Unsupervised Recognition and Clustering of Speech Overlaps in Spoken Conversations

Shammur Absar Chowdhury, Giuseppe Riccardi, Firoj Alam

Department of Information Engineering and Computer Science, University of Trento, Italy
{sachowdhury, riccardi, alam}@disi.unitn.it

Abstract

We are interested in understanding speech overlaps and their function in human conversations. Previous studies on speech overlaps have relied on supervised methods, small corpora and controlled conversations. The characterization of overlaps based on timing, semantic and discourse function requires an analysis over a very large feature space. In this study, the corpus of overlapped speech segments was automatically extracted from human-human spoken conversations using a large vocabulary Automatic Speech Recognizer (ASR) and a turn segmenter. Each overlap instance is automatically projected onto a high dimensional space of acoustic and lexical features. Then, we used unsupervised clustering to find the distinct and well-separated clusters in terms of acoustic and lexical features. We have evaluated recognition and clustering algorithms over a large set of real human-human spoken conversations. The clusters have been comparatively evaluated in terms of feature distributions and their contribution to the automatic classification of the clusters.

Index Terms: Overlapping Speech, Human Conversation, Discourse, Language understanding

1. Introduction

Over the last forty years, the study of human-human conversations has attracted interest from researchers in the fields of sociology, computational linguistics and speech science. In the early seventies, Sacks et al. [1] studied human conversations and found that the transition from one speaker to another should occur with minimum overlap or gap, the two turn-taking signals. A recent study [2] suggests that the timing of turn taking with overlaps and silence is less precise. This study also states that overlap is in fact a frequent phenomenon and is much more than just a turn-taking signal.

Overlaps represent a speaker’s behavior and intention in a regular conversation. Some overlaps indicate support for the current speaker to continue speaking while others are intended to break the flow of the conversation or to compete for turns [3]. The former class of overlaps is referred to as non-competitive and the latter, competitive.

Most previous studies have been conducted on controlled meeting corpora [4]. However, in this study, we focus on spoken conversations collected from a call center. Thus, the aim of this study is to analyze the natural distinctive statistical patterns that describe overlaps using unsupervised techniques.

We have analyzed acoustic and lexical features that discriminate individual clusters and compared them with the characteristics of features mentioned in previous literature on distinguishing overlaps.

In contrast with that of previous studies, the contribution of this study differs in a number of ways:

- Designing the corpus of speech overlaps using an automated approach.
- Extraction of a large set of acoustic and lexical features.
- Investigation of speech overlaps using unsupervised clustering.
- Analysis of discriminative characteristics of speech overlaps over acoustic and lexical features.

This paper is organized as follows. An overview of previous studies of overlaps is given in Section 2, followed by a description of data preparation procedures in Section 3. In Section 4, we discuss the experimental methodology used in this study. We present an analysis of our findings in Section 5 and provide conclusions in Section 6.

2. Related Work

Speech overlaps have been categorized in terms of speakers’ (non) competitiveness. Different verbal and non-verbal predictors have been proposed to indicate willingness to compete. According to [5], position of the overlap onset is an important feature along with some temporal features related to the position of overlaps. In [6], it is mentioned that speech rate, cut-offs and repetition are also important features used by speakers in competitive overlaps. In [3], authors observed that precise location does not describe a competitive overlap but the phonetic design plays the role. Later studies in [8, 9] supported this hypothesis. In [9], authors also stated that competitive overlaps are usually high in pitch and amplitude to grab the attention from the current speaker. Classification of competitive and non-competitive overlaps was studied in [10] using decision tree and they found that duration is the most distinguishing feature. The findings in [4,7] suggest that $f_0$ is the most common feature and is being higher in competitive overlaps.

3. Data

In this study, we have analyzed a corpus collected from a call center, which contains inbound Italian phone conversations between agents and customers. Each conversation was recorded over two channels at a sample rate of 8 kHz, 16bit speech samples and has an average duration of 395 seconds.

