INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Speech and Text Analysis for Multimodal Addressee Detection in Human-Human-Computer Interaction

Oleg Akhtiamov[1, 2], *Maxim Sidorov*[1], *Alexey Karpov*[2, 3] *and Wolfgang Minker*[1]

[1]Ulm University, Germany
[2]ITMO University, Russia
[3]SPIIRAS, St. Petersburg, Russia

oakhtiamov@gmail.com, maxim.sidorov@alumni.uni-ulm.de, karpov_a@mail.ru, wolfgang.minker@uni-ulm.de

Abstract

The necessity of addressee detection arises in multiparty spoken dialogue systems which deal with human-human-computer interaction. In order to cope with this kind of interaction, such a system is supposed to determine whether the user is addressing the system or another human. The present study is focused on multimodal addressee detection and describes three levels of speech and text analysis: acoustical, syntactical, and lexical. We define the connection between different levels of analysis and the classification performance for different categories of speech and determine the dependence of addressee detection performance on speech recognition accuracy. We also compare the obtained results with the results of the original research performed by the authors of the Smart Video Corpus which we use in our computations. Our most effective meta-classifier working with acoustical, syntactical, and lexical features reaches an unweighted average recall equal to 0.917 showing almost a nine percent advantage over the best baseline model, though this baseline classifier additionally uses head orientation data. We also propose a universal meta-model based on acoustical and syntactical analysis, which may theoretically be applied in different domains.

Index Terms: Off-Talk, speaking style, acoustical analysis, syntactical analysis, lexical analysis, text classification, spoken dialogue system

1. Introduction

Spoken dialogue systems (SDSs) have become significantly more complex and flexible over recent years and are now capable of solving a wide range of tasks. The requirements for SDSs depend on a particular application area; e.g., personal assistants in smartphones are meant to interact with a single user – the owner. Theoretically, the interaction between a user and such a system may be considered as a pure human-computer (H-C) dialogue. However, there is the possibility that the user is solving a cooperative task that requires some interaction with other people nearby, e.g., interlocutors may be negotiating how they will spend this evening, asking the system to show information about cafes or cinema and discussing alternatives. In this case, the system deals with a multiparty conversation which may include human-addressed utterances as well as machine-addressed ones, leading to the problem of addressee detection (AD) in human-human-computer (H-H-C) conversations [1]. Solving this problem, the system needs to determine whether it is being addressed or not and provide the addressee prediction to a dialogue manager so that it can control a dialogue flow more precisely; the SDS is supposed to give an immediate response in the case of a direct request, otherwise, the system is not supposed to participate in the dialogue actively.

Traditionally, user interfaces have been engineered to avoid addressee ambiguity by using a push-to-talk button, key words, or by assuming that all potential input utterances are system-addressed and rejecting those which cause a failure-to-recognize or a failure-to-interpret [2, 3]. These straightforward approaches are no longer applicable, since modern SDSs support essentially unlimited spoken input, i.e., input queries may be in arbitrary conversational form. Therefore, more sophisticated classification methods are required for AD.

The present paper is a continuation of our previous study on text-based AD [4] and includes three main contributions. The first contribution is an attempt to extract as much useful information from audio signal as possible. Relying on other modalities, e.g., on visual information, is not reasonable in certain applications in which users have no visual contact with the object they are talking to, e.g., while driving a car. The second contribution is to define the connection between different levels of speech and text analysis and the classification performance for different categories of speech. The third contribution is to update the results of an existing study. In our work, we analyse the Smart Video Corpus (SVC) and compare our results on the AD problem with the results obtained by the authors of the corpus. In their original research, the term 'Off-Talk detection' is used instead of AD [5].

The paper is organised as follows: in Section 2, we report on several existing studies and point out the main concepts and features which are considered to be important for AD. Section 3 describes the Smart Video Corpus and some basic steps of data preparation. In Section 4, we provide a description for different levels of speech and text analysis and propose several classification methods. Section 5 contains the comparative analysis of the proposed classification models as well as their comparison with the baseline classifiers proposed by the authors of the corpus. In Section 6, we analyse the performance of different models for different categories of speech, and finally, make concluding remarks and specify prospective directions of AD for future SDSs in Section 7.

