines

For the classification of overlaps, the first algorithm that we have applied is Sequential Minimal Optimization (SMO) − a Support Vector Machine (SVM) implementation of Weka (Hall et al., 2009). Prior to classification, feature values are normalized within [0,1] intervals. The SVM models are trained using the linear kernel and default parameters (i.e., $C = 1.0$).

4.4.2. Feed-Forward Neural Networks

The architecture of the fully-connected Feed-Forward Neural Network (FFNN) for the classification of overlaps is depicted in Fig. 7. In this architecture, the layers are densely connected, and each layer consists of a different number

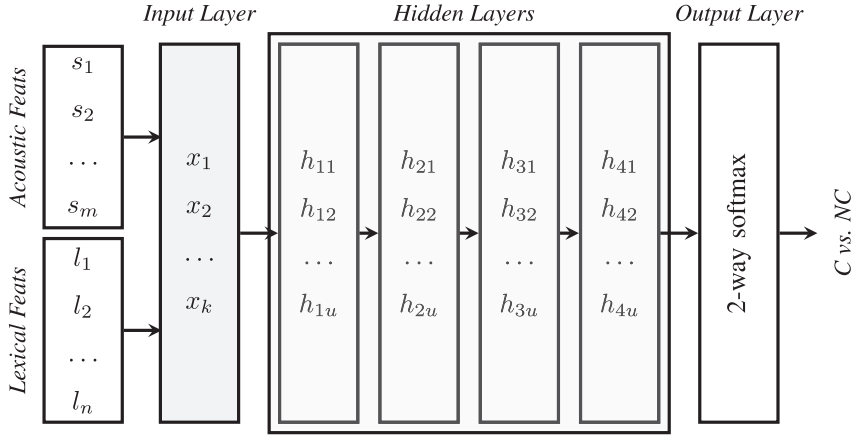

Fig. 7. The Feed-Forward Neural Network (FFNN) architecture for the classification of competitiveness in overlapping speech. u represents the number of units in each hidden layer. The input layer vector x is either acoustic S ($k = m$) feature vector, or lexical Bag-of-Ngram (BoN) ($k = n$) feature vector, or average word embedding (WE) vector ($k = z$), or the linear combination of the above ($k = m + n$ or $k = m + z$).

Please cite this article as: S. Chowdhury et al., Automatic classification of speech overlaps: Feature representation and algorithms, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.12.001

of units (u). The input is a vector x, which consists of individual feature sets − Acoustic (AC) or Bag-of-Ngrams (BoN) − or their linear combination. The input is mapped to the output y as shown in Eq. (11).

$$y = f(x) = g(W.x) \tag{11}$$

In the equation, 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 Feed-Forward Neural Network (FFNN), rectified linear units (ReLU) (Krizhevsky et al., 2012) or sigmoid are used as activation functions (depending on the features and the tasks). For the output layer, the softmax function is used.

4.4.3. Convolutional Neural Networks

The architecture of Convolutional Neural Network (CNN) is illustrated in Fig. 8. The input to the CNN is the lexical contents of the overlap segment extracted from both speakers, where $O = \{o_1, o_2, \ldots, o_n\}$ and $P = \{p_1, p_2, \ldots, p_m\}$ are the sequences of words from each. This lexical content from each speaker is then transformed into a sequence of indices by mapping each word $o_i \in O$ and $p_i \in P$ into an index in L. L is a shared look-up table, $L \in \mathbb{R}^{|V| \times D}$, where D is the dimensional word vector for each word in the vocabulary V.

In our experimental settings, the model parameter L is initialized using pre-trained word-embedding vectors, as described in Section 4.3.2. Using the look-up table L and the indexed sequence, we then create the input matrix for

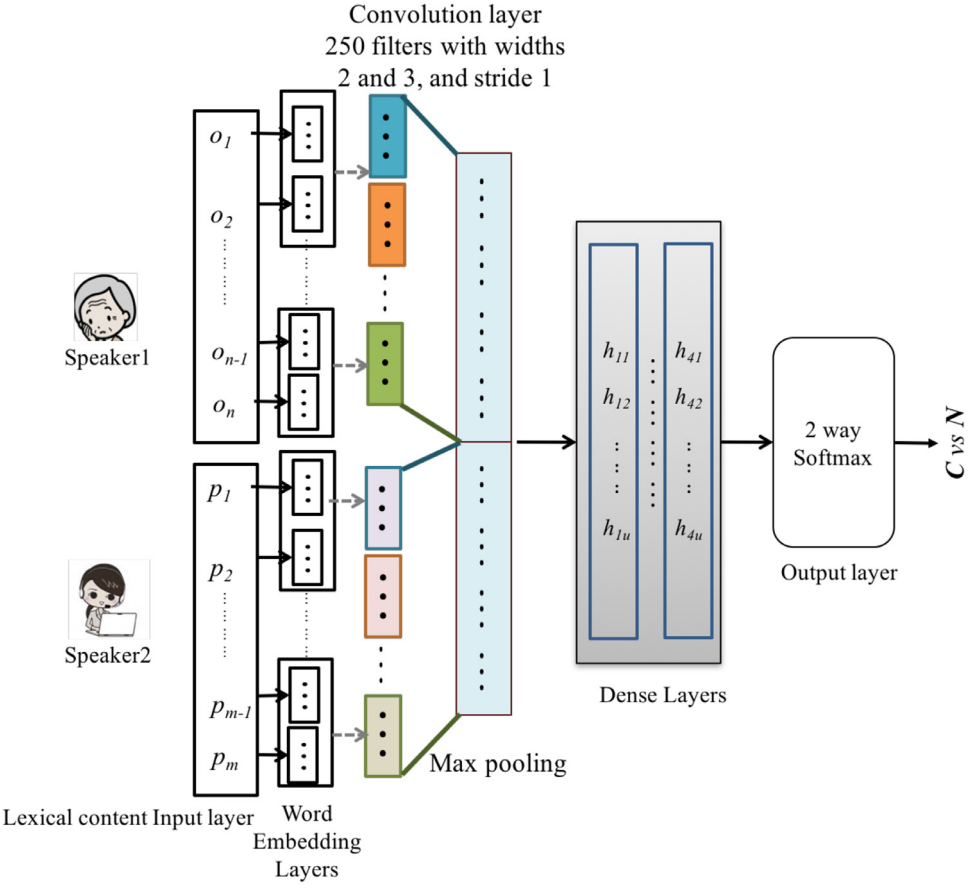

Fig. 8. The CNN architecture for the classification of competitiveness in overlapping speech using word embeddings (WE). u represents the number of units in each hidden layer. The input is the transcription (lexical content) of each speakers ($\{o_1 \ldots o_n\}$ and $\{p_1 \ldots p_m\}$ represents the words sequences for speaker 1 and 2 respectively) on which convolution has been applied, and later the features are concatenated after the max polling operation.