We selected 515 conversations based on maximum duration of the overlapped segments. The overlapped segments were detected using start and end time of each speaker’s turn and for each word unit within that turn. To get each speaker’s turns we passed the conversation to an automatic turn segmenter [11] followed by a large vocabulary Italian ASR to get automatic transcription from the corresponding turns. Then, we detected overlapping turns, where each turn has an alignment between the automated word level transcriptions and the speech recording. Using the overlapping turns, we also extracted words that are overlapped within the turns. Then, we extracted the speech signal from our overlapping speech instances using the start time of the first word in the overlap to the end time of the last word. Therefore, the overlap segment has two channels per
conversation. Following this approach, we have extracted 25132 instances of overlaps (average duration of 0.52s) from 515 conversations, where the duration of overlap is 3 hours and 38 minutes and total the duration of speeching time is 41 hours and 52 minutes. In the process of detecting overlapping segments, our approach differs from previous study [29], where they designed a classifier to detect speech overlap.

The ASR system was designed using a subset of the data, with around 100 hours of conversations, and a lexicon of size ~18K. The training data of the ASR was completely independent from the data set that was used in the study.

The MFCC features were spliced, taking 3 frames from each side of the current frame, followed by Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) feature-space transformations to reduce the feature space. For obtaining the best acoustic model using Kaldi [12], we used speaker adaptive training (SAT) followed by Maximum Mutual Information (MMI) training. The Word Error Rate (WER) for our ASR system is 31.78% using Kaldi [12], we used speaker adaptive training (SAT) feature set, we performed cluster evaluation and analyzed the only and acoustic merged with lexical features. For each designed three feature set combinations: acoustic only, lexical only and acoustic merged with lexical features. For each feature set, we performed cluster evaluation and analyzed the features based on the clustered output using feature ranking. Finally, we studied whether there was loss of any information due to the reduction of the feature dimension.

These low-level features were extracted with approximately 100 frames per second, with 25 milliseconds per frame. The 39 low-level features include frame energy, loudness, mel-frequency cepstral coefficients (MFCC1-12), voice quality, fundamental frequency (f0), exponentially smoothed f0-envelope, jitter-local, differential of jitter, shimmer-local, logarithmic harmonics-to-noise ratio (HNR) computed from auto-correlation, zero crossing rate of time signal, formant frequencies (f0-f5) and spectral features with different bands (0-250Hz, 0-650Hz, 250-650Hz, 1-4kHz), roll-off points (25%, 50%, 70%, 90%), centroid, flux, max-position and min-position. Delta and acceleration coefficients of these features have also been extracted. These low-level acoustic features were then projected onto 24 statistical functionals, which included range, absolute position of max and min, linear and quadratic regression coefficients and their corresponding approximation errors, moments-centroid, variance, standard deviation, skewness, kurtosis, zero crossing rate, peaks, mean peak distance, mean peak, geometric mean of non-zero values, and number of non-zeros.

As mentioned earlier, the overlap segment’s components appear in two channels. Therefore, we extracted the same number of features from each channel. The size of the feature vector in a single channel is computed as: \((39 + 39 + 39) \times 24 + 2 = 2808\). A total of 5616 features were obtained after merging the feature vectors from both channels.

4.1.2. Lexical features

Lexical features were extracted from automatic transcription using the ASR explained in Section 3. The lexical features were transformed into a bag-of-words (vector space model) [16]. The idea of this approach is to represent the words into numeric features. For this study, we extracted bigram features and selected the top 2000 frequent features to reduce features dimension. The frequencies in the feature vector were then transformed into tf-idf values – the product of the logarithmic term frequency (tf) and inverse document frequency (idf).

4.1.3. Feature Combination

Although feature combination has been widely used in other speech processing tasks, its relative contribution greatly varies depending on the data and experiments. For this study, we also analyzed the contribution of feature combination. As shown in Figure 1, after extracting acoustic and lexical features we merged the feature vectors into a single vector and then used that for clustering.

4.1.4. Dimensionality Reduction

Since the complexity of any pattern recognition algorithm also depends on the number of features, we reduced the feature space to limit the complexity and number of free parameters. A typical approach for feature reduction is to map higher dimensional feature spaces into lower dimensional spaces, while maintaining as much of the information as possible. In our study, we have used Principal Component Analysis (PCA), which is one of the fundamental feature reduction methods. After transforming the feature space using PCA, the usual approach is to take the leading p components that explain the data with 95% variance [17]. However, as a baseline study we took the leading p components with 99% variance. The value of p varies for different features sets. Hence, we reduced 63% acoustic, 11% lexical and 59%
acoustic+lexical features. The reason for obtaining a minimal reduction with lexical features is the weak correlation with feature dimensions and sparseness.

