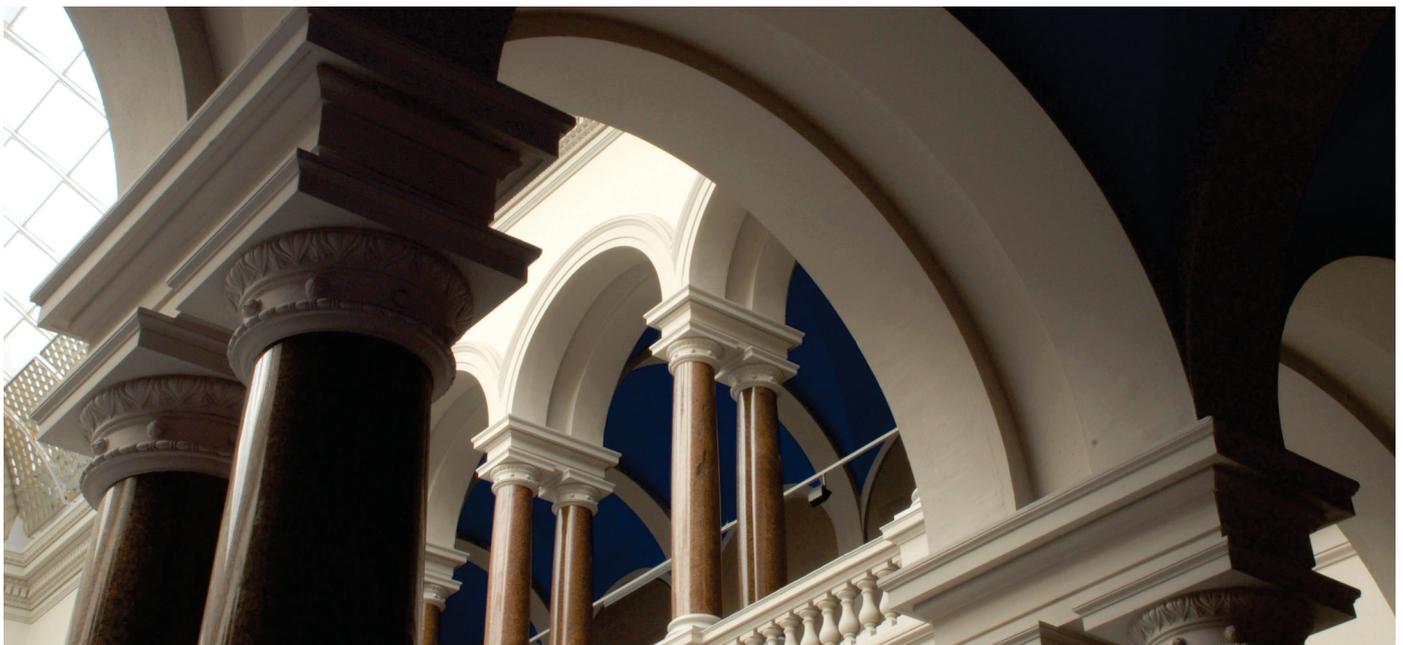


Athanasios Lykartsis, Margarita Kotti, Alexandros Papangelis, Yannis Stylianou

Prediction of Dialogue Success with Spectral and Rhythm Acoustic Features Using DNNS and SVMS

Conference paper | Accepted manuscript (Postprint)

This version is available at <https://doi.org/10.14279/depositonnce-9717>



Lykartsis, A., Kotti, M., Papangelis, A., & Stylianou, Y. (2018, December). Prediction of Dialogue Success with Spectral and Rhythm Acoustic Features Using DNNS and SVMS. 2018 IEEE Spoken Language Technology Workshop (SLT). <https://doi.org/10.1109/slt.2018.8639580>

Terms of Use

© © 2018 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

WISSEN IM ZENTRUM
UNIVERSITÄTSBIBLIOTHEK

Technische
Universität
Berlin

PREDICTION OF DIALOGUE SUCCESS WITH SPECTRAL AND RHYTHM ACOUSTIC FEATURES USING DNNS AND SVMs

Athanasios Lykartsis

Technische Universität Berlin
Audio Communication Group
Berlin, Germany

M. Kotti, A. Papangelis, Y. Stylianou†*

Toshiba LTD
Cambridge Research Laboratory
Cambridge, UK

ABSTRACT

In this paper we investigate the novel use of exclusively audio to predict whether a spoken dialogue will be successful or not, both in a subjective and in an objective manner. To achieve that, multiple spectral and rhythmic features are inputted to support vector machines and deep neural networks. We report results on data from 3267 spoken dialogues, using both the full user response as well as parts of it. Experiments show an average accuracy of 74% can be achieved using just 5 acoustic features, when analysing merely 1 user turn, which allows both a real-time but also a fairly accurate prediction of a dialogue successfulness only after one short interaction unit. From the features tested, those related to speech rate, signal energy and cepstrum are amongst the most informative. Results presented here outperform the state of the art in spoken dialogue success prediction through solely acoustic features.