Please cite this article as: S. Chowdhury et al., Automatic classification of speech overlaps: Feature representation and algorithms, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.12.001

each speaker's lexical content to be passed to the convolution layer. Applying the max-pooling operation, we obtain a higher-level feature representation, which is an equal-sized feature vector for each overlap instance. This vector is then passed to one or more hidden layers, followed by an output layer.

Since input transcriptions may vary in length, i.e. number of words, they are zero-padded for an equal length to perform the convolution. The convolution operation involves applying a series of filters $u \in R^{L.D}$ to a window of L words to produce a new feature representation according to Eq. (12).

$$h_t = f(u.x_{t:t+L-1} + b_t) \tag{12}$$

In the equation, $x_{t:t+L-1}$ is the concatenation of L input vectors, b_t is a bias term, and f is a nonlinear activation function − rectifier linear unit (ReLU).

We have applied the filters with sizes 2 and 3 to capture ngram information. This filtering has been applied to generate a feature mapping $h_i = [h_1, h_2, \ldots, h_{T+L-1}]$. Max-pooling operation − a down-sampling strategy (as shown in Eq. (13)) is applied to obtain an equal-sized higher-level feature representation.

$$m = [\mu_p(h_1), \mu_p(h_2), \ldots, \mu_p(h_N)] \tag{13}$$

In the equation, $\mu_p(h_i)$ is the max-pooling operation. It is applied to each window of p features in the feature mapping h_i. In the convolution and fully connected layers, we use ReLU as an activation function, and in the output layer, we use softmax activation function.

4.4.4. Long short-term memory networks

The long short-term memory (LSTM) architecture used for the classification of competitiveness in overlapping speech using word embeddings (WE) is illustrated in Fig. 9. The LSTM (Hochreiter and Schmidhuber, 1997) has a range of repeated modules for each time-step as in a standard recurrent neural network (RNN). At each time step, the output of the module is controlled using several gates such as *input, output*, and *forget*, which control the amount of information to get in and out to the LSTM cell. A sigmoid neural net layers and point-wise multiplication operators are used to design each gate. The LSTM memory cell is implemented using the following equations:

$$i_t = \sigma(W_{wi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{14}$$

$$f_t = \sigma(W_{wf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{15}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{wc}x_t + W_{hc}h_{t-1} + b_c) \tag{16}$$

$$o_t = \sigma(W_{wo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{17}$$

$$h_t = x_o \odot \tanh(c_t) \tag{18}$$

In the equations, σ is the point-wise logistic sigmoid function $\frac{1}{(1+e^{-x})}$, and $\odot$ is the point-wise multiplication of the two vectors; *i, f, o* and *c* are the input, forget, output gates, and memory cell vectors, respectively. The weight matrix W_* represents the weight vectors of different gates (e.g. W_{wi} is the input to input gate matrix). b_i, b_f, b_c, and b_o denote bias vectors. Therefore, the LSTM unit takes current input w at a time t and previous hidden state h_{t-1} and computes the next hidden state h_t.

In the architecture, we only output the hidden state vector of LSTM of each speaker at the last time step, i.e. $t = n$ for speaker 1 and $t = m$ for speaker 2. These state vectors (V_o for $speaker_1$ and V_p for $speaker_2$), with a fixed dimension, are then concatenated to form a vector which represents variable-length input sequences from both speakers as a single fixed-length vector. The fixed length vector is then passed to the dense layers of the architecture, and finally to the output layer. We use sigmoid activation functions in the dense layers and softmax activation function in the output layer.

4.4.5. Combined neural networks

In Fig. 10, we present the architecture of the combined neural network (FFNN-LSTM) that jointly uses acoustic (AC) and lexical (WE) information. The system takes an audio signal and transcription as input, and for each input modality we have different hidden representations, followed by a layer in which we combine the hidden representations. After the

Please cite this article as: S. Chowdhury et al., Automatic classification of speech overlaps: Feature representation and algorithms, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.12.001

