Will this dialogue be unsuccessful? Prediction using audio features

Margarita Kotti
Toshiba Research Europe Ltd
Cambridge Research Lab
margarita.kotti@crl.toshiba.co.uk

Alexandros Papangelis
Toshiba Research Europe Ltd
Cambridge Research Lab
alex.papangelis@crl.toshiba.co.uk

Yannis Stylianou*
Toshiba Research Europe Ltd
Cambridge Research Lab
yannis.stylianou@crl.toshiba.co.uk

ABSTRACT

This paper proposes a method to improve statistical spoken dialogue systems and specifically it aims to provide a way for early detection of unsuccessful dialogues using the audio stream. If an interaction is predicted as unsuccessful, this information could be used to update the policy or to forward the call to a human agent. A dataset of interactions between Amazon Mechanical Turk workers and a statistical spoken dialogue system is used. A total of 702 dialogues are recorded. Then, mel-frequency cepstral coefficients (MFCCs) are extracted from the user's speech signal, forming a "feature image" which is then given as input to a convolutional neural network comprising of 9 layers. The reported accuracy is 94.7%, and the system manages to predict that a dialogue will be unsuccessful for the 97.9% of the cases. With respect to accuracy, there is an improvement of 17.2%, compared to our previous work on predicting dialogue quality. We observe that for our task, convolutional neural networks can model temporal correlations given context information and that the cepstral domain is a useful and compact representation for convolutional neural networks.

KEYWORDS

statistical spoken dialogue systems, dialogue success, speech analysis, convolutional neural networks, audio channel

ACM Reference format:
Margarita Kotti, Alexandros Papangelis, and Yannis Stylianou. 2017. Will this dialogue be unsuccessful? Prediction using audio features. In *Proceedings of ICTIR' 17 Workshop on Search-Oriented Conversational AI (SCAI' 2017), Amsterdam, Netherlands, October 1, 2017,* 6 pages.
https://doi.org/

1 INTRODUCTION

Statistical Spoken Dialogue Systems (SDS) form a very active research area due to their many applications, such as phone banking, tutoring, assisted living, intelligent virtual agents, human-robot interaction, health care, and speech-driven interactive mobile applications. However, one major roadblock in SDS prototyping is the significant effort required for estimating the performance of a dialogue system. For example, dialogue success is difficult to determine, since real users are often unwilling to extend the interaction in order to give feedback or they do not even have a clear target into mind when interacting with the SDS. Using crowd-sourced subjects and then asking for their subjective rating has also proven to be problematic with respect to consistency [8] [25]. To overcome those drawbacks an innovative way of predicting success in spoken dialogue systems using the audio channel of the user is introduced.

To the best of the authors' knowledge there has been no research towards the direction of exploiting the audio channel per se for predicting dialogue quality, although there are some studies that propose the use of the emotion of the speech an indirect quality measure [18]. Most efforts within the quality evaluation community are towards defining user satisfaction metrics [32] or estimators of the dialogue performance [5]. Regarding the means to achieve the aforementioned quality estimation, researchers have resorted to the use of dialogue features, such as speech acts [33] or belief state, act, and turn-level features [24] [29] or linguistic features [23]. In all those cases, the audio channel is neglected altogether.

However, the invaluable contribution of the audio channel towards the improvement of SDS experience can be seen in previous works that exploit the audio channel, although with alternative objectives, outside the scope of predicting dialogue success. For example, the system in [6] uses pitch, energy and duration features to discriminate between 8 dialogue acts. The ultimate aim is to understand the structure of spoken language. In [4], prosodic and timing features are used to understand dialogue structure. In an additional work [26], prosodic and spectral features are used to detect sarcasm aiming to create a dialogue agent that is able to understand sarcasm. Additionally, phonetic distances have been used in [16] to locate repetition as a symptom of problematic communication. More recently, the authors of [14] calculated the mean intensity, jitter, shimmer, noise to harmonic ratio and speaking rate. Then, they matched the Text-to-Speech output to the input for each feature and showed that the synthesised English speech was better liked. In all those works, the speech signal has been used to augment the statistical SDS' efficiency.