Index Terms— dialogue success, acoustic features, deep neural networks, support vector machines

1. INTRODUCTION

In the last years, there has been much research into Statistical Spoken Dialogue Systems (SDS), as they increasingly find more and more applications such as assisted living [1], phone banking [2], intelligent virtual agents [3] and health care [4], to name just a few. Although many of the aforementioned systems for these applications use speech to communicate with the user, the audio channel is typically not directly used to enhance the user experience, for example by predicting dialogue success, as is the case in this paper. However, this acoustic information can be very helpful, especially at the early dialogue stages, as it can allow the system to engage in a different strategy, such as not repeating the same question if the dialogue is predicted unsuccessful, adapting the way the answer is presented to the user, adjusting the synthesised speech parameters, or even forwarding the user to a human agent. An auto-

matic way to measure task success would also be useful for evaluating conversations among humans, e.g., for evaluating agents in a call center as the authors of [5] underline.

In the dialogue success prediction area, various linguistic and belief state features [6], automatic speech recognition features [7], and acoustic features [8] can be used to obtain some information concerning the current state of the interaction. In human-computer dialogues, predicting the task success after just a first few turns of the conversation could avoid disappointment if the conversation is predicted unsuccessful. However, in the dialogue successfulness prediction area, most of the efforts use exclusively the linguistic channel [5] or the dialogue logs [9], disregarding the audio part altogether. In the work done previously by various authors [5, 7, 8, 10], it has been shown that employing all of these feature categories, separately or together, can provide satisfactory classification accuracy when analysing a dataset of interactions, where users communicate with such an automatic system. In a limited number of works the speech channel is used to measure user satisfaction. However, most efforts so far take into account the systems voice, instead of the users. For example, in [11] and in [12] it was found that automatic speech recognition (ASR) quality is a key component of user satisfaction.

There are certain possible shortcomings when ignoring the audio channel. For example, using belief state features to predict dialogue success implies that we have a robust mechanism to track the users' goal as it evolves over time, for each domain, i.e. subject, the system operates in. However, the user goal is most of the time hidden. Therefore, our aim is to infer if the goal is met based only on readily available attributes which can be extracted from speech. We observe that the exclusive use of the speech signal per se has numerous advantages: Firstly, the speech signal is often already analysed for ASR purposes (in the sense of feature extraction). Hence, by utilizing this information for predicting dialogue successfulness, we are receiving an extra gain at a very low (or even none, depending on the extracted features) computational cost. Secondly, speech features are domain-independent, in the sense that their values are not affected by the subject the conversation is about. If the user is dissatisfied

*A. Papangelis was with Toshiba during submission and is now at Uber.

†Y. Stylianou was with Toshiba at the time of submission, is currently at Apple Inc., and is also with the Department of Computer Science, University of Crete, Heraklion, Greece.

with the system, the quality of his or her voice is affected in the same way, whether the subject of the conversation is booking a hotel room, purchasing a product or receiving information. Thirdly, speech signal features are not affected by the ontology size of an SDS, in the sense that the number of slots and values of an ontology is not connected with the user’s cognitive state while interacting with the system. For all the reasons mentioned above, we choose to take a closer look at the gains that acoustic¹ features alone, extracted solely from the human user voice, can provide when predicting dialogue success. Such features have been used with great success in many other audio classification tasks [13].

In the next section, related work is given. Contributions are presented in Section 3 and the method is detailed in Section 4. We continue with the presentation of the results in Section 5, their discussion in Section 6 and sum up with conclusions and directions for future work.

2. RELATED WORK

The number of research effort that take into account acoustic features to improve the dialogue experience is limited, since, as already mentioned, most works disregard the audio channel. For example, in [14] features derived from the ASR output and the preceding system action are used, disregarding the audio channel, in order to predict the user satisfaction. A limited number of systems, like [15] have used multiple features and sparingly some acoustic descriptors for predicting problematic dialogues. More current attempts have been made, e.g., in [16], [17] and [18] to detect irregularities in dialogues using acoustic features. In other works, [19, 20] prosodic features were also used to detect quotations and dialogue acts, respectively. Prediction of dialogue success has also been carried out, using energy, cepstrum, and pitch-related features [8, 10]. Also, [21] have implemented a multimodal system and extracted a very broad range of acoustic features to predict the user’s cognitive state. However, predicting the cognitive state is out of the scope of this paper since the target values provided with the presented database refer to success. To the best of our knowledge, there are no works using a wider range of acoustic features to predict the success of the dialogue. To that end, we aim to use simple, thus easy to compute at low cost and without much data, acoustic features, suitable for a real-time SDS. Such features have shown promising results for alternative applications [22, 23, 24, 13].