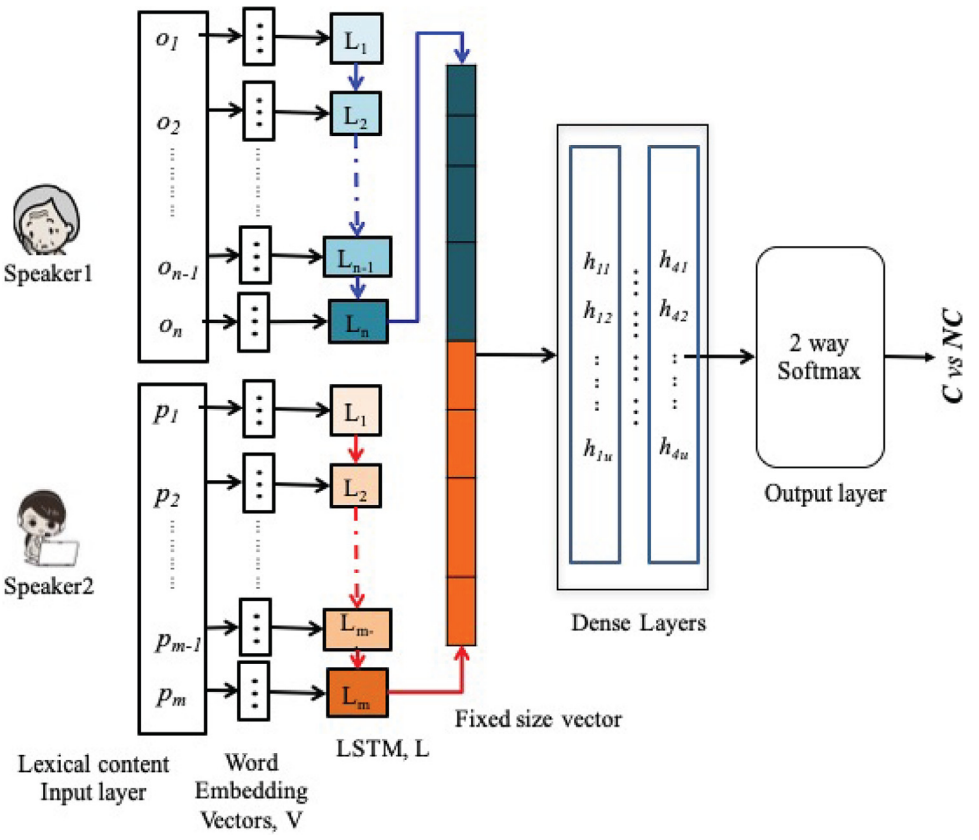

Fig. 9. The LSTM architecture for the classification of competitiveness in overlapping speech using word embedding features (WE). u represents the number of units in each hidden layer. The input is the transcription (lexical content) from each speakers ($\{o_1 \ldots o_n\}$ and $\{p_1 \ldots p_m\}$ represent the words sequences for speaker 1 and 2, respectively).

combined layer we can employ one or more hidden layers before the output layer. This architecture is heavily dependent on parameter tuning, which includes number of layers, number of hidden units in each layer, choices of activation function, such as ReLU or sigmoid; and optimization function such as SGD, Adadelta (Zeiler, 2012), Adagrad (Duchi et al., 2011), rsmprop and Adam (Kingma and Ba, 2014).

4.5. Evaluation methodology

There is no commonly agreed metric for the evaluation of overlap classification. Previous studies have used accuracy as the evaluation metric (Kurtic et al., 2010); however, due to the imbalanced class distribution, accuracy is not a good choice (Japkowicz and Shah, 2011). Consequently, in this paper, we use information retrieval metrics of precision (P), recall (R) and F_1-measure (see Eqs. (5)−(7)).

Since both classes are of interest, we compute macro-averaged F_1-measure, using macro-averaged precision (P_{av}) and recall (R_{av}) − an averages of P and R for both classes, respectively. Along with the F-measures, we are also reporting the Area under the ROC curve (AUC) (Bradley, 1997) to compare the best models using individual and combined feature sets. Model performances are compared for statistically significant differences using McNemar's test with p-value < 0.05.

5. Experiments and results

In this section, we present the results of experiments to identify the best representation of the information from speakers and their combination for the the classification of competitiveness using different classification algorithms. The results of the experiments are reported in Tables 4−6.

Please cite this article as: S. Chowdhury et al., Automatic classification of speech overlaps: Feature representation and algorithms, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.12.001

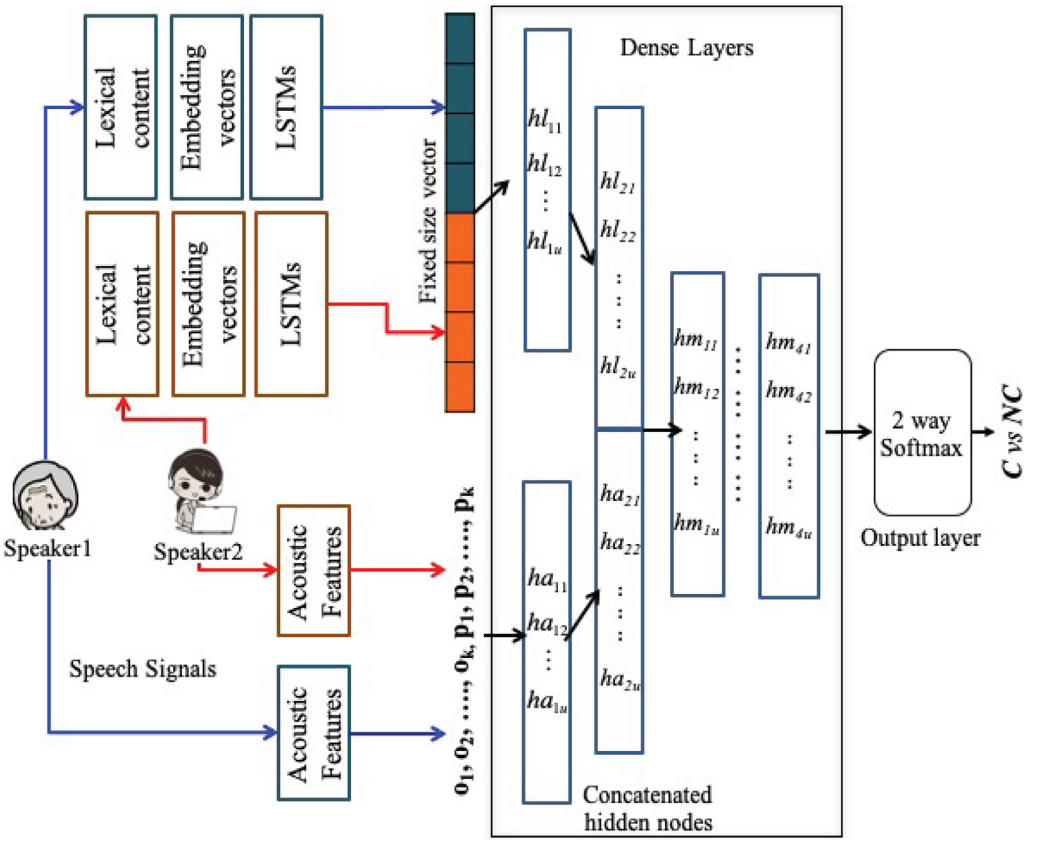

Fig. 10. The FFNN-LSTM architecture for the classification of competitiveness in overlapping speech using Word Embeddings (WE) and low-level acoustic features (AC) combined in a hidden space. u represents the number of units in each hidden layer.

5.1. Acoustic feature models

Table 4 presents the results for the classification with acoustic features using SVM and FFNN architecture. For the FFNN, we use four hidden layers with a different number of units (500, 400, 200, 50) and a sigmoid activation function in each layer. From the results, it is observed that overall performance of acoustic features improves by 2% when using FFNN compared to SVM. Closer analysis of the results reveals that the improvement is due to the increase in recall of competitive overlaps from 0.41 to 0.60. The improvement suggests that FFNN is able to capture the complexity of competitive overlaps more efficiently than a SVM model.

Table 4
Per class and the macro-averaged F_1-measures on the test set for Support Vector Machines (SVM) and Feed-Forward Neural Networks (FFNN) using acoustic features (AC). The baselines – majority and chance-level – are based on the distribution of classes in the training set.

