International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 2 ǁ 2022 ǁ PP. 21-25

Virtual Reality – A Human Computer Interface

*[1]Mrs.Apitha Kuchambal K.,MCA.,M.phil, *[2]Mrs.Padma Priya J.,MCA.,M.phil,.
*[3]Mrs.Rathina Kumari A.,MCA.,

[1]HOD & Assistant Professor, Department of Computer Application,
Arcot Sri Mahalakshmi Womens College,Villapakkam, TamilNadu, India.
[2]Assistant Professor, Department of Computer Application,
Arcot Sri Mahalakshmi Womens College, Villapakkam, TamilNadu, India.
[3]Assistant Professor, Department of Computer Application,
Arcot Sri Mahalakshmi Womens College, Villapakkam, TamilNadu, India.

Abstract:
The belief that humans will be able to interact with computers in conversational speech has long been a favorite subject in science fiction, reflecting the persistent belief that spoken dialogue would be the most natural and powerful user interface to computers. With recent improvements in computer technology and in speech and language processing, such systems are starting to appear feasible. There are significant technical problems that still need to be solved before speech-driven interfaces become truly conversational. This article describes robotics interact with humans through speech.

Keywords: *Address detection (AD), Spoken dialogue system (SDS), speaking style, interlocutors, acoustical, syntactical, and lexical speech.*

Date of Submission: 02-02-2022 Date of acceptance: 16-02-2022

I. INTRODUCTION

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 address 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.

1.1 Spoken dialogue systems (SDSs): spoken dialogue systems are 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 address detection (AD) in human-human-computer (H-H-C) conversations[1] .

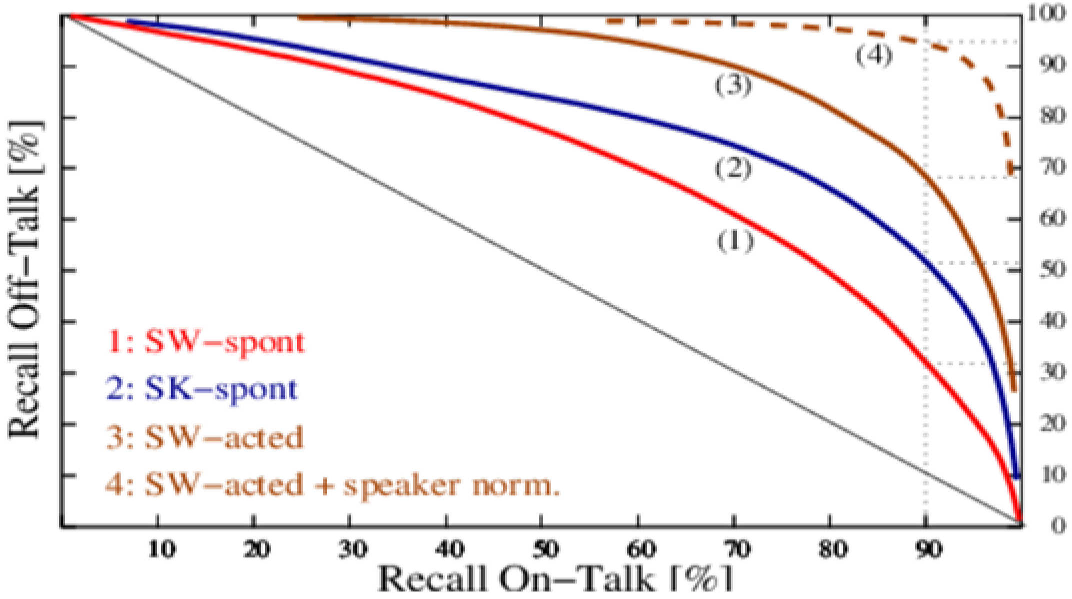
Solving this problem, the system needs to determine whether it is being addressed or not and provide the addressee prediction to a dialogue manager so that it can control a dialogue flow more precisely, the SDS is supposed to give an immediate response in the case of a direct request otherwise, the system 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. These straightforward approaches are no longer applicable, since modern SDSs support essentially unlimited spoken input. The present paper is a continuation of our previous study on text-based AD 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 about 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 analyssee the Smart Video Corpus (SVC) and compare our results on the AD problem with the results obtained by the authors of the corpus[5][6].

1.2 Ontalk and Offtalk: People talking to a computer can – the same way as while talking to another human – speak aside, either to themselves or to another person. On the one hand, the computer should notice and process such utterances in a special way; such utterances provide us with unique data to contrast these two registers: talking vs. not talking to a computer. By that, we can get more insight into the register 'Computer-Talk'. In this figure 1, we present two different databases, SmartKom and SmartWeb, and classify and analyse On-Talk (addressing the computer) vs. Off-Talk (addressing someone else) found in these two databases.An unintentional command activation when a human voice generates the same tone as a control signal . In their original research, the term 'OffTalk detection' is used instead of AD[2] .