3. CONTRIBUTIONS

Contributions of this paper include: i) The extraction of a broader pool of acoustic features, consisting a substantially larger dataset than that already found in related literature [8,

¹It must be mentioned here, that the term *acoustic features* in this paper refers to features extracted directly from the audio stream, as opposed to acoustic/ASR-related ones.

10], and the identification of the acoustic features most informative and useful for the task, ii) determining the minimum necessary number of dialogue turns (and, accordingly, length) to achieve a satisfactory prediction accuracy, and iii) whether it is more feasible to predict user satisfaction (i.e., if the user reported that they were content with course and outcome of the dialogue and interaction) or the dialogue’s objective success (if the goal set for the interaction was actually met).

4. METHOD

The method consists of the pipeline shown in Fig. 1. More information about the individual steps will be given in the following subsections. All experiments reported were performed on a Linux CentOS system using python 2.7.5.

4.1. Audio Data

The dataset consists of a series of human-machine interactions. Specifically, the interactions are calls between Amazon mechanical turk workers concerned with finding appropriate laptops using an SDS. Each person was given a set of preferences prior to the session and was asked to retrieve a laptop that fulfils those preferences. Regarding the laptop ontology, it comprises 9 slots, i.e., database attributes, 6 of which are informable, i.e., information the user provides with a mean of 3 values per slot, and 3 of which are requestable, i.e., information the system requests. The person then interacted with the dialogue system until the item was retrieved or the person hung up. We used a dataset of a total length of approximately one hour, comprising of 3267 dialogues. The audio files comprise light background noise and some artifacts, while being of telephone speech quality (with a sampling rate of 16000 Hz). The files are segmented per dialogue turn, i.e., every part of the utterance, in which the user interacts with the system, is available, as well as the full-length interaction. Therefore, using only some initial dialogue turns for classification would potentially speed up the classification procedure and lead an earlier recognition of dialogue success. We experimented with the whole dialogue files, as well as with their parts, i.e., turns, up to a maximum of 5 dialogue turns. Transcripts and details about the nature of the dataset can be found in [10]. An example of a dialogue is given in Table 1.

The labels for classification comprise two groups: *subjective*, denoting the success of the dialogue as expressed in the feedback of the user to the question “Did you find all the information you were looking for?”; and *objective* ones, resulting as the outcome of the system response offering the user a laptop adhering to the requirements provided by the user in the beginning. For example, if the goal is to find a low-price, light-weight laptop, if the final item offered to the user did in fact fulfil those two requirements, the dialogue was labelled as objectively successful. The two label sets were largely non-identical, i.e., the subjective satisfaction was not a con-

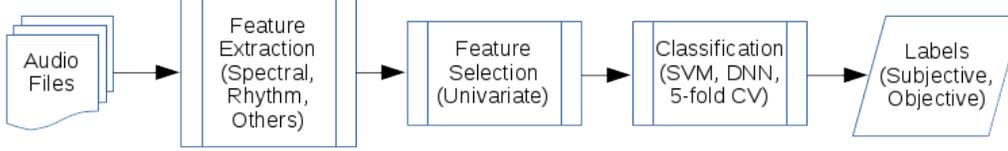


Fig. 1. acoustic feature extraction, selection and classification pipeline.

sequence of objective success, or vice-versa. Since the initial 3267 dialogues were highly unbalanced both for the case of objective (with a prior of 64% for the unsuccessful case) and for that of subjective (with a prior of 73% for the successful case) labels, we create balanced versions to avoid spurious high accuracy by randomly under-sampling the larger class to the size of the smaller. This results to a sample size of 1786 for the subjective and 2404 for the objective label data.

Criteria: You want to find a laptop in the large hard drive range that is in the mid weight range and has a battery in the exceptional rating category. Make sure you get the name and family the laptop belongs to.

System: Hello, I am here to help you find a suitable laptop. Now please tell me your requirements.

User: With large hard drive, ehh, mid weight, and exceptional battery.

System: Would you like your battery rating to be standard, good, or exceptional?

User: Exceptional!

System: Regarding weight, would you like something mid weight or heavy?

User: Mid.

System: The tecra-z50-a-106 is in the mid weight range, is in the moderate price range and has an exceptional battery rating.

User: What's family?

System: This laptop is called tecra-z50-a-106.

User: Thank you, good bye.

Table 1. The transcript of a dialogue between an Amazon mechanical turk worker and the statistical SDS for the case of a dialogue that was objectively unsuccessful (family info was not provided by the system, hence not meeting the given requirements), but subjectively successful.