Using the audio channel for predicting the dialogue success has distinct advantages, namely:

- it introduces no overheads. The user is interacting with the system anyway. So, subsequent errors due to problematic subjective ratings are avoided.
- speech signals properties are not dependent of the subject of the conversation for the case of information seeking (also known as slot-filling) problem, i.e. they are domain-independent; so there is no need or cost to retrain the proposed system

*Also with the Department of Computer Science, University of Crete, Greece

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
ICTIR' 17 Workshop on Search-Oriented Conversational AI (SCAI' 2017), October 1, 2017, Amsterdam, Netherlands
© 2017 Copyright held by the owner/author(s).
ACM ISBN .
https://doi.org/

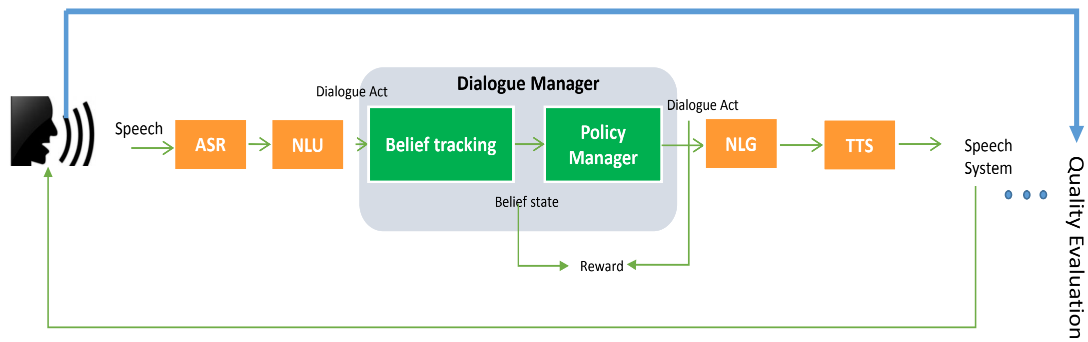

Figure 1: A typical statistical SDS architecture.

in order to use it to a different domain, for example to move from hotel booking to sales.
- it is not affected by the size, defined as the number of actions and slots of the SDS in hand
- acoustic feature computation may introduce no overheads, since a plethora of such features are computed for ASR anyway.

In the work presented here, the detection of unsuccessful dialogues is done by feeding Mel-frequency cepstral coefficients (MFCCs) of the user's utterance to convolutional neural networks (CNNs). A total of 702 dialogues are used, 70% of which are used for training and the rest 30% for testing. Those dialogues are collected using a statistical SDS which interacts with users through Amazon Mechanical Turk [1]. The aim is to locate the unsuccessful dialogues [13] [31] as early as possible, since that would be of practical meaning to real-world applications, so as to forward to a human agent for example or to update the policy to ask for alternative information for instance via implementing a user model [15] or for updating the reward function. Also, if a dialogue is detected not to be progressing effectively, other parts of ths SSDS, other than the policy can be updated. For example, the NLG may be more detailed or the TTS can be adapted to the user, thus exhibiting entainment [14]. To achieve the early intervention, just the first 5 turns are retained. From those turns, MFCCs are extracted and then concatenated to form a matrix of 12 (coefficients) x 5 (turns). Those can be seen as "feature images", which are then fed as input to CNNs. The latter neural networks have proved to be efficient initially for image processing tasks, [11] [28] [30], whereas lately also their suitability for audio classification problems has been studied [2] [3] [9] [21] [27]. The CNN architecture chosen for this paper includes two convolution layers and one max-pooling layer. A final overall accuracy of 94.7% is reached, whereas 97.9% of the unsuccessful testing calls are classified correctly.

The rest of the paper is organised as follows: in Section 2 a description of the research question is given along with the details of the dataset we study and then in Section 3 the extracted features along with the CNN architecture is detailed. Results are presented in Section 4; conclusions and future work are detailed in Section 5.

2 PROBLEM FORMULATION

2.1 Research Question

A typical statistical SDS architecture can be seen in Figure 1. In conventional systems, the speech signal is used just for Automatic Speech Recognition (ASR) transcription and then it is discarded. What this paper proposes is an alternative exploitation of the speech signal: a system that takes as input acoustic features, extracted directly and real-time from the user's speech and produces as output a prediction on whether the dialogue at hand will be unsuccessful. This is depicted by the blue line in Figure 1.