Model	Features	C	NC	**Macro-F_1**
Baseline	Majority	0.00	**0.87**	0.44
Baseline	Chance	0.25	0.76	0.50
SVM	AC	0.44	0.85	0.64
FFNN	AC	**0.50**	0.81	**0.66**

Please cite this article as: S. Chowdhury et al., Automatic classification of speech overlaps: Feature representation and algorithms, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.12.001

Table 5
Per class and the macro-averaged F_1-measures on the test set for Support Vector Machines (SVM) and Feed-Forward Neural Networks (FFNN) using lexical features as Bag-of-Ngrams (BoN); and Convolutional Neural Networks (CNN) and Long Short-Term Memory networks (LSTM) using lexical features as Word Embeddings (WE). The baselines – majority and chance-level – are based on the distribution of classes in the training set.

Model	Features	C	NC	**Macro-F_1**
Baseline	Majority	0.00	**0.87**	0.44
Baseline	Chance	0.25	0.76	0.50
SVM	BoN	0.43	0.82	0.63
FFNN	BoN	0.36	0.83	0.59
SVM	WE	0.46	0.86	0.67
FFNN	WE	**0.54**	0.82	**0.69**
CNN	WE	0.41	0.80	0.61
LSTM	WE	0.45	0.83	0.64

Table 6
Per class and macro-averaged F_1 on the test set using feature combination. AC – low-level Acoustic features, BoN – Lexical features in bag-of-ngrams representation, WE – Lexical features in word embedding representation. The baselines – majority and chance-level – are based on the distribution of classes in the training set.

Model	Features	C	NC	**Macro-F_1**
Baseline	Majority	0.00	**0.87**	0.44
Baseline	Chance	0.25	0.76	0.50
SVM	AC+BoN	0.48	0.83	0.66
FFNN	AC+BoN	0.51	0.84	0.67
SVM	AC+WE	0.49	0.86	0.67
FFNN	AC+WE	**0.54**	0.86	**0.70**
FFNN-LSTM	AC+WE	0.50	0.81	0.68

5.2. *Lexical feature models*

Results using lexical features are reported in Table 5. The Bag-of-Ngrams representation yield better performances with SVM rather than the feed-forward network FFNN:BoN with hidden layers (200, 400, 300) and ReLU activation function. However, in comparison to performance of SVM:AC and SVM:FFNN models in Table 4, both models (SVM:BoN and FFNN:BoN) yield inferior performance. One possible explanation for such weak performance of lexical features – BoN – is the fact that the lexical patterns describing non-competitive cases form a more closed set of lexical patterns compared to competitive cases; as non-competitive overlaps frequently contain back-channelling and acknowledgments. Moreover, from an experimental point of view, BoN representation of lexical content is a very basic one.

To test whether more advanced feature representation techniques can yield better models, we experiment with word embeddings (WE) to train SVM and FFNN models (hidden layers: 500,50,50,30; activation function: sigmoid). SVM:WE and FFNN:WE models outperform SVM:BoN and FFNN:BoN by 4% and 10% macro-F_1, respectively. The results indicate that BoN representation of lexical content lacks discriminative information, which is available in averaged word embeddings.

Please cite this article as: S. Chowdhury et al., Automatic classification of speech overlaps: Feature representation and algorithms, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.12.001

Lexical content from each channel $\mathbf{t_i}$ is a vector of n_i words $(w_1^{[i]}, \ldots, w_n^{[i]})$, in which each word has a fixed d-dimensional (d=300) representation in the word embedding space $\mathbf{E}$ (extracted from the model described in Section 4.3.2). Therefore, a speaker's content is represented as a $d \times n_i$ matrix $\mathbf{V_i}$, where column j is a word $w_j^{[i]}$. For SVM and FFNN, we average these column-wise matrices $\mathbf{V_i}$ to get the lexical feature represented as a fixed d-dimensional vector $\mathbf{h_i}$ in the word embedding space $\mathbf{E}$; and then concatenated the vectors to obtain a final representation of dimension $2 \times d = 600$.

Word embeddings representation already outperforms bag-of-ngrams representation; however, it might be the case that we lose predictive information by averaging the embeddings. The hypothesis is tested by training CNN model described in Section 4.4 using word-embedding representation.

The performance of the CNN architecture (four hidden layers with units u (600, 300, 300, 100)) with word embeddings is reported in Table 5. While the performance of CNN:WE is higher than FFNN:BoN by significant $\approx 2\%$, compared to other models, the results are significantly lower. The performance of CNN architecture is heavily dependent on the parameter tuning − number of hidden layers and their sizes, number of filters, learning rate, etc. In this experiment, we have experimented only with the number of hidden layers: h is tuned for $h = \{2, 3, 4\}$ layers with a fixed list of neurons. Thus, there is space for further improvement using a different architecture and a different set of parameters.

To understand the long term contextual dependencies of word sequences for classifying competitive overlaps, we experiment with an LSTM architecture (see Section 4.4). In this architecture, we use a dimension of 256 for the state vectors of each speaker (V_o and V_p), and concatenate them to form V, where $V = V_o + V_p$. The vector V is then passed to the dense layer containing five hidden layers with units 1500, 300, 300, 800, and 50, respectively, and a sigmoid activation function.

The LSTM:WE model outperforms all the models with BoN representation, and CNN:WE model. However, it does not outperform SVM:WE and FFNN:WE models. Overall, similar to acoustic feature experiments, FFNN:WE model significantly outperforms all other algorithms and lexical content representations.

5.3. Combined feature models

The experiment on the combination of lexical and acoustic features comprise of training SVM and FFNN algorithms on the combinations of AC with BoN and WE representations (AC+BoN and AC+WE, respectively). The FFNN:AC+BoN architecture has 4 hidden layers with 500, 300, 200, 50 units each and a ReLU activation function; whereas FFNN:AC+WE architecture has 4 hidden layers with 500, 50, 30, 30 units each and a sigmoid activation function. The results are reported in Table 6.

Combination of acoustic features (AC) with BoN representation yields models superior in performance than each representation alone. However, combination of acoustic features with WE significantly outperforms the BoN models. The pattern is the same for both SVM and FFNN algorithms; even though FFNN:AC+BoN and SVM:AC+WE have the same performance. Overall, the FFNN:AC+WE architecture yields the best performing models that combine acoustic and lexical features linearly. Due to the imbalanced distribution of labels, the majority baseline achieves the highest F_1 score for the non-competitive (NC) overlap class (0.87); however, the difference from the FFNN:AC+WE on the same class (0.86) is not statistically significant.