As can be seen from the example dialogue of Table 1, it was quite common in the database to provide most of the information you need to ask the system during the first turn. In the example dialogue that would be: "With large hard drive, ehh, mid weight, and exceptional battery." The rest of the turns are short, such as solely one word reply ("Exceptional" or "Mid" in the example dialogue). We expect that feature extraction at the utterance level would be less stable compared to sentence-level feature extraction. Longer interactions refer to the 1st turn, due to the nature of the dataset.

4.2. Feature Extraction

Since we wanted to primarily investigate the effectiveness of low-level acoustic information (which is computationally easy and cheap to obtain, this being an advantage for a possible real-time application), we extracted three categories of features, using the librosa library [25].

The first category comprises features providing spectral information. In all cases, X refers to the signal magnitude spectrum (as a result of a DFT), K refers to the total number of frequency bins, for which k is a running index, whereas N refers to the total number of time frames, for which n is a running index. Those are:

- **The three first Mel-frequency cepstral coefficients (MFCCs)** (excluding the zeroth coefficient), as a compact spectral representation (here, X' , K' and k' refer to the mel-warped magnitude spectrum):

$$MFCC^{1-3}(n) = \sum_{k'=1}^{K'} \log(|X'(k', n)| \cdot \cos(j(k - 0.5) \frac{\pi}{K'})) \quad (1)$$

- **Spectral flux**, indicating spectral shape change:

$$SF(n) = \frac{\sqrt{\sum_{k=0}^{K/2-1} (|X(k, n)| - |X(k, n-1)|)^2}}{K/2} \quad (2)$$

- **Spectral flatness**, indicating noisiness or tonalness:

$$SFL(n) = \frac{\frac{K/2}{\sqrt{\prod_{k=0}^{K/2-1} |X(k, n)|}}}{\frac{2}{K} \sum_{k=0}^{K/2-1} |X(k, n)|} \quad (3)$$

- **Spectral contrast**, indicating difference between high and low frequency content, (refer to [26] for more information on this feature's extraction).

- **Spectral centroid**, the spectral center of weight,

$$SC(n) = \frac{\sum_{k=0}^{K/2-1} k \cdot |X(k, n)|^2}{\sum_{k=0}^{K/2-1} |X(k, n)|^2} \quad (4)$$

- **Spectral rolloff**, indicating the limit frequency, from which on only very little energy is existent in the spectrum (κ is here equal to 0.85, i the frequency index of a specific bin):

$$SR(n) = i \quad \text{with} \quad \sum_{k=0}^i |X(k, n)| = \kappa \cdot \sum_{k=0}^{K/2-1} |X(k, n)| \quad (5)$$

- **Spectral bandwidth**, describing the signal effective bandwidth.

$$SB(n) = \sum_{k=0}^{K/2-1} |X(k, n)| * (i - SC) \quad (6)$$

Also, temporal features are computed, such as:

- **Root mean square (RMS) energy value**, an indicator for the energy of the signal (the equation here refers to one frame of samples; i_s is the sample start index, i_e end index inside the block, n refers to the frame index, K is the sample amount in the frame):

$$RMS(n) = \sqrt{\frac{1}{K} \sum_{i=i_s(n)}^{i=i_e(n)} x(i)^2} \quad (7)$$

- **Zero crossing rate**, showing tonalness or noisiness (the indexes are as for the previous temporal feature, the RMS, and f_s refers to the sample frequency):

$$ZCR(n) = \frac{2 \cdot (i_e(n) - i_s(n))}{f_s \cdot \sum_{i=i_s(n)}^{i=i_e(n)} |sign[x(i)] - sign[x(i-1)]|} \quad (8)$$

- **Pitch**, i.e., for voice, the fundamental frequency, as estimated in librosa using parabolic interpolation [27].

For more information on all those features, the interested reader can refer to [28]. Using those features has the goal of capturing energy, tonal and spectral information providing, e.g., clues to the speaker's voice characteristics or their tone of voice (e.g. loud vs. quiet). Those might change during a call or have different values depending on the speaker's cognitive state, which we assume to be related to dialogue success. All of the above mentioned features are based on either a Short-Time Fourier transform (STFT) or on a frame-based processing with the same parameters. For this we use a *hann* window with length of 32ms, and an overlap of 50%. These values resulted after a parameter grid search, but are also characteristic for similar speech processing tasks, as they capture instantaneous values of the signal in a time frame, where the signal can be deemed to be stationary. All other parameters of the features (wherever that applies) are the default ones provided by the library. The feature values over the whole audio track are then aggregated using their mean and standard deviation values, as well as their skewness and kurtosis (4 subfeatures).

Furthermore, we use beat histogram features as in [22], based on different novelty functions. This procedure has the following steps: i) Calculate the temporal trajectory of 7 of the instantaneous, frame-based features mentioned (flux, flatness, centroid, rolloff, bandwidth, RMS and pitch); ii) Compute the first difference function over time for them, stressing changes in the respective features; iii) Create a beat histogram by using an autocorrelation function on each of the novelty

functions, and retaining periodicities in the area between 0.5 and 10 Hz, since that is the approximate expected frequency area for speech rates and rhythms [29].

