Unsupervised Recognition and Clustering of Speech Overlaps in Spoken Conversations

Shammur A. 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, we discover and characterize speech overlaps using unsupervised techniques. Overlapping segments of human-human spoken conversations were extracted and transcribed using a large vocabulary Automatic Speech Recognizer (ASR). Each overlap instance is automatically projected onto a high-dimensional space of acoustic and lexical features. Then, we used unsupervised clustering to discover distinct and well-separated clusters that may correspond to different discourse functions (e.g., competitive, non-competitive overlap). We have evaluated recognition and clustering algorithms over a large set of real human-human spoken conversations. The automatic system separates two classes of speech overlaps. 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, in between the turns. Recent studies [2] suggested that the timing of turn-taking using overlaps and silence is less precise. It also argued that, overlap is actually a frequent phenomenon and is a lot 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 her or his turn 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 have focused on spoken conversations collected from a call center. Thus, the focus 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 the overlaps.

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

- Investigation of speech overlaps using unsupervised clustering.
- Extraction of very a large set of lexical features and acoustic features.
- Analysis of the speech overlaps' discriminative characteristics 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. Finally, 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. In previous work 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 F0 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 395.90 seconds.

As mentioned earlier, our data preparation process was completely automatic and we selected 515 conversations based on maximum duration of the overlapped segments. The overlapped segments are 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 components per conversation. Following this approach, we have extracted 25132 instances of overlaps from 515 conversations, where total overlap duration is 3 hours and 38 minutes and total speaking time is 41 hours and 52 minutes.

The ASR system was designed using a portion of the data set 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.

To train the ASR, we extracted the MFCC features, and the model was trained using Kaldi [12]. We obtained the best results by using the speaker adaptive training (SAT) that splices 3 frames on each side of the current frame. Linear Discriminant analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) feature-space transformations were then used to reduce the feature space then followed by MMI training. The Word Error Rate (WER) for our ASR system is 31.78% on the test set split using a tri-gram language model. The perplexity of the language model is 87.69.

4. Methodology

The workflow of clustering and feature analysis is shown in Figure 1. The overlap segment's components are shown for each channel. Acoustic and lexical features were extracted from both channels and then combined separately. In addition, we investigated the relevance of the combination such as acoustic and lexical features. We designed three feature set combinations: acoustic type only, lexical type only and acoustic together with lexical types. For each feature set, we performed cluster evaluation and analyzed the features based on the clustered output using a feature ranking approach. In addition, we tried to understand whether we are losing any information by reducing the feature dimension.

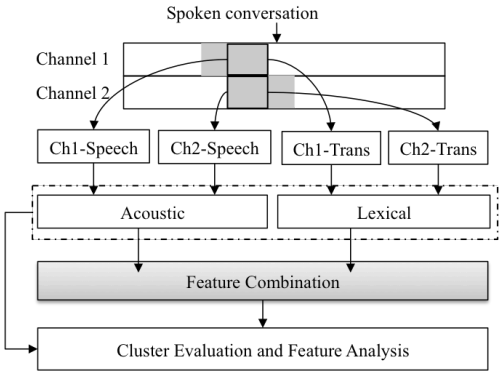

Figure 1: The overlap classification system.

4.1. Features

4.1.1. Acoustic features

We extracted a large number of acoustic features, motivated by their success in paralinguistic task [13,14]. The process is to extract a large number of low-level descriptors (LLD) and then project onto statistical functionals, which we have done by using openSMILE [15].

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 (probability of voicing computed from autocorrelation), fundamental frequency (F0), exponentially smoothed F0-envelope, jitter-local (pitch period length deviations), differential of jitter, shimmer-local (amplitude deviations between pitch periods), logarithmic harmonics-to-noise ratio (HNR) computed from auto-correlation, voice quality (probability of voicing), spectral features with different bands (0-250Hz, 0-650Hz, 250-650Hz, 1-4kHz), spectral roll-off points (25%,50%,70%,90%), centroid, flux, max-position and min-position, zero crossing rate of time signal and formant frequencies (F0-F3). Delta and acceleration coefficients of these features have also been extracted.

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

As mentioned earlier, overlap segment's components appear in two channels; therefore we extracted same features from both channels. The size of the feature vector in single channel is $(39 + \Delta 39 + \Delta\Delta 39)$ LLD $\times$ 24 functionals = 2808. After combining we ended up with 5616 features.

4.1.2. Lexical features

Lexical features were extracted from automatic transcription using the ASR explained in Section 3. The lexical features transformed into bag-of-words (vector space model) for clustering. Bag-of-words is a numeric representation of text that has been introduced in text categorization [16]. The idea of this approach is to represent the words into numeric features. For this study, we extracted bigram features and select top 2K frequent features to reduce the load of the large dictionary. We have not used higher order n-gram due to the limitation of the utterance length in the overlapped segment. The frequency in the feature vector was then transformed into tf-idf - logarithmic term frequency (tf) times inverse document frequency (idf).