2. Related Work

There exist several studies investigating the separate roles of acoustical [6], textual [7], and visual [8] information towards the AD problem. It was determined that people combine prosodic, lexical, and gaze cues to specify desirable addressees [1]. Other works report that the way users talk to an SDS essentially depends on the overall system performance [5] and

http://dx.doi.org/10.21437/Interspeech.2017-501

how people see the system (as a human-like robot or as an information kiosk) [9]. Modern SDSs are still far from perfection, and users tend to change their normal manner of speech and talk to the system as if they were talking to a child [10], making it easier to understand, and, therefore, prosodic information plays a significant role in AD. The fact that prosodic features use no lexical, context, or speaker information makes prosody a universal modality for applications nowadays [6]. Simultaneously with future SDS improvement, prosodic features will become less representative, and future systems will thus rely more on textual and gaze information. It was shown that addressee and response selection in multiparty conversations between humans can be successfully performed by analysing lexical content and conversational context with recurrent neural networks [11].

The following features are representative for AD in existing SDSs (according to their relative contribution in descending order): acoustical, automatic speech recognition (recognized text and recognition confidence), dialogue state, gaze direction, and beamforming [1]. In the present study, we observe only those features which can be extracted from audio signal.

3. Experimental Data

The SVC data (part of the Smart Web Project) has been collected within large-scaled Wizard-of-Oz experiments and models the H-H-C conversation in German between two users and a multimodal SDS. The corpus includes queries in the context of a visit to a Football World Cup stadium in 2006. A user was carrying a mobile phone, asking questions of certain categories (transport, sights, sport statistics, and also open-domain questions) and discussing the obtained information with another human whose speech is not presented in the corpus. The data comprises 3.5 hours of audio and video, 99 dialogues (one unique speaker per dialogue), 2193 automatically segmented utterances with manual transcripts, and 25 073 words in total. The labelling of addressees was carried out for each word; four word classes were specified: On-Talk (NOT) – computer-addressed speech, read Off-Talk (ROT) – reading information aloud from the system display, paraphrased Off-Talk (POT) – retelling the information obtained from the system in arbitrary form, and spontaneous Off-Talk (SOT) – other human-addressed speech. No requirements regarding Off-Talk were given in order to obtain a realistic H-H-C interaction.

In our research, all features are extracted at the utterance level in contrast to the original study in which the authors analysed word-level features initially, though they also transformed predictions of word-based classifiers into meta-features at the utterance level. Frankly speaking, their meta-classifiers perform utterance-based AD as well as our models do. An utterance label is calculated as the mode of word labels in the current utterance. After performing the word-to-utterance label transformation, we obtain 1087 NOT, 474 SOT, 323 POT and 309 ROT utterances. We consider a two-class task only (On-Talk vs. the three Off-Talk classes), since it is equivalent to the AD problem. Experiments with a four-class task may be found in the original paper [5]. After merging the three Off-Talk classes into one and performing the word-to-utterance label transformation, we obtain 1078 On-Talk and 1115 Off-Talk utterances.

4. Classification

4.1. Speech analysis

The main idea of using acoustical information for AD is the fact that people make their speech louder, more rhythmical, and easier to understand in general once they start talking to an SDS. There is no standard feature set for acoustical AD. Several research groups analysed different sets [1, 5], and therefore, we decided to use a highly redundant paralinguistic attribute set to perform feature selection afterwards. We extract 6373 acoustical attributes for each utterance by applying the openSMILE toolkit and the feature configuration of the INTERSPEECH 2013 Computational Paralinguistics Challenge [12]. After that, we calculate the coefficients of the normal vector of a linear support vector machine (SVM) for each fold and set them as attribute weights. We sort the attributes according to their weights and perform recursive feature elimination removing the 50 attributes with the lowest weights per step. As a classifier, we apply a liner SVM implemented in RapidMiner Studio 7.3 [13]. It turned out that the optimal number of attributes was approximately 1000 in each fold, therefore, it was decided to use the first 1000 attributes with the highest weights. The selected features are speaker-dependent, however, they are much less sensitive to a specific domain in comparison with lexical attributes.