4.2. Clustering

To find well-separated clusters of speech overlaps in our dataset, we used K-means [18] where data points are classified as belonging to one of K groups. For reproducibility and transferability, we used Weka’s implementation [19]. Members of the clusters are determined by comparing the data point with each group’s centroid and assigning it to the nearest one. The reason for choosing K-means is that it is highly recommended for large datasets [20, 21] and is one of the simplest methods. However, one of main limitations of K-means is in choosing the value of K in prior. Therefore, we used cascaded K-means, which uses Calinski-Harabasz [22] criterion to determine the best value of K that represents the dataset.

For each value of K, cascaded K-means calculates the between-group dispersion, BGSS, within-cluster sum of squares, WGSS, and Calinski-Harabasz (CH) value or index, using Equations [1-3].

\[
BGSS = \sum_{k=1}^{K} n_k \| G(k) - G \|^2
\] (1)

\[
WGSS = \sum_{k=1}^{K} \sum_{i \in C_k} \| M_i^{(k)} - G^{(k)} \|^2
\] (2)

\[
CH = \frac{(N-K) \times BGSS}{(K-1) \times WGSS}
\] (3)

K is the number of clusters, N is number of observations, \(G^{(k)}\) is the barycenter of cluster \(C_k\), G is the barycenter of the entire dataset, \(n_k\) is the number of elements in the \(C_k\), \(I_k\) is the set of the indices of the observations belonging to \(C_k\), and \(M_i\) is the \(i^{th}\) observation of element in \(C_k\).

Figure 2 shows the values of CH corresponding to the number of clusters K. The results of our experiments are shown in Table 1.

The optimal number of K using acoustic and acoustic+lexical feature sets is 2, as can be seen in Figure 2. The clustering difference between the two feature sets is minimal. The optimal number for k, which we obtained using lexical features, is 4. The CH value for the cluster is significantly smaller compared to the CH values with acoustic features. The minimal separability of lexical features could be due to the sparseness and recognition error of the ASR.

We applied PCA feature reduction method to the acoustic and acoustic+lexical feature sets and with reduced dimensions we obtained 2 clusters for each set. We then calculated the cluster agreement between the feature sets using kappa (κ) statistics [23]. We found that the agreement between the original and reduced dimensions is fairly reasonable. The agreements for the acoustic and acoustic+lexical feature sets are 92% and 91%, respectively. This indicates that feature reduction is important in reducing computational cost with minimal loss of information.

To check the validity of the clusters using cascaded K-means, we used another well-known clustering algorithm – Spectral Clustering [24, 25]. Using this algorithm, we also found 2 clusters. Then, we compared the clusters generated by these two clustering algorithms for acoustic and acoustic+lexical feature sets using κ measure. The agreements between the two algorithms on acoustic and acoustic+lexical feature sets are 90% and 87%, respectively. Detailed results of Spectral Clustering algorithm have not been included in this paper, in favor of brevity.

5. Analysis and Discussions

We analyzed different features based on the cluster decision of acoustic features, where cluster 0 (C0) and cluster 1 (C1) contain 37% and 63% of overlapping instances, respectively. The members of the clusters were analyzed using duration distribution of speech overlaps and top-ranked acoustic features. Based on this cluster decision, we extracted and analyzed lexical features. In doing so, we compared our observations with those of previous studies to determine whether our clusters resembled competitive or non-competitive overlaps.

5.1. Duration Distribution

Figure 3 shows the distribution of overlap durations for C0 and C1. It can be seen that C1 contains instances of overlaps with short durations whereas C0 has instances with comparatively long durations. The authors in [10] and [26] state that non-competitive overlaps tend to be shorter and resolved soon after the second speaker has recognized the overlap, whereas competitive overlaps are persistent because speakers keep on speaking despite overlapping. Therefore, it can be inferred that competitive overlaps have longer durations than non-competitive overlaps.

With duration a key distinguishing feature, we observed that there is clear distinction between C0 and C1. We also observed that the median duration distribution of C1 is very close to the minimum distribution of C0. The minimum, median, third quartile and maximum durations, in milliseconds, of the clusters are C0 - {300, 740, 950, 3590} and C1 - {40, 330, 430, 850}, in that respect. The vertical lines in the figure indicate the median of the respective distributions.