Equation (4.2) shows the creation of a beat histogram BH_{IF} for an instantaneous feature IF with a novelty function $NF = diff(IF)$, where σ^2 is the sample variance, μ the sample mean, n the index for the total amount of instantaneous frame values for the feature N , and k a shifting index:

$$BH_{IF} = \frac{1}{\sigma^2 \cdot N} \sum_{n=1}^{N-k} (NF_n - \mu)(NF_{n+k} - \mu) \quad (9)$$

At the end, from the resulting beat histograms (for each novelty function), we extract the aggregate values and the rate, i.e. the most prominent periodicity value, resulting in 5 subfeatures per novelty function. For brevity reasons, we do not include any figures here concerning those novelty functions and their beat histograms, but the interested reader may refer to [22] for more information. With those features we intend to capture information relating to the speaker's rhythmic voice characteristics, for example the regularity of speech or the speech rate. These might also be influenced through the satisfaction of the user with the SDS, and they are expected to be useful.

The total feature count is 84, comprising $3 * 4 = 12$ MFCCs, $4 * 6 = 24$ spectral, $3 * 4 = 12$ temporal/energy features and $5 * 7 + 1 = 36$ rhythm features (one of them being the tempo value extracted through the tempogram toolbox with the same STFT parameters as before, s. [30]). It is reminded that we consider 4 subfeatures per instantaneous feature and 5 subfeatures for each novelty function based beat histogram, that is why the number of features is multiplied by 4, respectively 5 in the above calculation.

4.3. Classification and Feature Selection

Classification is performed using two classifiers, which are the ones who led to the optimal results: a support vector machine with radial basis function kernel (RBF-SVM) classifier through the sklearn module [31], as well as a deep neural network (DNN) of 2 hidden layers in tensorflow [32].

The mode of classification is binary, i.e. successful/non-successful, for both the objective and the subjective success prediction case. In order to test the stability of the proposed approach, 5-fold cross validation is used and the average accuracy and standard deviation of it across the 5 folds is reported in the next section. We used a development set design: The samples were separated in a training, development and test set in a ratio of 0.8 – 0.1 – 0.1. The test set samples were kept absolutely separate from the rest, the development set samples were used for the classifier parameter tuning. Prior to classification, the features are z-score standardised, separately for each set. After performing initial experiments with the 84

features mentioned in 4.2, we decided to employ methods of dimensionality reduction to check if we can achieve the same accuracy, but with a reduced feature set, thus rendering our approach computationally less intensive. To that end, a filter feature selection method [33] is implemented through the sklearn package, namely univariate feature selection, so as to capture both linear and non-linear dependencies between the features and the labels. We experiment with retaining the best n features by gradually reducing their number, resulting to a minimal set of just 5 features.

Concerning the classifier setting, we experimented with several alternatives to reach the best accuracy. For the RBF-SVM, the hyperparameters are determined through grid search in the development set, resulting in $C = 2$ and a γ of 0.2 for the 5 features and 0.012 for the 84 features. To find the optimal architecture for the DNN, several alternatives have been tested. For 2 hidden layers we tried having 68 neurons for the first hidden layer and 24 for the second; or 50 neurons for the first layer and 100 for the second; or finally 24 neurons for the first layer and then 68 for the second. Out all those, the best performing architecture for both cases was found to be the neural network where the first hidden layer had 64 neurons and the second one 28 neurons. Additional architectures having more hidden layers, such as 5, 5, 5, 5 or 89, 28, 11, 5 or finally 68, 24, 12 were suffering from data starvation, thus leading to poor results. Hence, we resorted to the 64, 28 neurons architecture. The Adam algorithm [34] was used as optimiser. A dropout value of 0.9, a learning rate of 0.01 and softsign as the activation function were used.

5. EXPERIMENTAL RESULTS

Figures 2 and 3 show the mean accuracy and standard deviation over the 5 folds of the cross validation achieved for the SVMs and the DNNs using the minimal set of 5 features, for an increasing amount of turns (one to five, where the number of turns is cumulative, i.e., the 2 turn case is turn 1 and turn 2 together; and then altogether; Due to time issues, we have not experimented with single turns apart from the first one, both for the subjective- and the objective-label data. These results refer to the performance on the *test* set. Finally, Table 3 provides an overview of the results mentioned before, but also includes the accuracy results on the development set. Concerning the performed feature selection, the most informative features are the standard deviation of the three MFCCs and the tempo and mean of the RMS-based beat histograms.