4.2. Text analysis

The text obtained with automatic speech recognition (ASR) allows us to carry out syntactical and lexical analysis. In this paper, most text-based computations are performed by using manual transcripts (it is assumed that our recognizer has word recognition accuracy close to 100%). We also test our system in conjunction with a real recognizer (Google Cloud Speech API) with word recognition accuracy of around 80% and analyse three additional ASR-based features besides text: recognition confidence, number of recognized words and utterance length. The underlying idea is that computer-addressed speech matches the ASR pattern better than human-addressed speech does. For these three attributes, we apply the same classifier as for the acoustical features.

4.2.1. Syntactical analysis

We perform two stages of text analysis. The first stage is syntactical analysis which allows us to determine differences in the structure of human- and computer-addressed sentences. The underlying idea is that the syntax of machine-addressed speech possesses more structured patterns in comparison with the syntax of human-addressed speech. As a representation of syntactical structure, we apply part-of-speech (POS) *n*-gram. Firstly, we perform POS tagging by using spaCy 1.8 [14] and obtain utterances in which each word is replaced by one of 15 universal POS tags. After that, we extract uni- bi-, tri-, tetra-, and pentagrams and weight them by using the following term weighting methods: Inverse Document Frequency (IDF), Gain Ratio (GR), Confident Weights (CW), Second Moment of a Term (TM2), Relevance Frequency (RF), Term Relevance Ratio (TRR), and Novel Term Weighting (NTW) [15]. The obtained syntactical attributes are language-dependent, however, they are much less sensitive to a specific domain in comparison with lexical features. We apply three classification algorithms which demonstrated high performance for other text classification tasks [15]: *k* Nearest Neighbours (KNN) [16], Fast Large Margin (Liner SVM-based classifier – SVM-FLM)

[17], and Rocchio (Centroid classifier) [18]. The first two classifiers were implemented in RapidMiner Studio 7.3, the third one was developed in C++.

4.2.2. Lexical analysis

The second stage of text analysis is lexical analysis which allows us to determine typical lexical units for each class. In other words, this kind of analysis shows *what* has been said, while acoustical and syntactical analysis indicate *how* it has been said. We perform the same procedure of text classification as it was shown for syntactical analysis with a single distinction: we deal with real words instead of POS tags. Firstly, we apply two linguistic filters implemented in tm (R package for text mining): stemming and stop-word filtering. Then, we extract uni-, bi-, and trigrams, weight them by using the seven term weighting methods and apply the three classification algorithms mentioned above.

4.3. Data fusion

In order to get benefits from all the levels of speech and text analysis, we carry out data fusion. Combining the ASR additional information and the acoustical attributes, we perform feature-level fusion, while a meta-classifier based on a linear SVM is applied for different combinations of the acoustical, syntactical and lexical models. Feature-level fusion for these groups of attributes shows poor results due to the high dimensionality of a new feature vector after concatenation. As input features, each meta-classifier receives the classification confidence scores of the models included in it. In order to train meta-models, we split each training set into two sets in a proportion of eight to two. The first set is used for training single models, and the second one provides unique information for training the corresponding meta-model. An example of a meta-classifier is depicted in Figure 1. In the original research, a linear discriminant classifier was applied for single models as well as at the meta-level [19].

5. Experimental Results

For statistical analysis, we carry out leave-one-group-out cross validation splitting the entire corpus into 14 folds (7 speakers for each and one more speaker to the fold with the least number of utterances) so that the proportion of classes remains equal in each fold. All statistical comparisons are drawn by using a *t*-test with a confidential probability of 0.95. Unweighted average recall has been chosen as the main performance criterion in order to make a correct comparison with the original research.

Figure 2 illustrates the performance of different classifiers. An average performance value and a standard deviation are calculated for each model. The ASR additional information (ASR info) and the acoustical attributes (ac) demonstrate a significant dependence on speakers and also show the lowest performance of 0.668 and 0.822 respectively which becomes significantly higher up to 0.828 after their feature-level fusion.