In specific, the aim is to predict well in advance when a dialogue will be a failure. Here with failure we mean that either i) the user hanged up, or ii) the interaction reached the maximum number of turns, that is 30 for this specific scenario, or that iii) the retrieved item is not one that complies with the set of preferences given to the user for the specific dialogue. Accordingly, here success is the task success and not related to the user satisfaction. By predicting the unsuccessful dialogues, interactions that are predicted to fail can be handled in a different way, as explained in the Introduction Section. The concern is on predicting the unsuccessful dialogues, rather that the successful ones, since the unsuccessful dialogues demand an intervention. That is, the cost of misclassification between the two categories is not the same. Also, an early intervention is speculated to increase more the user satisfaction, rather than a late one. For that reason, we decided to use just the first few turns of each dialogue. An additional advantage of retaining more than one turns is related to the system's ability to recognise recovery from error. Training over a number of turns, rather than just one turn, is more possible to lead to a system that learns to identify cases when a failed turn recovery is possible. The number of turns is determined by the histogram presented in Figure 2, where the blue bars correspond to the case of successful and the red ones to the case of unsuccessful dialogues. It can be deducted from Figure 2 that not all the turns have the same information burden for predicting the success. The median number of turns is 6 and distribution of the number of call is left-skewed, with a limited number of calls having a high turn number. In addition, the calls that have more than 5 turns are mostly unsuccessful. In fact, the median number of successful turns is 5. For all those reasons, we retain just the first 5 turns.

So, the research question is whether or not the first 5 turns can be predictive of the dialogue failure.

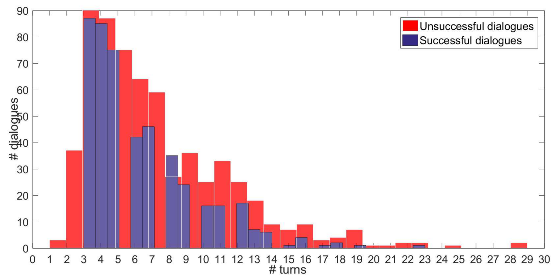

Figure 2: Overlaid histograms of the number of turns for successful and unsuccesful dialogues.

2.2 Dataset

The dataset consists of 702 spoken dialogues between humans and a statistical SDS. The users were recruited via Amazon mechanical turk. The interactions of the Amazon mechanical turk workers concerned finding appropriate Toshiba laptops. Each person, therefore, was given a set of preferences, as can be seen in upper part of Tables 1 and 2. Regarding the Toshiba laptop ontology it comprises 6 informable slots with a mena of 3 values per slot, and 3 requestable slots. The person then interacted with the dialogue system until the item was retrieved or until the person decided to hung up. The dialogue is deemed successful if the Toshiba laptop retrieved at the end of the interaction meets the criteria originally presented to the user. An example of a successful dialogue can be seen in Table 1, whereas the case of an unsuccessful dialogue is demonstrated in Table 2. Regarding the technical details of the audio files, they are stored in a wave format, the sampling rate is 16KHz, and the bit rate is 256kbps mono.

Criteria: You want a laptop for business use that has a standard battery rating. Make sure you get its weight, and its dimension.

System: Hello, I am here to help you find a suitable laptop. Now please tell me your requirements.
User: Eh rm...
System: Sorry, I can't hear you.
User: I want a laptop for business use with standard battery.
System: Satellite Pro C70-B-34L is a very good computer for business computing. It has a standard battery.
User: What is the weight?
System: The weight of Satellite Pro C70-B-34L is 2.7 kg.
User: What is its dimension?
System: Satellite Pro C70-B-34L is a well designed computer, it's diagonal is about 19.4 inches.
User: Rm...
System: Satellite Pro C70-B-34L is a very good computer for business computing.
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 successful dialogue.

Criteria: You want a laptop that has a large hard drive and an exceptional battery rating. You don't care the purpose of use of the laptop. Make sure you get its weight, and its dimension.

System: Hello, I am here to help you find a suitable laptop. Now please tell me your requirements.
User: I want an expensive laptop with a large hard drive.
System: Model number A-15N in the Portege family R30 series, is a very good computer. It has a five hundred GB hard drive, and its price is £1070.
User: And what is its weight?
System: Model number A-15N in the Portege family R30 series, is a very good computer. It has a five hundred GB hard drive, and its price is £1070.
User: How much does it weight?
System: The price of A-15N Portege is £1070.
User: What is its weight?
System: Portege-R30-A-15N, is a very good computer. It has a five hundred GB hard drive, and its price is £1070.
User: How much does it weight?
System: The price of A-15N Portege is £1070.
User: Thank you, good bye.