The FFNN-LSTM architecture combines lexical and acoustic features in hidden space. The performance of the model is inferior to the linear combination with FFNN. Nevertheless, the model achieves the best recall (0.66) for competitive overlap category. The complexity of the model with respect to the parameter tuning must be acknowledged, as it is a major factor determining the performance. The FFNN-LSTM architecture presented in Fig. 10 consists of hidden layers with units in lexical module as (500,200), acoustic module as (900,500) and merged module as (50,800,800,50). The numbers report performance after tuning for the number of neurons in the hidden layers with a fixed set, S={50, 100, 200, 300, 500, 800, 1500}, while keeping the number of hidden layers in the architecture fixed.

6. Discussion

Modeling the discourse of overlapping speech is crucial for uncovering human interaction dynamics, improving naturalness of spoken dialog systems, and extraction of behavioral cues for downstream affective computing models.

Please cite this article as: S. Chowdhury et al., Automatic classification of speech overlaps: Feature representation and algorithms, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.12.001

In this study, we have designed and evaluated different linear and non-linear computational models for classifying discourse of overlaps as competitive and non-competitive, using a large ecological conversational data. While doing so, we have experimented with different input representations from both speakers (acoustic, lexical and the surrounding context).

Unlike all the previous research, mentioned in Section 2, this study explores the use of lexical content for overlap discourse classification using both bag-of-ngrams (BoN) and word embedding (WE) representations, with Support Vector Machines (SVM) and various neural network models − feed-forward (FFNN), convolutional (CNN), and long short-term memory (LSTM) architectures. The acoustic cues (AC) are represented as high-dimensional low-level descriptors projected on statistical functionals and modeled using both SVM and feed-forward (FFNN) networks.

The comparative analysis indicates that lexical content is a powerful and a very competitive feature for classifying the intention behind the overlapping speech. Even though most previous studies have focused on acoustic features, especially prosodic features, our study suggests that using a proper representation of lexical content, such as averaged word embedding features of overlapper and overlappee, a performance of $F_1 = 0.69$ can be achieved in an unbalanced data sets with a small feature vector dimension (compared to the best acoustic model with $F_1 = 0.66$). The importance of the lexical content as a feature is also reflected from the observation in Section 3.3, where we have observed that the choice of words while initiating an overlap can indicate the intention of the speaker.

Our investigation further suggests that a simple Feed-Forward Neural Network model is capable of extracting discriminative cues from acoustic and lexical embedding features. Moreover, it is able to combine the feature to outperform other more complex models with the $F_1 - 0.70$ (see Fig. 11).

To decouple classifier performance from class skewness, we additionally compute area under ROC curve (*AUC*) (see Fig. 12) for FFNN:AC, FFNN:WE and FFNN:AC+WE models.

The FFNN:AC+WE model that combine acoustic and lexical features has an *AUC* of 0.81, whereas for the FFNN:AC and FFNN:WE models the *AUC* are 0.75 and 0.80, respectively. Consequently, we conclude that FFNN:AC+WE model can discriminate competitive and non-competitive overlaps better that individual model representations.

This study shows the importance of the representation of lexical content. The experiments open new research questions for future work, such as the effect of combining different representations of lexical content together with the goal of overlap classification. The acoustic feature representation depends on the extraction of low-level descriptors, one future research direction to explore is the use of raw audio signals to design end-to-end systems that are completely void of any hand-crafted feature designs.

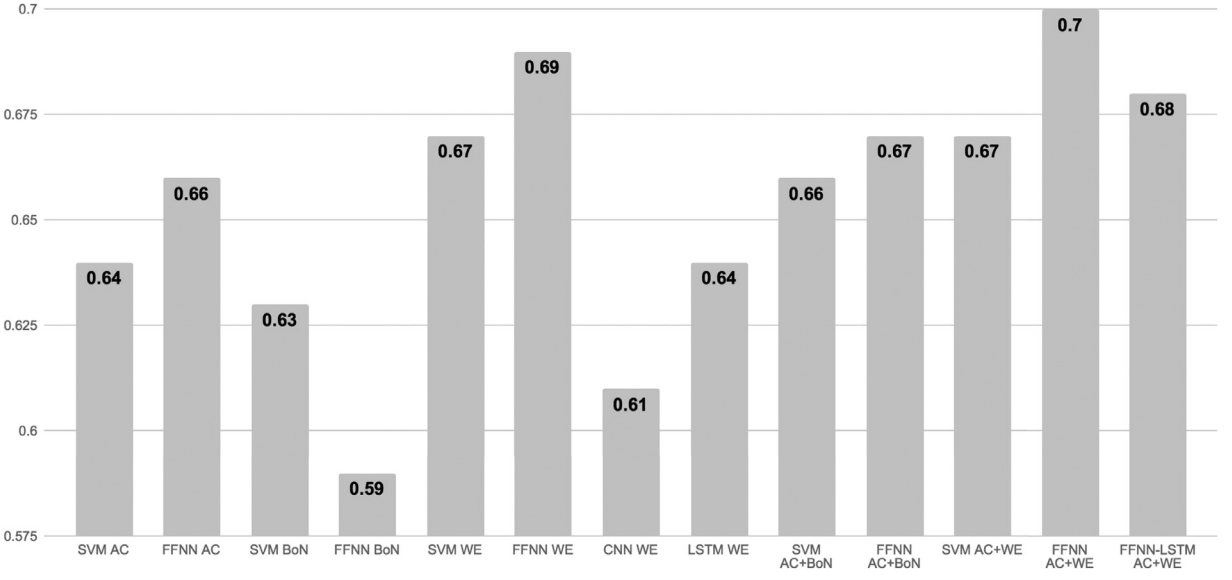

Fig. 11. Summary of the results as macro-F_1 on the test set. Support Vector Machines (SVM), Feed-Forward Neural Networks (FFNN), Convolutional Neural Networks (CNN) and Long Short-Term Memory networks (LSTM) using Acoustic (AC), Lexical Bag-of-Ngrams (BoN), and Lexical Word Embeddings (WE) features and their combinations (AC+BoN and AC+WE).