6. DISCUSSION

The best performance achieved for the both the SVM and the DNN case is that of considering only one turn for the subjective labels. In that case, the reported SVM accuracy is $74\% \pm 15\%$, whereas the corresponding accuracy for the DNN is $72\% \pm 2\%$. For the objective case, the best SVM

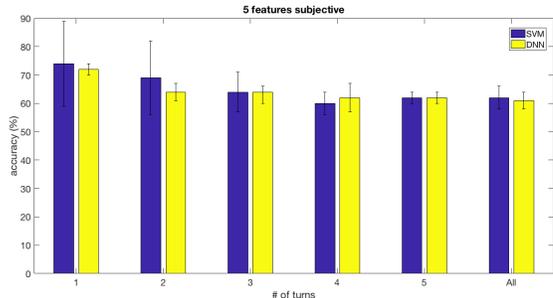


Fig. 2. SVM and DNN mean and standard deviation of accuracy for different turns, subjective labels, test set.

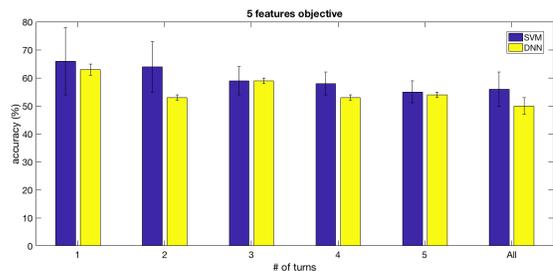


Fig. 3. SVM and DNN mean and standard deviation of accuracy for different turns, objective labels, test set.

accuracy is $66\% \pm 12\%$ when only one turn is considered, for the DNN the best performance is achieved when also 1 turn is considered and is equal to $63\% \pm 2\%$. Hence, it can be deduced that the objective success of the dialogue is more difficult to predict. This could be partially due to the acoustic features capturing more information about the speaker's cognitive state or voice characteristics, which could give indications as to whether they are satisfied or not. Additionally, the DNN architecture applied for the subjective results seems not to provide optimal results for the objective case. Also, the SVM exhibits much larger variance for the accuracy between the folds. Also, the results on the development set (seen in table 3) are in almost any case worse than on the test set. The overall worse results using the DNN can be attributed to data deprivation.

Regarding the turns used, it can be observed that there is not a very pronounced effect of the audio information size used, but it can be said that accuracy decreases with the rising amount of turns generally. This may be partially attributed to the telephone speech quality of the data introducing more redundant information the longer the processed become, which might be reflected into their aggregate values; it is, however, a very positive result, since one quick interaction part can be almost as meaningful as a longer file, essentially reducing substantially the amount of data necessary to make good predictions. This is most important for on-line, real-time sys-

Setting/Number of Turns	1	2	3	4	5	All
SVM development set, subjective labels	68 (12)	64 (11)	60 (8)	59 (5)	58 (5)	59 (5)
SVM development set, objective labels	66 (10)	62 (8)	58 (7)	56 (6)	55 (4)	55 (4)
DNN development set, subjective labels	67 (1)	61 (1)	60 (3)	59 (4)	60 (2)	64 (1)
DNN development set, objective labels	60 (1)	61 (1)	50 (1)	50 (1)	50 (1)	50 (1)

Table 2. Classification results, development set, average accuracy and standard deviation in percent (rounded). The number of turns reported is cumulative.

Setting/Number of Turns	1	2	3	4	5	All
SVM test set, subjective labels	74 (15)	69 (13)	64 (7)	60 (4)	62 (2)	62 (4)
SVM test set, objective labels	66 (12)	64 (9)	59 (5)	58 (4)	55 (4)	56 (6)
DNN test set, subjective labels	72 (2)	64 (3)	64 (3)	62 (5)	62 (2)	61 (3)
DNN test set, objective labels	63 (2)	53 (1)	59 (1)	53 (1)	54 (1)	50 (4)

Table 3. Classification results, test set, average accuracy and standard deviation in percent (rounded). The number of turns reported is cumulative.

tems since meaningful predictions can be reached using just some seconds of the captured speech. Another possible reason could be that over the whole phone call, the cognitive state of the user may vary significantly. For example, the call may start with a neutral user that gets satisfied for the first few turns while providing a lot of information and then dissatisfied towards the end of the call, providing responses with fewer words. User cognitive state evolves over time during a phone call and this is what is expected. For example, in [35] the authors found that emotions evolve as time progresses, whereas the authors of [36] that analyse call centre dialogues, try to predict a category per utterance, rather than over the whole dialogue.

Regarding feature selection, experiments also took place when all the 84 features are used. In that case, the results are deteriorated for most of the cases. For example, for the already reported case of subjective results when 1 turn is considered, the SVM accuracy is $65\% \pm 10\%$, whereas the DNN accuracy is $72\% \pm 2\%$. Moreover, the feature selection results corroborate what has been found in previous studies [22], namely that basic statistics of spectral features and the tempo are informative descriptors of the speaker characteristics. For example, speech rate as predicted through the RMS tempo, seems to provide important information, as it can probably capture partially if the speaker becomes agitated or not [37].