4.1.3. Feature Combination

Feature combination has been widely used in other speech-processing task and its relative contribution varies greatly depending on the data and experiments. For this study, we also wanted to understand the contribution of feature combination. As shown in Figure 1, after extracting acoustic and lexical features we combined 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 depends on the number of features, therefore we tried to reduce the feature space to reduce the complexity and number of free parameters. Typical approach for feature reduction is to map higher dimensional feature space into lower dimensional space, while keeping as much information as possible. In our study, we have used principle component analysis (PCA), which is the fundamental and most widely used feature reduction. 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 leading p components with 99% variance. Hence, we reduced 63% acoustic, 11% lexical and 59% acoustic+lexical features. The reason of obtaining minimal reduction with lexical features is the weak correlation with feature dimensions and sparseness.

4.2. Clustering

To find the well-separated clusters of speech overlaps in our dataset we used K-means [18] where data points are classified as belongings to one of K-group. For reproducibility and transferability we used weka's implementation [19]. Members of the clusters are determined by comparing the data point with each groups centroid and assigning to the nearest one. The reason to choose K-means is that it is highly recommended for large dataset [20, 21] and is one of the simplest methods. However, one of main downside of K-means is 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.

In its process, for each k value, it calculates the between-group dispersion, BGSS; within-cluster sum of squares, WGSS and Calinski-Harabasz (CH) value or index, using Equation [1-3]

$$BGSS = \sum_{k=1}^{K} n_k \left\| G^{\{k\}} - G \right\|^2 \quad (1)$$

$$WGSS = \sum_{k=1}^{K} \sum_{i \in I_k} \left\| M_i^{\{k\}} - G^{\{k\}} \right\|^2 \quad (2)$$

$$CH = \frac{(N-K) * BGSS}{(K-1) * WGSS} \quad (3)$$

where **K** is the number of clusters, **N** is number of observations, $\mathbf{G^{\{k\}}}$ is the barycenter of cluster C_k, **G** is the barycenter of the whole dataset, $\mathbf{n_k}$ is the number of elements in the C_k. $\mathbf{I_k}$ is the set of the indices of the observations belonging to C_k $\mathbf{M_i}$ is the ith observation of element in C_k.

Figure 2 shows the values of CH corresponding to the number of cluster K and 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 and their clustering difference is very minimal. Using lexical feature, we obtained the optimal number for k is 4, and the CH value for the cluster is significantly less compare to CH values with acoustic features. The minimal separability of lexical features could be due to the sparseness and the recognition error of ASR.

We applied PCA feature reduction method on acoustic and acoustic+lexical feature sets and with reduced dimensions we obtained 2 clusters in each set. We then calculated the cluster agreement of different feature sets using kappa (**κ**) statistics [23]. We found that the agreement between original and reduced dimensions is quite reasonable. The agreements for acoustic and acoustic+lexical feature sets are 92% and 91% respectively. This indicates that feature reduction helps in terms of computational cost without any loss of information.

To check the validity of clusters using cascaded k-means we used another well-known clustering algorithm, which is Spectral Clustering [24, 25]. Then, compared the clusters generated by two clustering algorithms for acoustic and acoustic+lexical feature sets using **κ** measure. The agreement between the two algorithms on acoustic and acoustic+lexical feature sets are 90% and 87% respectively. For the sake of simplicity, we do not show the details cluster results of Spectral Clustering algorithm in this paper.

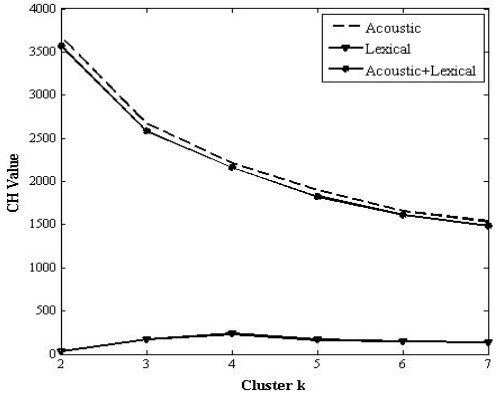

Figure 2: Calinski-Harabasz (CH) value for cluster decision

Feature Set	K	CH	W	B
Acoustic	2	3681.12	42.31	155749.22
Lexical	4	232.66	1.38	320.04
Acoustic +Lexical	2	3568.28	43.71	155985.56
Acoustic + lexical with PCA	2	2754.82	21.28	58631.95

Table 1: Cluster evaluation using different feature sets, K - number of cluster, CH - Calinski-Harabasz value, W – weighted within-cluster sum of squares, B – between group dispersion

5. Analysis and Discussions

We analyzed different features based on the cluster decision of acoustic features, where cluster 0 (C0) and cluster 1 (C1) contains 37% and 63% of overlapping instances respectively. The members of 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 tried to correlate our observation with previous studies to see whether our clusters represented 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 of [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 the occurrence of overlap. Therefore, it can be inferred that competitive overlaps have longer durations than non-competitive overlaps.