We have determined that the best models for syntactical and lexical analysis include POS tagging + trigrams + RF term weighting + SVM-FLM classifier and stemming + unigrams + RF term weighting + SVM-FLM classifier respectively. There is no significant difference between the acoustical and the syntactical model (synt) which demonstrates a performance of 0.836. All the lexical classifiers show the highest results among single models. Stemming (lex s) reduces the dimensionality of the text classification task by 20% (the average dictionary size falls from 1381 to 1108) keeping the AD performance at the level of the lexical model without linguistic filtering (lex) which reaches a result of 0.911, while stop-word filtering (lex f) significantly decreases the performance to 0.883.

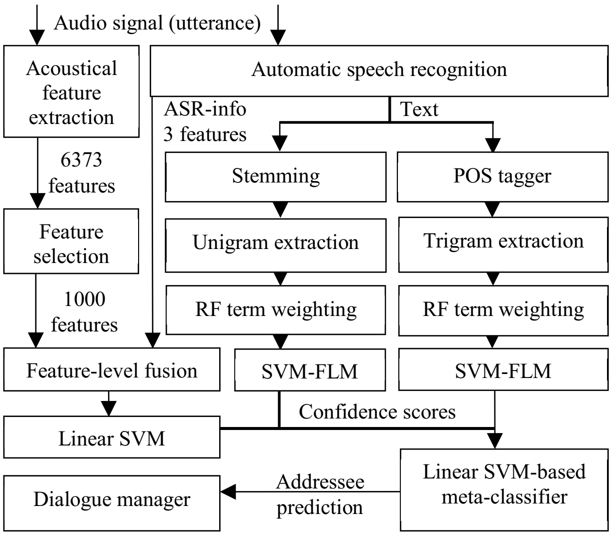

Figure 1: *Scheme of a meta-classifier.*

Each meta-classifier (meta) significantly outperforms each single model included in it. The most effective meta-classifier analysing the information at all three levels reaches a performance of 0.929 demonstrating a statistically significant advantage over the other models. The scheme of this meta-classifier is depicted in Figure 1. The performance of another meta-classifier working with acoustical and syntactical information is significantly lower and equal to 0.886. However, the main advantage of this meta-model is domain-independence and a higher degree of universality in comparison with the most effective meta-model which is domain-dependent.

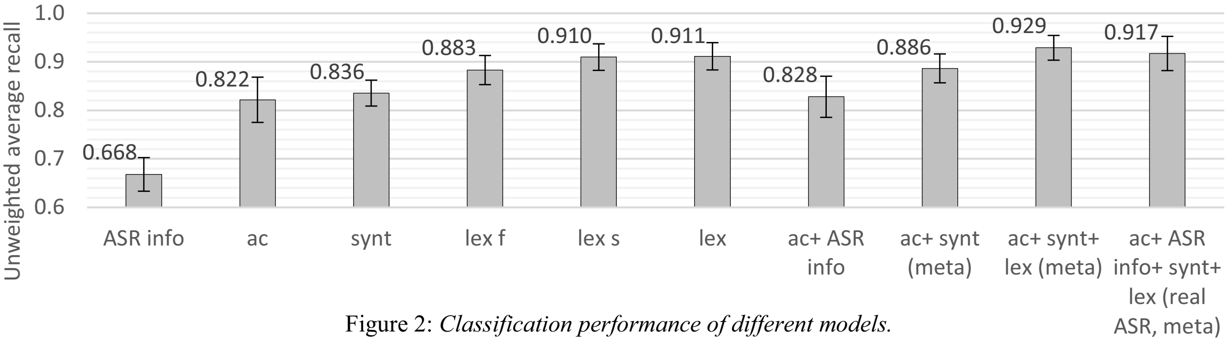

Figure 2: *Classification performance of different models.*

There is no significant difference between the most effective meta-classifier using the textual information obtained from the manual transcripts and the analogical meta-model working with the real ASR and showing a performance of 0.917.

We tried to reproduce the original experiment [5] as precisely as possible. We excluded four speakers which had technical problems, then we randomly split the remaining speakers into a training (58 speakers) and a test set (37 speakers) until we obtained approximately the same number of utterances in the respective sets as they were in the original research. Figure 3 demonstrates that all the proposed models outperform the corresponding baselines analysing related groups of features, particularly, our most effective meta-classifier reaches a performance of 0.917 showing almost a nine percent advantage over the most effective baseline classifier, though this baseline model additionally uses head orientation data [5].