Please cite this article as: S. Chowdhury et al., Automatic classification of speech overlaps: Feature representation and algorithms, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.12.001

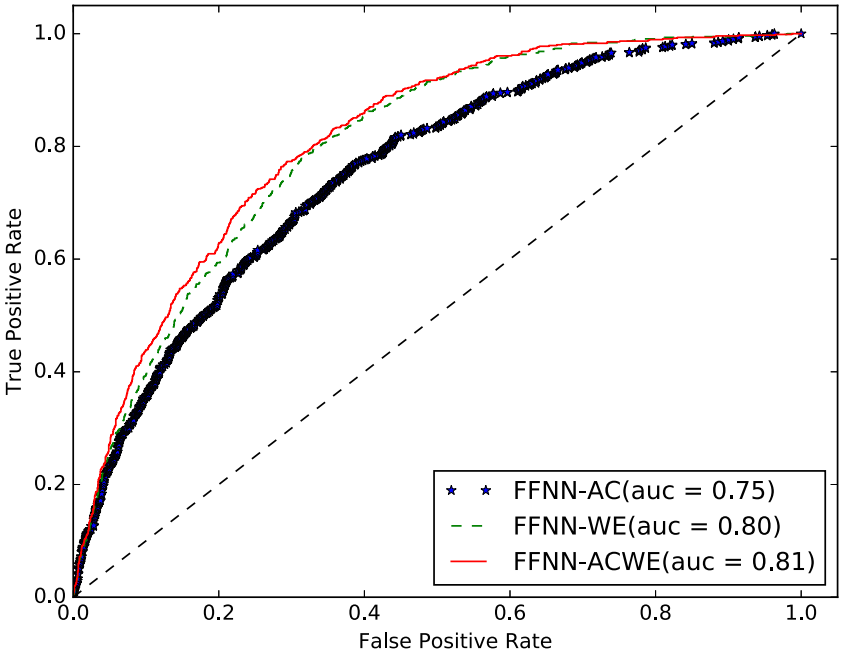

Fig. 12. Receiver Operating Characteristic (ROC) curve and area under the curve (AUC) for Feed-Forward Neural Network (FFNN) using acoustic (AC) and lexical (WE) features and their combination (AC+WE).

7. Applications of overlap classification

The automatic classification of speech overlaps can give us an insight into speakers' dominance and aggression (West, 1979), empathic behavior and basic emotions (Alam et al., 2016), behavioral summarization of a conversation (Stepanov et al., 2015), among others. Last but not least, the ability to qualify automatically a speech overlap as competitive or not is crucial for action selection (and response generation) in the most advanced human−machine conversational systems. In this section, we describe an application of automatic classification of speech overlaps to the characterization of the user experience in spoken conversations.

In Chowdhury et al. (2016a), the authors predict user satisfaction from turn-taking features that include the overlap categories. While the authors' goal was to predict user satisfaction, our goal is to identify the contribution of automatic overlap classifier for the task. Consequently, different from Chowdhury et al. (2016a), here we additionally present user satisfaction performance using automatic overlap features only, and compare it to the other feature sets (reported in Chowdhury et al., 2016a). We first briefly summarize the study of Chowdhury et al. (2016a) and then describe the novel extensions.

A brief review. The study of Chowdhury et al. (2016a) investigates the importance of different turn-taking features, including percentage of competitive and non-competitive overlaps, for predicting user satisfaction states as positive, negative, or neutral. The pipeline of the system for predicting user satisfaction from speakers' audio signals is presented in the Fig. 13. The system first performs Automatic Speech Recognition (ASR) on each speaker channel, which detects speech vs. non-speech segments prior to recognition. The output of ASR is a time-aligned word transcription, which is used to extract turn-taking features. The *Turn-Taking Feature Extraction* component consists of several steps. *Turn segmentation and labeling* step aligns the inter-pausal units (IPUs) − segments of consecutive

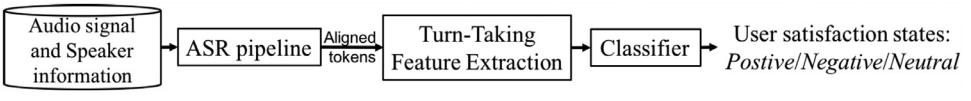

Fig. 13. Architecture of the system for automatic prediction of user satisfaction (as negative, positive and neutral).

Please cite this article as: S. Chowdhury et al., Automatic classification of speech overlaps: Feature representation and algorithms, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.12.001

Table 7
F_1 measures for automatic prediction of user satisfaction state using lexical (L), prosodic (P), turn-taking (T) and overlap ($\widetilde{O}$) features.

Experiments	Pos	Neg	Neu	Overall
Lexical (L) (Chowdhury et al., 2016a)	0.44	0.58	0.35	0.48
Prosodic (P) (Chowdhury et al., 2016a)	0.33	0.32	0.52	0.40
Turn-Taking (T) (Chowdhury et al., 2016a)	0.55	0.52	0.63	0.58
Overlap ($\widetilde{O}$) $\widetilde{O} \subset T$	0.45	0.36	0.55	0.50

tokens with no less than 50 ms gaps in between − from each channel and labels them as non-overlapping turns (from either agent or customer), overlaps, or different varieties of silences such as pauses, gaps, etc. The generated sequence is then passed to the *Discourse labeling* step.

The discourse module includes an *Overlap classifier* (SVM:AC model Chowdhury et al., 2015b which is used as one of the baselines for this study); and a classifier that categorizes turns into *Dialog Act Semantic Dimensions* (Chowdhury et al., 2016b) such as *feedback, social obligation, general purpose dialog acts*, etc. Using the sequence of overlap labels and a set of dialog act dimensions, the system generate various turn-taking features that include speaker and conversation level features such as participation equality, dialog act dimension rates with respect to speakers' speech duration, percentages/median durations of different turn types including overlaps, silences and non-overlapping turns, etc.

Apart from turn-taking features (T), the authors also experiment with basic prosodic and lexical features. The prosodic features (P) were extracted using both speaker channels and linearly merged to form the feature vector. The prosodic feature set includes pitch, loudness, and voice-probability together with their derivatives, which are projected onto statistical functionals. Lexical features (L), on the other hand, are extracted from both speaker channel using automatic transcription and represented as bag-of-word or frequency-based tf-idf vectors. Sequential Minimal Optimization (SMO) implementation of Support Vector Machines was used to train the classification models.