Table 2: The transcript of a dialogue between an Amazon mechanical turk worker and the statistical SDS for the case of an unsuccessful dialogue.

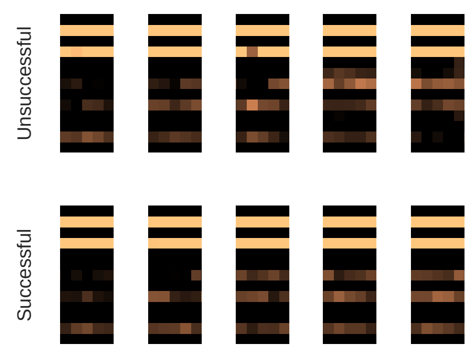

Figure 3: Example "feature images": the MFCC heatmaps for the first 5 turns. The x axis of each heatmap is the number of turn; there are 5 turns depicted. The y axis of the heatmap is the number of MFCC; the first row corresponds to MFCC 1 and the last one to MFCC 13. In this figure, the top row is for the unsuccessful calls and bottom row for the successful ones.

3 METHOD

To tackle with the research question of whether or not the speaker utterances derived by first 5 turns can be predictive of the dialogue failure, we propose to extract low level audio features which are then feed as input to neural networks, specifically convolutional neural networks (CNNs). In specific, we extracted 13 MFCCs. For the MFCC implementation we used the Auditory Toolbox [22], with a 512-sample FFT, 256-sample window, and the lowest frequency at 133.33 Hz. The use of MFCCs is a classic choice in speech processing, and specifically for the human-machine dialogue case [7] [12] [17]. For each utterance we then retain the average value for each MFCC coefficient over time. Averaging is a first solution; in the future we are aiming to use more sophisticated feature summarisation techniques, that construct the feature space automatically, such as evolutionary algorithms. However, even just the simple averaging can still provide basic information about changes in speaking style. For example, it can detect a change in the frequency bands utilised, which can be indicative for example of an user that is becoming impatient. Also, a magnitude increase can be suggestive of a user that is becoming angry, or bored, if a decrease is observed. Additionally, this feature can facilitate capturing the increase in the arousal among successive utterances, where an abrupt change among successive utterances can be an indicator of a communication problem.

We then produce "feature images" by creating a matrix of 13 mean MFCCs along the 5 first turns. Accordingly, those "feature images" have a size of 13x5. For visualisation purposes, those 13x5 feature images are presented as heat-maps in Figure 3 for both the case of unsuccessful (top row) as well as the successful (bottom row) calls.

Those "feature images" are then fed as input to CNNs. By using a CNN we aim to find a structure to those "feature images". Convolutional deep neural networks have consistently shown more robustness to noise and background contamination [19]. One more natural advantage of using CNN is that it's invariant against translations of the variations in frequencies.

Here, the CNN is developed on Matlab Version 2016b (The MathWorks). Specifically, the architecture we resorted is:

- An input layer of size 13 x 5.
- A convolution layer of 100 filters (or feature maps) of size 2 x 2.
- A rectified linear unit (RELU) layer
- A pooling layer of size 4 x 2.
- A convolution layer of 100 filters of size 3 x 2.
- A RELU layer
- A fully connected layer with an output size of 2.
- A softmax layer
- A fully connected layer with an output size of 2 (as is the number of classes).

This architecture can be graphically seen in Figure 4. Regarding the pooling layer, it performs max pooling and the stride is 4 x 2. Regarding the options for training the neural network, stochastic gradient descent with momentum is used as the learning algorithm, the learning rate is 0.02 the batch-size is set to 200, and the maximum number of epochs is 5000.

4 EXPERIMENTAL RESULTS

We used the aforementioned 702 dialogues, 70% of which are used for training and the rest 30% for testing. The size of the corpus is not large, however, this paper aims to serve as a proof of concept and we are currently working actively towards collecting more data. This is neither trivial nor low-cost. The dataset in unbalanced, with 90% of the dialogues being unsuccessful and the remaining 10% successful. This can be attributed to i) the challenging nature

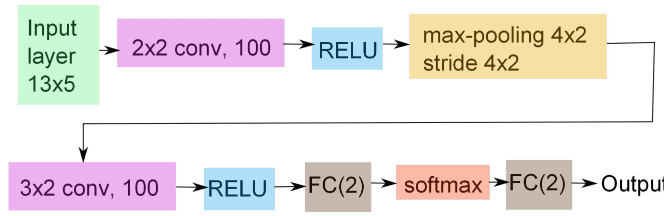

Figure 4: The proposed CNN architecture: 9 layers.