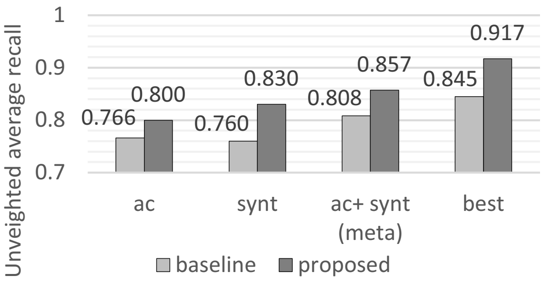

Figure 3: *Comparison with the results of the original research.*

6. Analysis

The obtained meta-model analysing acoustical and syntactical features may be theoretically applied in different domains, since it uses the attributes containing no lexical information. The most effective meta-classifier considers also lexical content. The text-based models are less speaker-dependent in comparison with the acoustical model but also language-dependent (syntactical model) and even domain-dependent (lexical model). The lexical models demonstrate the highest results among single models for the particular domain. The following groups of lexical terms have the highest RF weights and are therefore considered to be important: question words and polite requests for On-Talk, pronouns (particularly, second person), indirect speech, colloquial words and interjections for Off-Talk. It turned out that lexical AD is not sensitive to different word forms, since stemming does not influence the classification performance, while stop-word filtering decreases the performance removing some important terms, e.g., pronouns.

Solving the two-class task (NOT vs. the other categories of speech) and comparing the classification performance for separate categories of speech in Figure 4, we see that the text-based classifiers have the strongest confusion between NOT and SOT and significant confusion between NOT and POT that leads us to the conclusion that the more spontaneous the speech is, the worse the text-based models work. The acoustical and the ASR-info-based classifier possess the strongest confusion between NOT and ROT and significant confusion between NOT and POT, meaning that the more limited the speech is, the worse results these models demonstrate.

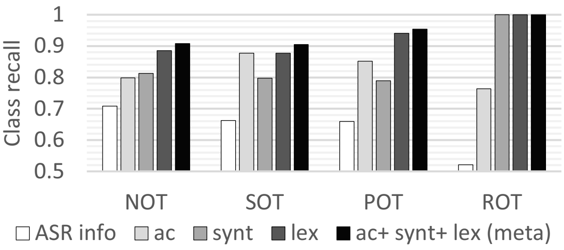

Figure 4: *Classification performance for different categories of speech.*

7. Conclusions and Future Work

The comparison with the original research has shown that utterance-based AD provides more context information and thus leads to higher results than word-based AD does. The classification performance may be further improved; we are planning to integrate head orientation data into the present research to perform a more complete comparison with the baseline [5].

Due to the comprehensive utterance-level analysis with several stages of speech and text processing, even relatively simple machine learning models are able to demonstrate effective results for the AD problem. More complex models such as deep neural networks require a larger amount of training data, which is difficult to obtain during the process of collecting a corpus of realistic human-human-computer interaction. However, it is possible to use out-of-domain data, e.g., for textual modality, to train a word2vec model which would extract word embedding vectors from the raw text. Such a feature extractor may be domain-independent [20], and it would be possible to replace the stages of syntactical and lexical analysis by a single text-based model which might be a recurrent neural network processing utterances as sequences of word embedding vectors and returning addressee predictions [21].

It is necessary to keep in mind that the more advanced the SDS turns out to be, the more naturally users behave, and the less it should rely on acoustical information while detecting addressees. Text and dialogue state will remain reliable, and therefore, we are planning to focus on conversational context-based AD for multiparty SDSs in our future work [11].

8. Acknowledgements

This research is partially supported by DAAD together with the Ministry of Education and Science of the Russian Federation within the Michail Lomonosov Program (project No. 8.704.2016/DAAD), by RFBR (project No. 16-37-60100), by the Government of the Russian Federation (grant No. 074-U01), and by the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" which is funded by DFG.