An extension of the study. As a contribution to the summarized study, we include the performance of the classifier trained on automatically extracted overlap features. The overlap feature set ($\widetilde{O}$) is a subset of turn-taking features (T) and includes: percentage of overlaps, percentage of C and NC with respect to the total duration of overlaps, median duration of C and NC overlaps in the conversation, probabilities of each speakers' turn after a competitive/non-competitive event, and probabilities of C/NC events after each speakers' turns.

The performance of the prediction models on each of the feature sets T, P, L and $\widetilde{O}$ are reported in Table 7. The overlap feature set ($\widetilde{O}$) outperforms lexical (L) and prosodic (P) feature sets overall and Pos and Neu classes. Whereas for the Neg user-satisfaction class, $\widetilde{O}$ features only outperform the prosodic feature set. Overall, the results indicate that overlap categories are more predictive of user satisfaction than prosody or lexical content.

8. Conclusion

In this study, we have designed, explored and evaluated different linear and non-linear computational models for classifying discourse of overlaps as competitive *vs.* non-competitive. While doing so, we have experimented with different input representations (acoustic and lexical and context) from both speakers' channel and evaluated them on unscripted and in-the-wild spoken conversations. The designed neural network architectures for speech overlap classification are able to efficiently combine the information from both lexical and acoustic modalities. Both linear (FFNN:AC+WE) and hidden-space (FFNN-LSTM:AC+WE) feature combinations significantly outperform all the other models. The study suggests that the lexical content of the overlapping speech, with appropriate feature representation, is a powerful tool for classifying the intent behind the overlap events. The results also indicate that the FFNN architecture is able to capture the complexity of competitive overlaps while also being effective in distinguishing non-competitive overlaps, as it achieves the best overall F_1-measure of 0.70.

In future work, we plan to further explore the lexical features, their representation and combination, as very few studies on overlaps have considered lexical information before. Whereas for acoustic features, we plan to explore the utility of raw audio signals to train deep neural networks and harvest their power in unsupervised feature extraction.

Please cite this article as: S. Chowdhury et al., Automatic classification of speech overlaps: Feature representation and algorithms, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.12.001

Acknowledgments

The research leading to these results has received funding from the European Union − Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 610916 − SENSEI − http://www.sensei-conversation.eu/.

Supplementary material

Supplementary material associated with this article can be found in the online version at 10.1016/j.csl.2018.12.001.

References

Alam, F., Chowdhury, S.A., Danieli, M., Riccardi, G., 2016. How interlocutors coordinate with each other within emotional segments? In: Proceedings of the 2016 International Conference on Computational Linguistics, COLING.
Alam, F., Danieli, M., Riccardi, G., 2018. Annotating and modeling empathy in spoken conversations. Comput. Speech Lang. 50, 40–61.
Alam, F., Danieli, M., Riccardi, G., 2019. Automatic Labeling Affective Scenes in Spoken Conversations. Springer International Publishing, pp. 109–130.
Alam, F., Riccardi, G., 2013. Comparative study of speaker personality traits recognition in conversational and broadcast news speech. In: Proceedings of the 2013 INTERSPEECH, pp. 2851–2855.
Baroni, M., Dinu, G., Kruszewski, G., 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In: Proceedings of the 2014 Association for Computational Linguistics, ACL, 1, pp. 238–247.
Bazzanella, C., 1996. Repetition in Dialogue. 11. Walter de Gruyter.
Beňuš, Š., Gravano, A., Hirschberg, J., 2011. Pragmatic aspects of temporal accommodation in turn-taking. J. Pragmat. 43 (12), 3001–3027.
Bian, J., Gao, B., Liu, T.-Y., 2014. Knowledge-powered deep learning for word embedding. Machine Learning and Knowledge Discovery in Databases. Springer, pp. 132–148.
Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 30 (7), 1145–1159.
Carletta, J., 1996. Assessing agreement on classification tasks: the kappa statistic. Comput. Linguist. 22 (2), 249–254.
Chowdhury, S.A., 2017. Computational Modeling of Turn-Taking Dynamics in Spoken Conversations. University of Trento.
Chowdhury, S.A., Danieli, M., Riccardi, G., 2015. Annotating and categorizing competition in overlap speech. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, ICASSP. IEEE.
Chowdhury, S.A., Danieli, M., Riccardi, G., 2015. The role of speakers and context in classifying competition in overlapping speech. In: Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association.
Chowdhury, S.A., Riccardi, G., 2017. A deep learning approach to modeling competitiveness in spoken conversation. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE.
Chowdhury, S.A., Riccardi, G., Alam, F., 2014. Unsupervised recognition and clustering of speech overlaps in spoken conversations. In: Proceedings of the Workshop on Speech, Language and Audio in Multimedia.
Chowdhury, S.A., Stepanov, E., Riccardi, G., 2016a. Predicting user satisfaction from turn-taking in spoken conversations. In: Proceedings of the 2016 INTERSPEECH.
Chowdhury, S.A., Stepanov, E.A., Riccardi, G., 2016b. Transfer of corpus-specific dialogue act annotation to iso standard: is it worth it? In: Proceedings of the 2016 International Conference on Language Resources and Evaluation, LREC.
Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20 (1), 37–46.
Danieli, M., Bazzanella, C., SpA, L., Torino, I., 2002. Linguistic markers in coming to understanding. In: Proceedings of the VIII Meeting of AIIA (Associazione Italiana Intelligenza Artificiale), AIIA 2002, pp. 10–13.
Danieli, M., Riccardi, G., Alam, F., 2015. Emotion unfolding and affective scenes: a case study in spoken conversations. In: Proceedings of the Emotion Representations and Modelling for Companion Systems (ERM4CT) 2015. ICMI.
Duchi, J., Hazan, E., Singer, Y., 2011. Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. 12 (Jul), 2121–2159.
Duncan, S., 1972. Some signals and rules for taking speaking turns in conversations. J. Personal. Soc. Psychol. 23 (2), 283.
Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., Bengio, S., 2010. Why does unsupervised pre-training help deep learning? JMLR 11 (Feb), 625–660.
Ericsson, K.A., Simon, H.A., 1984. Protocol Analysis. MIT-Press.
Eyben, F., Weninger, F., Gross, F., Schuller, B., 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In: Proceedings of the Twenty-First ACM International Conference on Multimedia. ACM, pp. 835–838.
Fleiss, J.L., 1975. Measuring agreement between two judges on the presence or absence of a trait. Biometrics 651–659.
French, P., Local, J., 1983. Turn-competitive incomings. J. Pragmat. 7 (1), 17–38.
Goldberg, J.A., 1990. Interrupting the discourse on interruptions: an analysis in terms of relationally neutral, power-and rapport-oriented acts. J. Pragmat. 14 (6), 883–903.
Gravano, A., Hirschberg, J., 2011. Turn-taking cues in task-oriented dialogue. Comput. Speech Lang. 25 (3), 601–634.
Gravano, A., Hirschberg, J., 2012. A corpus-based study of interruptions in spoken dialogue. In: Proceedings of the 2012 INTERSPEECH.