In total, it could be argued that the findings can be very helpful in providing assistance to an SDS in a fully automatic way: Solely by using a pre-trained model on a medium-sized amount of data, a real-time analysis of the audio signal (using fast algorithms and few features) could allow an estimate of the dialogue success with high probability (74%) only after one short interaction unit. Those results, although not directly comparable due to the different datasets and/or features used in other studies [8], [10], [7], provide a higher improvement compared to the baseline, given the dataset priors. Specifically, [8] reported results 6% better than the prior for the sub-

jective labels as well as 6.5% for the objective labels. For [10] only the objective labels are provided, with a performance 2% higher than the prior, while the results from [7] are not comparable, as no prior is given and the evaluation measure is F-Score. In our case, we achieve a 24% increase (compared to the 50% prior) in accuracy for the subjective labels, and 16% for the objective ones.

7. CONCLUSIONS AND FUTURE WORK

In this paper various acoustic features have been extracted, some of them used for the first time in this context, and then selected just a small subset of 5 of them for the prediction of dialogue success. It was possible to show that a satisfactory classification accuracy can be achieved when using even one dialogue turn and that the subjective success is more accurately predicted. The features related to RMS and MFCCs are the most informative. Our results are slightly better than the state-of-the-art for dialogue success prediction based on audio features only, but there is room for further improvement.

In the future, we aim to consider more data and extract further acoustic features in real-time, using the essential library [38], as well as state-of-the-art methods such as i-vectors which might be useful in identifying speaker characteristics even for short utterances and telephone quality speech [39, 40]. Another possibility is to use PCA-reduced (mel-)spectrograms as input to a convolutional neural network (CNN), as was similarly done in [41] for onset detection. Furthermore, information pertinent to the emotional state of the user [42] could provide cues to user satisfaction. Finally, user studies capturing the perceived dialogue success after each turn could be conducted in order to examine how the user state changes with every dialogue turn.

8. REFERENCES

- [1] François Portet, Michel Vacher, Caroline Golanski, Camille Roux, and Brigitte Meillon, “Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects,” *Personal and Ubiquitous Computing*, vol. 17, no. 1, pp. 127–144, 2013.
- [2] Bayan AbuShawar and Eric Atwell, “Usefulness, localizability, humanness, and language-benefit: additional evaluation criteria for natural language dialogue systems,” *International Journal of Speech Technology*, vol. 19, no. 2, pp. 373–383, 2016.
- [3] Yoichi Matsuyama, Arjun Bhardwaj, Ran Zhao, Oscar Romeo, Sushma Akoju, and Justine Cassell, “Socially-aware animated intelligent personal assistant agent,” in *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2016, pp. 224–227.
- [4] Mandy Korpusik and James Glass, “Spoken language understanding for a nutrition dialogue system,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1450–1461, 2017.
- [5] David Reitter and Johanna D Moore, “Predicting success in dialogue,” 2007.
- [6] Stefan Steidl, Christian Hacker, Christine Ruff, Anton Batliner, Elmar Nöth, and Jürgen Haas, “Looking at the last two turns, i’d say this dialogue is doomed—measuring dialogue success,” in *International Conference on Text, Speech and Dialogue*, 2004, pp. 629–636.
- [7] David Vandyke, Pei-Hao Su, Milica Gasic, Nikola Mrkšić, Tsung-Hsien Wen, and Steve Young, “Multi-domain dialogue success classifiers for policy training,” in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 763–770.
- [8] Alexandros Papangelis, Margarita Kotti, and Yannis Stylianou, “Predicting dialogue success, naturalness, and length with acoustic features,” in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, 2017, pp. 5010–5014.
- [9] Sebastian Möller, Klaus-Peter Engelbrecht, and Robert Schleicher, “Predicting the quality and usability of spoken dialogue services,” *Speech Communication*, vol. 50, no. 8-9, pp. 730–744, 2008.
- [10] Margarita Kotti, Alexandros Papangelis, and Yannis Stylianou, “Will this dialogue be unsuccessful? prediction using audio features,” in *ICTIR 2017 Workshop on Search-Oriented Conversational AI*, 2017.
- [11] Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan, “Automatic online evaluation of intelligent assistants,” in *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2015, pp. 506–516.
- [12] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos, “Understanding user satisfaction with intelligent assistants,” in *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, New York, NY, USA, 2016, pp. 121–130.
- [13] Francesc Alías, Joan Claudi Socoró, and Xavier Sevilano, “A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds,” *Applied Sciences*, vol. 6, no. 5, pp. 143, 2016.
- [14] Stefan Ultes, Paweł Budzianowski, Inigo Casanueva, Nikola Mrkšić, Lina Rojas-Barahona, Pei-Hao Su, Tsung-Hsien Wen, Milica Gašić, and Steve Young, “Domain-independent user satisfaction reward estimation for dialogue policy learning,” in *Eighteenth Annual Conference of the International Speech Communication Association*, 2017, pp. 1721–1725.
- [15] Marilyn Walker, Irene Langkilde, Jerry Wright, Allen Gorin, and Diane Litman, “Learning to predict problematic situations in a spoken dialogue system: experiments with how may i help you?,” in *1st North American chapter of the Association for Computational Linguistics conference*, 2000, pp. 210–217.
- [16] Joseph Tepperman, David Traum, and Shrikanth Narayanan, “Yeah right: Sarcasm recognition for spoken dialogue systems,” in *Ninth International Conference on Spoken Language Processing*, 2006.
- [17] Narichika Nomoto, Hirokazu Masataki, Osamu Yoshioka, and Satoshi Takahashi, “Detection of anger emotion in dialog speech using prosody feature and temporal relation of utterances,” in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.
- [18] José Lopes, Giampiero Salvi, Gabriel Skantze, Alberto Abad, Joakim Gustafson, Fernando Batista, Raveesh Meena, and Isabel Trancoso, “Detecting repetitions in spoken dialogue systems using phonetic distances,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
- [19] Alessandra Cervone, Catherine Lai, Silvia Pareti, and Peter Bell, “Towards automatic detection of reported speech in dialogue using prosodic cues,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