Considering duration as a key distinguishing feature, we observed that there is a 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.

5.2. Acoustic and Lexical Feature Analysis

For the analysis of acoustic features we used Relief [27] feature selection technique to rank the features. The top ranked low-level acoustic features include logarithmic harmonic to noise ratio (logHNR) with its delta and acceleration coefficients, F0 envelope, shimmer-local, jitter-local, spectral features, etc. Whereas the statistical functionals include range, standard deviation, mean of peak, linear regression with error coefficients, centroid, etc. Figure 4 shows some of the top ranked low-level features projected on statistical functionals as described in Table 2. From the figure, we can see how two clusters differ in their distributions; the mean values for C1 are always lower compare to C0.

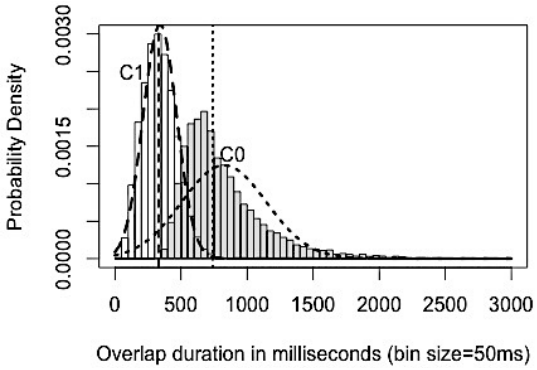

Figure 3: Overlap duration distribution of the two clusters

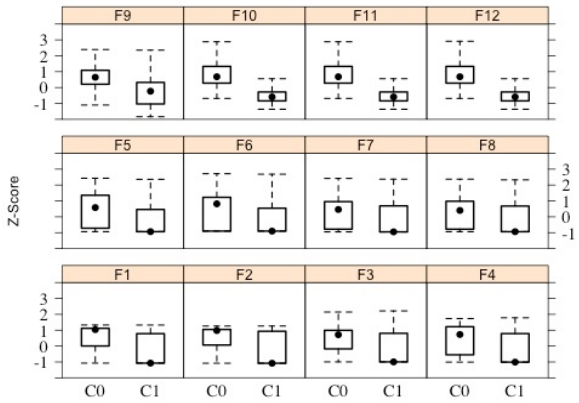

Figure 4: Selected acoustic features (F1-12) and their z-score distribution in C0 and C1. Box-plots, representing the mean, max, upper and lower inner fences of top ranked features. Outliers have been removed for readability.

The significant difference in the means of the voice quality features (F1-F4, F6, F7) 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 understand that this feature has not been used before in the analysis of overlaps. Other commonly used features for categorizing overlaps are F0, loudness and energy. By observing the values of F0 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. For example, the token "sì/yes" is present in C1 with a frequency of 2506, three times as much as that of C0. It can also be noted that, in comparison with C0, C1 has a homogenous lexicon, giving C0 its long tail as shown in Figure 5.

Feat.	Description
F1	logarithmic harmonic to noise (logHNR) ratio with delta coefficient projected to statistical range
F2	logHNR projected to statistical range
F3	logHNR with delta coefficient projected to statistical mean of peak
F4	logHNR projected to statistical standard deviation
F5	logHNR with linear error computed as the difference of the linear approximation and the actual contour
F6	F0 envelope projected to statistical mean of peak
F7	local shimmer with centroid
F8	local jitter with centroid
F9	F0 envelope projected to geometric mean of non-zero values
F10	first formant with number of non-zero values
F11	loudness with number of non-zero values
F12	log energy with delta coefficient projected to non-zero values

Table 2: Acoustic features and their description

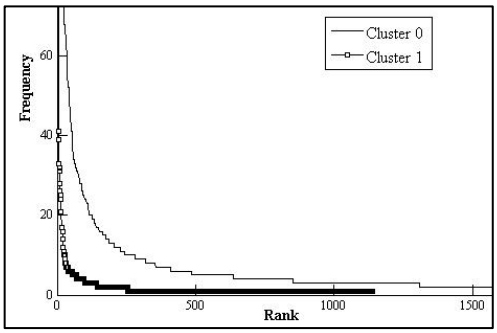

Figure 5: Zipf's plot with bigrams for overlapped clusters. Frequency is plotted as a function of frequency rank.