Please cite this article as: S. Chowdhury et al., Automatic classification of speech overlaps: Feature representation and algorithms, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.12.001

Graves, A., Mohamed, A.-r., Hinton, G., 2013. Speech recognition with deep recurrent neural networks. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP. IEEE, pp. 6645–6649.
Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software: an update. ACM SIGKDD Explor. Newsl. 11 (1), 10–18.
Hammarberg, B., Fritzell, B., Gaufin, J., Sundberg, J., Wedin, L., 1980. Perceptual and acoustic correlates of abnormal voice qualities. Acta Oto-Laryngol. 90 (1−6), 441–451.
Heldner, M., Edlund, J., 2010. Pauses, gaps and overlaps in conversations. J. Phonet. 38 (4), 555–568.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.
Hripcsak, G., Rothschild, A.S., 2005. Agreement, the *F*-measure, and reliability in information retrieval. J. Am. Med. Inf. Assoc. 12 (3), 296–298.
Japkowicz, N., Shah, M., 2011. Evaluating Learning Algorithms. Cambridge University Press.
Jefferson, G., 1982. Two Explorations of the Organization of Overlapping Talk in Conversation. Tilburg University, Department of Language and Literature.
Jefferson, G., 2004. A sketch of some orderly aspects of overlap in natural conversation. Pragmat. Beyond New Ser. 125, 43–62.
Joachims, T., 1998. Text categorization with support vector machines: learning with many relevant features. In: Nédellec, C., Rouveirol, C. (Eds.), Proceedings of the European Conference on Machine Learning. ECML-98. Lecture Notes in Computer Science. 1398, Springer Berlin Heidelberg, pp. 137–142. doi: 10.1007/BFb0026683.
Kingma, D., Ba, J., 2014. Adam: A Method for Stochastic Optimization. arXiv preprint arXiv:1412.6980.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Proceedings of the 2012 Advances in Neural Information Processing Systems, pp. 1097–1105.
Kurtic, E., Brown, G.J., Wells, B., 2010. Resources for turn competition in overlap in multi-party conversations: speech rate, pausing and duration. In: Proceedings of the 2010 INTERSPEECH, pp. 2550–2553.
Kurtić, E., Brown, G.J., Wells, B., 2013. Resources for turn competition in overlapping talk. Speech Commun. 55 (5), 721–743.
Lee, C.-C., Lee, S., Narayanan, S.S., 2008. An analysis of multimodal cues of interruption in dyadic spoken interactions. In: Proceedings of the 2008 INTERSPEECH, pp. 1678–1681.
Lee, C.-C., Narayanan, S., 2010. Predicting interruptions in dyadic spoken interactions. In: Proceedings of the 2010 International Conference on Acoustics, Speech, and Signal Processing, ICASSP. IEEE, pp. 5250–5253.
Liu, S., Yang, N., Li, M., Zhou, M., 2014. A recursive recurrent neural network for statistical machine translation. In: Proceedings of the 2014 Association for Computational Linguistics, ACL, pp. 1491–1500.
Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. In: Proceedings of the 2013 International Conference on Learning Representations. Available as arXiv preprint:1301.3781.
Mikolov, T., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. Adv. Neural Inf. Process. Syst.
Nickolls, J., Dally, W.J., 2010. The GPU computing era. Micro, IEEE 30 (2), 56–69.
Oertel, C., Wlodarczak, M., Tarasov, A., Campbell, N., Wagner, P., 2012. Context cues for classification of competitive and collaborative overlaps. In: Proceedings of the 2012 Speech Prosody.
Řehůřek, R., Sojka, P., 2010. Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. ELRA, Valletta, Malta, pp. 45–50.
Sacks, H., Schegloff, E.A., Jefferson, G., 1974. A simplest systematics for the organization of turn-taking for conversation. Language 696–735.
Schegloff, E.A., 2000. Overlapping talk and the organization of turn-taking for conversation. Lang. Soc. 29 (01), 1–63.
Schegloff, E.A., 2001. Accounts of conduct in interaction: interruption, overlap, and turn-taking. Handbook of Sociological Theory. Springer, pp. 287–321.
Schuller, B., Batliner, A., Steidl, S., Seppi, D., 2011. Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. Speech Commun. 53 (9), 1062–1087.
Shriberg, E., Stolcke, A., Baron, D., 2001a. Can prosody aid the automatic processing of multi-party meetings? evidence from predicting punctuation, disfuencies, and overlapping speech. In: Proceedings of the ISCA Tutorial and Research Workshop (ITRW) on Prosody in Speech Recognition and Understanding.
Shriberg, E., Stolcke, A., Baron, D., 2001b. Observations on overlap: findings and implications for automatic processing of multi-party conversation. In: Proceedings of the 2001 INTERSPEECH, pp. 1359–1362.
Stepanov, E., Favre, B., Alam, F., Chowdhury, S., Singla, K., Trione, J., Béchet, F., Riccardi, G., 2015. Automatic summarization of call-center conversations. In: Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015).
Truong, K.P., 2013. Classification of cooperative and competitive overlaps in speech using cues from the context, overlapper, and overlappee. In: Proceedings of the 2013 INTERSPEECH, pp. 1404–1408.
Wells, B., Macfarlane, S., 1998. Prosody as an interactional resource: turn-projection and overlap. Lang. Speech 41 (3−4), 265–294.
West, C., 1979. Against our will: male interruptions of females in cross-sex conversation*. Ann. N. Y. Acad. Sci. 327 (1), 81–96.
Zeiler, M. D., 2012. Adadelta: An Adaptive Learning Rate Method. arXiv preprint:1212.5701.

Please cite this article as: S. Chowdhury et al., Automatic classification of speech overlaps: Feature representation and algorithms, Computer Speech & Language (2018), https://doi.org/10.1016/j.csl.2018.12.001