Table 3: Confusion matrix for the CNN using the 13 x 5 MFCC "feature images" when considering the first 5 user utterances.

		Predicted outcome	
		Unsuccessful	Successful
True outcome	Unsuccessful	185	4
	Successful	7	12

of the dialogue management task and ii) to the fact that parts of the dataset were built with an suboptimal policy (that was then improved via interaction with the user). Care was taken so that the ratio of 90%/10% unsuccessful/successful dialogues is retained in both the training and the test set. Dialogues are selected in a random manner. This is in line with the purpose of the paper to detect the unsuccessful dialogues that need intervention, rather than discriminating between successful and unsuccessful ones. To address the problem of imbalanced dataset, we followed the advice found in the literature [10]: several figures-of-merit are reported, so as to capture as many aspects of the system performance as possible; the focus is on detecting just one class (the unsuccessful calls); and the balanced case is reported (through random sub-sampling) for the standard experimental configuration.

The confusion matrix when using the MFCC "feature images" for the first 5 user utterances as input to CNNs described by Fig. 4 can be seen in Table 3. The overall classification accuracy is 94.7%, F_1 measure is 68.6% whereas for the case of unsuccessful dialogues 97.9% of the testing calls are classified correctly and the remaining 2.1% is predicted as successful. For the case of successful dialogues, although this is not the focus of this work, 63.2% of the cases are correctly classified and the remaining 36.8%, although successful is predicted as unsuccessful. If we reduce the test set to make it balanced, so as to have the same number of unsuccessful and successful calls, then accuracy equals 71.1%. Given the nature of this work, that is i) to early detect an interaction that will be unsuccessful and ii) it's better to predict and successful call as unsuccessful rather than vice versa, the presented results are promising. Commending on the number of layers of the CNN, the proposed architecture was found to be the more efficient. Addition of more convolution layers led to performance degradation.

A second set of experiments took place to examine the case of using less user turns, that is producing predictions using less data. In this case, two sub-cases are identified: in the first case, we consider only using the user turns 1 to 3 and for the second case the turns 3 to 5. Accordingly, the size of the "feature images" reduces from 13 x 5 to 13 x 3 for both cases. The rest of the experimental

Table 4: Confusion matrix for the CNN using the 13 x 3 MFCC "feature images" when considering the first 3 user utterances.

True outcome		Predicted outcome: Unsuccessful	Predicted outcome: Successful
	Unsuccessful	182	7
	Successful	9	10

protocol remains the same, besides the CNN architecture that has to be slightly modified so as to accommodate the updated input size. The difference with the one presented in Figure 4, is that the second convolution layer has a size of 2 x 2; and the pooling size is 4 x 1 instead of 4 x 2, since less turns are available.

When the first 3 turns are considered, the confusion matrix can be seen in Table 4. The overall classification accuracy is 92.3% and F_1 measure is 55.6%. For the case of unsuccessful dialogues 96.3% of the testing calls are classified correctly and the remaining 3.7% is predicted as successful. It can be deducted that when the first 3 user turns are exploited instead of the first 5 ones, there is a relatively small drop of 1.6%, in the accuracy of correctly predicting the unsuccessful calls.

When the last 3 turns are considered, that is the turns number 3, 4 and 5, the confusion matrix can be seen in Table 5. The overall classification accuracy is 94.2%, F_1 measure is 68.4%, whereas for the case of unsuccessful dialogues 96.8% of the testing calls are classified correctly and the remaining 3.2% is predicted as successful. When the last 3 user turns are taken into account instead of the first 5 ones, there is a relatively small drop of 1.%, in the accuracy of correctly predicting the unsuccessful dialogues. Overall, it can be deducted that it is feasible to reduce the number of turns taken into account, here from 5 to 3, with a relatively small drop in the accuracy.