- [20] Raul Fernandez and Rosalind W Picard, "Dialog act classification from prosodic features using support vector machines," in *Speech Prosody 2002, International Conference*, 2002.
- [21] Leimin Tian, Johanna Moore, and Catherine Lai, "Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*, 2016, pp. 565–572.
- [22] Athanasios Lykartsis and Stefan Weinzierl, "Using the beat histogram for speech rhythm description and language identification," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
- [23] Mayank Bhargava and Tim Polzehl, "Improving automatic emotion recognition from speech using rhythm and temporal feature," *arXiv preprint arXiv:1303.1761*, 2013.
- [24] Athanasios Lykartsis, Stefan Weinzierl, and Volker Dellwo, "Speaker identification for swiss german with spectral and rhythm features," in *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio*, 2017.
- [25] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in python," in *14th python in science conference*, 2015, pp. 18–25.
- [26] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai, "Music type classification by spectral contrast feature," in *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*. IEEE, 2002, vol. 1, pp. 113–116.
- [27] Julius O Smith et al., *Spectral audio signal processing*, vol. 1334027739, W3K, 2011.
- [28] Alexander Lerch, *An introduction to audio content analysis: Applications in signal processing and music informatics*, John Wiley & Sons, 2012.
- [29] Aniruddh D Patel, *Music, language, and the brain*, Oxford university press, 2010.
- [30] Peter Grosche and Meinard Müller, "Tempogram toolbox: Matlab implementations for tempo and pulse analysis of music recordings," in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR), Miami, FL, USA*, 2011, pp. 24–28.
- [31] Fabian Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [32] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al., "Tensorflow: A system for large-scale machine learning," in *OSDI*, 2016, vol. 16, pp. 265–283.
- [33] Isabelle Guyon and André Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.
- [34] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization. arxiv.org," 2014.
- [35] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.
- [36] Laurence Devillers and Laurence Vidrascu, "Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs," in *Ninth International Conference on Spoken Language Processing*, 2006.
- [37] Volker Dellwo, Adrian Leemann, and Marie-José Kolly, "Speaker idiosyncratic rhythmic features in the speech signal," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.
- [38] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez Gutiérrez, Sankalp Gulati, Perfecto Herrera Boyer, Oscar Mayor, Gerard Roma Trepas, Justin Salamon, José Ricardo Zapata González, and Xavier Serra, "Essentia: An audio analysis library for music information retrieval," in *Britto A, Gouyon F, Dixon S, editors. 14th Conference of the International Society for Music Information Retrieval*, 2013, pp. 493–498.
- [39] Ahilan Kanagasundaram, Robbie Vogt, David B Dean, Sridha Sridharan, and Michael W Mason, "I-vector based speaker recognition on short utterances," in *12th Annual Conference of the International Speech Communication Association*, 2011, pp. 2341–2344.
- [40] Mohammed Senoussaoui, Patrick Kenny, Najim Dehak, and Pierre Dumouchel, "An i-vector extractor suitable for speaker recognition with both microphone and telephone speech.," in *Odyssey*, 2010, p. 6.
- [41] Jan Schlüter and Sebastian Böck, "Musical onset detection with convolutional neural networks," in *6th International Workshop on Machine Learning and Music (MML), Prague, Czech Republic*, 2013.
- [42] Dimitrios Ververidis and Constantine Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech communication*, vol. 48, no. 9, pp. 1162–1181, 2006.