6. Conclusion

In this study, we designed an automatic system that separates the speech overlaps into two classes using unsupervised approach. Our data preparation was done with an automated process using a cascade of speech segmenter and ASR system. For clustering, we extracted a large number of acoustic features from overlapped segments and lexical features from automatic transcription. Our findings suggest that acoustic features play an important role for discovering well-separated clusters compared to lexical features. The voice quality features especially logHNR, jitter and shimmer are the most discriminating features in clustering the overlaps. From our analysis we found that instances of C0 have a higher probability to be competitive overlap whereas C1 to be non-competitive. Our observation on lexical features, obtained from the clustering decision of acoustic features, suggests that the frequency of filler and affirmative words are higher in C1 compare to C0. More investigation is needed to understand the role of 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.

8. References

[1] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation", Language, 50(4): 696–735, 1974.

[2] Heldner, M., & Edlund, J. "Pauses, gaps and overlaps in conversations." Journal of Phonetics, 38(4), 555-568, 2010.

[3] French, P., Local, J., "Turn-competitive incomings", J. Pragmatics 7, 701–715, 1983.

[4] Kurtić, E., Brown, G. J., & Wells, B, "Resources for turn competition in overlapping talk", Speech Communication, 55(5), 721-743, 2013.

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

[6] Schegloff, E., "Overlapping talk and the organisation of turn-taking for conversation", Lang. Soc. 29, 1–63, 2000.

[7] K. P. Truong, "Classification of cooperative and competitive overlaps in speech using cues from the context, overlapper, and overlappee," in Proceedings of Interspeech, 2013.

[8] Wells, B., McFarlane, S., "Prosody as an interactional resource: Turn-projection and overlap", Lang. Speech 41, 265–294, 1998.

[9] Hammarberg, B., Fritzell, B., Gaufin, J., Sundberg, J., & Wedin, L., "Perceptual and acoustic correlates of abnormal voice qualities", Acta otolaryngologica, 90(1-6), 441-451, 1980.

[10] Kurtic´, E., Brown, G.J., Wells, B., "Resources for turn competition in overlap in multi-party conversations: Speech rate, pausing and duration", Proc. Interspeech 2010. Makuhari, Japan, 2010.

[11] Ivanov, A. V., & Riccardi, G, "Automatic turn segmentation in spoken conversations", In Interspeech 2010, pp. 3130-3133, 2010.

[12] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, Karel Vesely, "The Kaldi speech recognition toolkit", Proc. ASRU, 1-4, 2011.

[13] B. Schuller, "Voice and speech analysis in search of states and traits," in Computer Analysis of Human Behavior. Springer, pp. 227–253, 2011.

[14] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge", Speech Communication, vol. 53, no. 9, pp. 1062–1087, 2011.

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

[16] T. Joachims, Text categorization with support vector machines: Learning with many relevant features. Springer, 1998.

[17] Alpaydin, Ethem, "Introduction to machine learning", MIT press, 2004.

[18] Arthur, David and Vassilvitskii, Sergei, "k-means++: The advantages of careful seeding", Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, pp. 1027-1035, 2007.

[19] I.H. Witten and E. Frank, "Data Mining: Practical machine learning tools and techniques", Morgan Kaufmann, 2005.

[20] Huang, Zhexue, "Extensions to the k-means algorithm for clustering large data sets with categorical values", Data Mining and Knowledge Discovery 2.3, pp. 283-304, 1998.

[21] Abbas, O. A., "Comparisons Between Data Clustering Algorithms", International Arab Journal of Information Technology (IAJIT), 5(3), 2008.

[22] T. Calinski, J. Harabasz, "A dendrite method for cluster analysis", Communications in Statistics-theory and Methods, vol. 3, no.1, pp. 1-27,1974.

[23] Donner, Allan, and Neil Klar. "The statistical analysis of kappa statistics in multiple samples", Journal of clinical epidemiology, vol. 49, no.9, 1053-1058, 1996.

[24] L. Zelnik-Manor and P. Perona, "Self-Tuning Spectral Clustering", Advances in Neural Information Processing Systems 17, pp. 1601-1608, 2005, (NIPS'04).

[25] Von Luxburg, Ulrike, "A tutorial on spectral clustering", Statistics and computing, Springer, Vol-17, Number-4, pp. 395-416, 2007.

[26] Jefferson, G., "A sketch of some orderly aspects of overlap in natural conversation", PRAGMATICS AND BEYOND NEW SERIES, 125, 43-62, 2004.

[27] Igor Kononenko, "Estimating Attributes: Analysis and Extensions of RELIEF", In European Conference on Machine Learning, 171-182, 1994.