A third set of experiments refers to using less MFCCs. That is based on the fact that the first 3 MFCCs as well as the last one seem to obtain values that are almost the same for both unsuccessful and successful dialogues. Accordingly, we want to test their discriminative power. If those coefficients are omitted, then the size of the "feature images" reduces to 9 x 3, where 9 is the original MFCCs minus the 1st, 2nd, 3rd, and 13th coefficients. The rest of the experimental protocol remains the same, besides the CNN architecture that has to be slightly modified so as to accommodate the updated input size. Now the second convolution layer has a size of 2 x 2. The respective confusion matrix can be seen in Table 6. The overall classification accuracy is 92.3%, F_1 measure is 46.7%, whereas for the case of unsuccessful dialogues 97.9% of the testing calls are classified correctly and the remaining 2.1% is predicted as successful. Compared to the previous case of using all the MFCCs, it can be seen that using less MFCCs does not have an impact on the case of unsuccessful dialogues, however, it has an adverse effect for the case of successful ones. Given that MFCC computation is not an expensive one and in many SDS systems such audio features are computed anyway as part of the ASR, we propose to use the larger "feature images" that contain all 13 MFCCs.

Let us note that although for this specific dataset we utilised specifically the first 5 turns as a case-study, that is not to be the case for a real SSDS. In that case, the turns window can slide over

Table 5: Confusion matrix for the CNN using the 13 x 3 MFCC "feature images" when considering the last 3 user utterances.

True outcome		Predicted outcome: Unsuccessful	Predicted outcome: Successful
	Unsuccessful	183	6
	Successful	6	13

Table 6: Confusion matrix for the CNN using the 9 x 5 MFCC "feature images" when considering the first 5 user utterances.

True outcome		Predicted outcome: Unsuccessful	Predicted outcome: Successful
	Unsuccessful	185	4
	Successful	12	7

the dialogue and produce results as the the dialogue evolves. So, for example, not only the turns 1-5 can be used to detect a problem at the 5th turn, but also the turns 2-6, to detect a problem at the 6th turn, 3-7 to detect a problem at the 7th turn, 4-8 to detect a problem at the 8th turn etc. The reason why for this specific work we confined ourselves to the first 5 turns is the nature of the dataset. As already stated, the dataset is a small one, collected not with the optimal policy. Accordingly, as can be seen in Figure 2, there were not many dialogues where we could consider for example turns 4-8. However, this is not a limitation, and the method can be applied over a sliding window along the dialogue.

Comparison with previous work is not available, due to the novel area of this study. To the best of the author's knowledge, the only previous attempt has the one made by the authors of this paper in [20]. In the case of [20], a larger dataset of 1456 calls was used. However, here we had to select the 702 calls that included at least 5 turns. In [20], instead of the first few turns, all the available user turns are taken into account. Moreover, the audio features are the root mean square and the pitch, for both of which statistical properties are extracted, such as mean, minimum, maximum. To take context into account, additional features between consecutive turns are calculated as well. The best performing classifier is the Gaussian Process Reggressor and the respective accuracy equals 77.5%. Accordingly, the proposed system accomplishes an absolute 17.2% improvement for the case of 13 x 5 MFCC "feature images".

To discuss the limitations of this work, this system refers to a slot-filling dialogue system that works on one specific domain, here Toshiba laptops. Although it is expected to be transferable to other slot-filling problems, such as hotel search or flight booking, the case of an open-domain system will possibly require a more complex adaptation, such as noise reduction, use of a greater number and more diverse audio features and a dynamic number of turns to be taken into account.

5 CONCLUSIONS AND FUTURE WORK

This paper tackles with the novel area of predicting if a statistical SDS will fail, using only the first few user utterances. The problem of estimating the accuracy of a dialogue system is an active research area, however, to the best of the authors' knowledge, there is no other research group that uses directly and exclusively the audio channel. Specifically, we constructed "feature images" of consecutive MFCCs that were then given as input to CNNs. A high accuracy of correctly predicting 97.9% of the testing calls as unsuccessful is reported. It has also been shown that the system can produce reliable results, even with less context, i.e. less user utterances. Thus informative decisions can be made with as few as just 3 turns.

In the future, we plan to exploit a larger audio dataset. Given that our SSDS is becoming more efficient, we expect to have a better balanced dataset, to avoid over-fitting as well. We also aim to use spectral domain features such as linear frequency spectograms or mel spectrograms, since those are a natural fit for CNNs in terms of natural interpretation and frequency discrimination. Finally, we aim to incorporate the unsuccessful call prediction to the policy manager to find a new way to optimise the system act, in case that the call is predicted to be unsuccessful. Also, we plan challenge our target value, that is the successful or unsuccessful dialogue and examine the case of more fine-grained classes, i.e. the degree of success. Additionally, one way to go forward is to take into account not only the dialogue success, but also the user satisfaction.