9. References

[1] T. J. Tsai, A. Stolcke and M. Slaney, "A study of multimodal addressee detection in human-human-computer interaction," *IEEE Transactions on Multimedia*, vol. 17, no. 9, pp. 1550-1561, Sept. 2015.

[2] J. Dowding, R. Alena, W. J. Clancey, M. Sierhuis, and J. Graham, "Are you talking to me? Dialogue systems supporting mixed teams of humans and robots," *Proc. AAAI Fall Symp.: Aurally*

Informed Performance: Integrating Mach. Listening Auditory Presentation Robot. Syst., Washington, DC, USA, pp. 22–27, Oct. 2006.

[3] T. Paek, E. Horvitz, and E. Ringger, "Continuous listening for unconstrained spoken dialog," *Proc. ICSLP*, B. Yuan, T. Huang, and X. Tang, Eds., vol. 1, pp. 138–141, Oct. 2000.

[4] O. Akhtiamov, R. Sergienko and W. Minker, "An approach to Off-Talk detection based on text classification within an automatic spoken dialogue system," *Proceedings of the 13th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2016)*, Lisbon, Portugal, Vol. 2, pp. 288-293, July 2016.

[5] A. Batliner, C. Hacker and E. Noeth, "To talk or not to talk with a computer," *Journal on Multimodal User Interfaces*, vol. 2, no. 3, pp. 171-186, 2008.

[6] E. Shriberg, A. Stolcke, and S. Ravuri, "Addressee detection for dialog systems using temporal and spectral dimensions of speaking style," *Proceedings INTERSPEECH 2013*, pp. 2559–2563, Aug. 2013.

[7] H. Lee, A. Stolcke, and E. Shriberg, "Using out-of-domain data for lexical addressee detection in human-human-computer dialog," *Proc. North Amer. ACL/Human Language Technol. Conf.*, pp. 221–229, June 2013.

[8] M. Johansson, G. Skantze, and J. Gustafson, "Head pose patterns in multiparty human-robot team-building interactions," *Proceedings of the 5th International Conference on Social Robotics (ICSR 2013)*, Bristol, UK, pp. 351-360, October 2013.

[9] M. K. Lee, S. Kiesler, and J. Forlizzi, "Receptionist or information kiosk: how do people talk with a robot?," *Proc. 2010 ACM Conf. Comput. Supported Cooperative Work*, pp. 31–40, 2010.

[10] B. Schuller et al., "The INTERSPEECH 2017 computational paralinguistics challenge: addressee, cold & snoring", *Proceedings INTERSPEECH 2017*, Stockholm, Sweden, 2017.

[11] H. Ouchi and Y. Tsuboi, "Addressee and response selection for multi-party conversation," *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, pp. 2133–2143, November 2016.

[12] B. Schuller et al., "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," *Proceedings INTERSPEECH 2013*, Lyon, France, August 2013.

[13] A. Ben-Hur and J. Weston, "A user's guide to support vector machines," *Data Mining Techniques for the Life Sciences*, pp. 223-239, Humana Press 2010.

[14] spaCy library. https://github.com/explosion/spaCy

[15] R. Sergienko, M. Shan and W. Minker, "A comparative study of text preprocessing approaches for topic detection of user utterances," *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*, Portorož (Slovenia), pp. 1826-1831, May 2016.

[16] Y. Zhou, Y. Li and S. Xia, "An improved KNN text classification algorithm based on clustering," *Journal of computers*, vol. 4, no. 3, pp. 230-237, 2009.

[17] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, C. J. Lin, "Liblinear: a library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[18] T. Joachims, "A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization," no. CMU-CS, pp. 96-118, CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE, 1996.

[19] W. Klecka, Discriminant analysis, 9 edn. SAGE PUBLICATIONS Inc., Beverly Hills, 1988.

[20] J. Pennington, R. Socher, C. Manning, "GloVe: global vectors for word representation," *in Proc. EMNLP*, Doha, Qatar, vol. 14, pp. 1532–1543, 2014.

[21] S. Ravuri, A. Stolcke, "Recurrent neural network and LSTM models for lexical utterance classification," *in Proc. Interspeech*, pp. 135-139, 2015.