Figure 1: Evaluation on-Talk vs. Off-Talk for the different databases



II. TYPES OF SPEECH ANALYSIS

2.1 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], 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 [8]. 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 [9].

Figure 2: Speech Waves

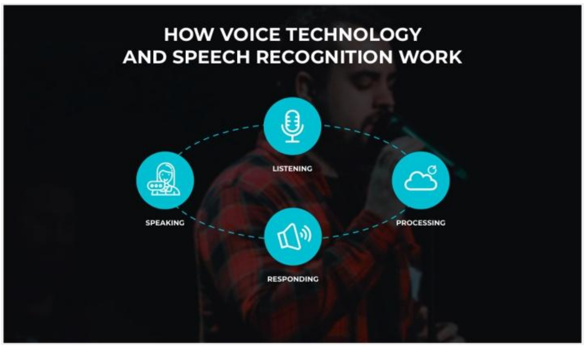
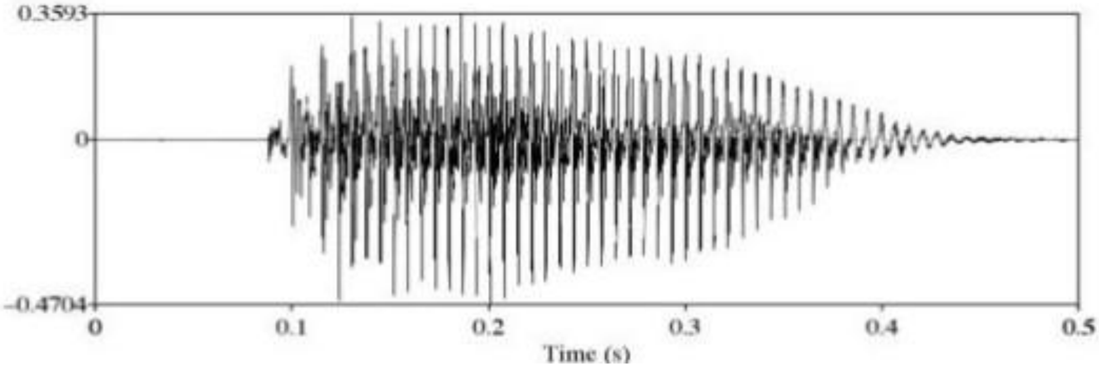


Figure 3: Speech Recognition

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.

2.2 Acoustical Speech: Any acoustic property of a speech sound that may be recorded and analyzed, as its fundamental frequency or formant structure. Acoustic is the study of the physical properties of speech**,** and aims to analyse sound wave signals that occur within speech through varying frequencies, amplitudes and durations. One way we can analyse the acoustic properties of speech sounds is through looking at a waveform. Pressure changes can be plotted on a waveform, which highlights the air particles being compressed and rarefied, creating sound waves that spread outwards[4]. A tuning fork being struck can provide an example of the pressure fluctuations in the air and how the air particles oscillate (move in one direction rhythmically) when we perceive sound.

Figure 3: A waveform of a vowel – Ogden 2009: 30



2.2.1 Frequency vs Amplitude: The frequency (pitch) and amplitude (‘loudness’ or intensity) of a sound can be analysed on a waveform. Frequency can be calculated through the number of cycles on a periodic waveform with a repeating pattern. The higher the number of cycles per second, the higher the frequency and perceived pitch. Frequency is usually expressed in Hertz (Hz)[7]. Analyzing the amplitude of a waveform tells us how intense or ‘loud’ a sound is, and how much the air particles deviate. It is conventionally expressed in decibels (dB). The x-axis on a waveform corresponds to the time frame in which the sound was produced (usually in seconds or milliseconds), and the y-axis represents the amplitude.

2.2.2 Sine Waves vs Complex Waves: Sine waves are waveforms that have very simple, regular repeating patterns. The number of ‘cycles’ in the waveform (the number of complete repetitions in the period waveform) reflects the number of times the vocal folds have opened within the time frame displayed. This is known as the fundamental frequency (f0), which is measured in Hertz (Hz). A frequency of 200Hz means that there are 200 hundred complete cycles per second within the waveform, so 200 times the vocal folds have opened. In reality, most speech sound waves have a rather complex pattern, and are known as complex waves. These are made up of two or more simple sine waves, and the fundamental frequency can also be calculated on complex waveforms by counting the number of cycles per second on a waveform.