5.2. Acoustic and Lexical Feature Analysis

For the analysis of acoustic features we used Relief feature selection technique [27] to rank the features, which has been useful in paralinguistic task [28]. It calculates the weight of the
features based on the nearest instances of the same and different classes. The top-ranked low-level acoustic features include logarithmic Harmonic to Noise Ratio (logHNR), $f_0$ envelope, shimmer-local, jitter-local, and spectral features, with their delta and acceleration coefficients, whereas the statistical functionals include range, standard deviation, mean of peak, linear regression with error coefficients, and centroid. Figure 4 shows some of the top-ranked low-level features projected on statistical functionals as described in Table 2. From the figure, it can be seen how the two clusters differ in their distributions – the mean values for C1 are always lower than those for C0.

Some of the voice quality features show significant difference in their distribution between two clusters. This indicates that these features play an important role in detailing the patterns in each cluster. logHNR is a feature which is widely used to analyze disorders such as hoarseness and depression. However, we observed that this feature has not been applied before to the analysis of overlaps.

Other commonly used features for categorizing overlaps are $f_0$, loudness and energy. By observing the values of F10 in Figure 4, it can be inferred that the mean value of C0 is higher than that of C1. This inference is extended to apply to the values of F11 as well. This, coupled with observations from previous research, provides the grounds for the conclusion that our C0 exhibits patterns similar to competitive overlaps. By studying the most frequent lexical features, it can be noted that filler and affirmative words are present in both clusters but that C1 has higher frequencies than C0 has. For example, the token “sì/yes” is present in C1 with a frequency of 2506, three times as much as that of the same token in C0. It can also be noted that, in comparison with C0, C1 has a homogenous lexicon, as demonstrated by the long tail of C0 in Figure 5.

### Table 2: Acoustic features and their description

<table>
<thead>
<tr>
<th>Feat.</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>Logarithmic harmonic to noise (logHNR) ratio with delta coefficient projected to statistical range</td>
</tr>
<tr>
<td>F2</td>
<td>LogHNR projected to statistical range</td>
</tr>
<tr>
<td>F3</td>
<td>LogHNR with delta coefficient projected to statistical mean of peak</td>
</tr>
<tr>
<td>F4</td>
<td>LogHNR projected to statistical standard deviation</td>
</tr>
<tr>
<td>F5</td>
<td>LogHNR with linear error computed as the difference of the linear approximation and the actual contour</td>
</tr>
<tr>
<td>F6</td>
<td>$f_0$ envelope projected to statistical mean of peak</td>
</tr>
<tr>
<td>F7</td>
<td>Local shimmer with centroid</td>
</tr>
<tr>
<td>F8</td>
<td>Local jitter with centroid</td>
</tr>
<tr>
<td>F9</td>
<td>$f_0$ envelope projected to geometric mean of non-zero values</td>
</tr>
<tr>
<td>F10</td>
<td>First formant with number of non-zero values</td>
</tr>
<tr>
<td>F11</td>
<td>Loudness with number of non-zero values</td>
</tr>
<tr>
<td>F12</td>
<td>Log energy with delta coefficient projected to non-zero values</td>
</tr>
</tbody>
</table>

Figure 5: Zipf’s plot with bigrams for C0 and C1 clusters. Frequency is plotted as a function of frequency rank.

### 6. Conclusion

In this study, we designed an automatic system that divides speech overlaps into two classes using an unsupervised approach. We prepared our data with an automated manner, by cascading a turn segmenter and an ASR system. For clustering, we extracted a large number of acoustic features from overlapped segments, and lexical features from automatic transcriptions. Our findings suggest that acoustic features play a more important role than lexical features in discovering well-separated clusters. Voice quality features, especially logHNR, jitter and shimmer, are the most discriminative in clustering overlaps. Based on previous work and from our analysis, we found that instances of C0 have a higher chance of being competitive overlaps, while instances of C1 are more likely to be of non-competitive nature. Our observation of lexical features, which are obtained from the clustering decision of acoustic features, is that the frequencies of filler and affirmative words are higher in C1 than the frequencies of such words in C0. These features may help in supervised classification, which we will investigate further. Future study includes manual annotation and studying contextual features as well as understanding the roles of acoustic and lexical features.

### 7. Acknowledgements

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