2.2.3 Periodic vs Aperiodic sound waves: The types of speech sounds that would appear as a periodic sound wave are voiced sounds, such as vowels or nasals. Since such sounds have regularly repeating waveforms, they can also be decoded through 'Fourier analysis' which breaks down the component sine waves. This type of graph is called a spectrum, which does not measure time. Instead, the x-axis measures frequency, and the y-axis represents the sound pressure level. The fundamental frequency on this type of graph can be worked out by selecting the lowest frequency component of this complex wave[5]. This is usually the first complete peak on the spectrum. From this fundamental frequency peak, harmonics occur at evenly spaced integer multiples. Harmonics are known as the 'natural resonances' within the vocal tract, which are the amplified frequencies. On the spectrum, these correspond to each peak[5].

2.2.4 Voicing on a spectrogram: A fundamental frequency can be calculated due to the regular openings of the vocal folds as they vibrate. On a waveform, this would be highlighted by a periodic sound wave. On a spectrogram, there are two specific visual elements to look out for, which resemble a voiced sound. The first is the vertical striations (they look like vertical wavy lines on the spectrogram), which correspond to this opening of the vocal folds, and when air flows through them every time[4]. The other visual clue is the dark horizontal bands which are typical of vowels, approximants and nasals. These are called formants, which are the natural resonances of the vocal tract (earlier, they were described as harmonics)[5]. The size and shape of the vocal tract can be modified to allow these formants to vary. This can be done by changing the tongue position, lip position, etc.

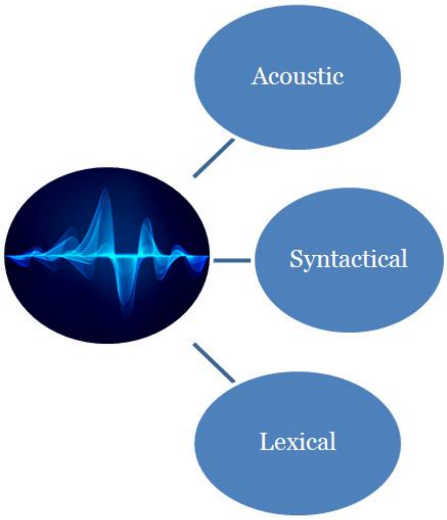
2.3 Syntactical Speech: We perform two stages of text analysis. The first stage is syntactical analysis which allows us to determine differences in the structure of human- and computer-addressed sentences. The underlying idea is that the syntax of machine-addressed speech possesses more structured patterns 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 [4] 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) [10]. The obtained syntactical attributes are language-dependent, however, they are much less sensitive to a specific domain in comparison with lexical features.

2.4 Lexical Speech: 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[1]. Lexicon gives the pronunciation of the words in that language. For the present system, as the Acoustic Models are in native languages the lexicon has to be generated in the native languages. To build the native Lexicon from the foreign Lexicon, the Artificial Neutral Network Phoneme Converter is used, which maps the phones of English to the phones of Indian Languages depending on the contextual information[11]. 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[4].

III. 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 4: Speech Analysis



IV. 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. 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 [2], 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 [1]. 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.

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]. 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.

[3]. 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.

[4]. B. Schulleret al., "The INTERSPEECH 2013 computational paralinguistic challenge: social signals, conflict, emotion, autism, "Proceedings INTERSPEECH 2013, Lyon, France, August 2013.

[5]. Stuart K.Card, Thomas P.Moran, Allen Newell, "Human –Computer Interaction", 1986.

[6]. Dan Diaper ,Neville Stanton, The Handbook of Task Analysis for "Human –Computer Interaction", 2003.

[7]. G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Process. Mag. 29, 82–97 (2012).

[8]. C.T. Ishi, S. Matsuda, T. Kanda, T. Jitsuhiro, H. Ishiguro, Nakamura S. Hagita N, Robust speech recognition system for communication robots in real environments, in 2006 6th IEEE-RAS International Conference on Humanoid Robots, Genoa (IEEE, 2006), pp. 340–345.

[9]. O. Mubin, J. Henderson, C. Bartneck, You just do not understand me! Speech recognition in human robot interaction, in Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), Aalborg (IEEE, New York, 2014), pp. 637–642.

[10]. 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), Portoroz (Slovenia), pp. 1826-1831, May 2016.

[11]. B. Yegnanarayana, S.P. Kishore, and A.V.N.S. Anjani, Neural network models for capturing probability distribution of training data, in Int. Conference on Cognitive and Neural Systems, (Boston), p. 6(A), 2000.