REFERENCES

[1] 2017. Amazon Mechanical Turks, Artificial Intelligence. (June 2017). https://www.mturk.com/mturk/welcome

[2] Ossama Abdel-Hamid, Abdel rahman Mohamed, Hui. Jiang, Li Deng, Gerald Penn, and Dong Yu. 2014. Convolutional Neural Networks for Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22, 10 (October 2014), 1533–1545. https://doi.org/10.1109/TASLP.2014.2339736

[3] Ossama Abdel-Hamid, Abdel rahman Mohamed, Hui. Jiang, and Gerald Penn. 2012. Applying Convolutional Neural Networks concepts to hybrid NN-HMM model for speech recognition. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4277–4280. https://doi.org/10.1109/ICASSP.2012.6288864

[4] Alessandra Cervone, Catherine Lai, Silvia Pareti, and Peter Bell. 2015. Towards automatic detection of reported speech in dialogue using prosodic cues. In *INTERSPEECH '15*. 3061–3065.

[5] Layla El Asri, Romain Laroche, and Olivier Pietquin. 2014. Task completion transfer learning for reward inference. *Proc of AAAI Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence* (2014).

[6] Raul Fernandez and Rosalind W. Picard. 2002. Dialog act classification from prosodic features using support vector machines. In *Proceedings of Speech Prosody (SP '02)*. 291–294.

[7] Shinya Fujie, Kenta Fukushima, and Tetsunori Kobayashi. 2005. Back-channel feedback generation using linguistic and nonlinguistic information and its application to spoken dialogue system. In *9th European Conference on Speech Communication and Technology*. 889–892.

[8] Milica Gašić, Catherine Breslin, Matthew Henderson, Dongho Kim, Martin Szummer, Blaise Thomson, Pirros Tsiakoulis, and Steve Young. 2013. On-line policy optimisation of Bayesian spoken dialogue systems via human interaction. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 8367–8371. https://doi.org/10.1109/ICASSP.2013.6639297

[9] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel rahman Mohamed ;, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patric Nguyen, Tara N. Sainath, and Brian Kingsbury. 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine* 29, 6 (November 2012), 82–97. https://doi.org/10.1109/MSP.2012.2205597

[10] Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, et al. 2006. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering* 30, 1 (2006), 25–36.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, Fernando Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger (Eds.). Curran Associates, Inc., 1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[12] Akinobu Lee, Keisuke Nakamura, Ryuichi Nisimura, Hiroshi Saruwatari, and Kiyohiro Shikano. 2004. Noice robust real world spoken dialogue system using GMM based rejection of unintended inputs. (2004).

[13] Piroska Lendvai, Antal van den Bosch, Emiel Krahmer, and Marc Swerts. 2002. Multi-feature error detection in spoken dialogue systems. *Language and Computers* 45, 1 (2002), 163–178.

[14] Rivka Levitan, Štefan Beňuš, Ramiro H. Gálvez, Agustín Gravano, Florencia Savoretti, Marian Trnka, Andreas Weise, and Julia Hirschberg. 2016. Implementing Acoustic-Prosodic Entrainment in a Conversational Avatar. In *INTERSPEECH '16*. 1166–1170.

[15] Diane J. Litman and Shimei Pan. 2002. Designing and Evaluating an Adaptive Spoken Dialogue System. *User Modeling and User-Adapted Interaction* 12, 2 (2002), 111–137.

[16] Jose Lopes, Giampiero Salvi, Gabriel Skantze, Alberto Abad, Joakim Gustafson, Fernando Batista, Raveesh Meena, and Isabel Trancoso. 2015. Detecting Repetitions in Spoken Dialogue Systems Using Phonetic Distances. In *INTERSPEECH '15*. 1805–1809.

[17] Ramón López-Cózar, Ángel de la Torre, José C Segura, Antonio J Rubio, and Juan M López-Soler. 2002. A new method for testing dialogue systems based on simulations of real-world conditions.. In *INTERSPEECH'02*.

[18] Michael McTear, Zoraida Callejas, and David Griol. 2016. *Evaluating the Conversational Interface*. Springer International Publishing, Cham, 379–402. https://doi.org/10.1007/978-3-319-32967-3_17

[19] Vikramjit Mitra and Horacio Franco. 2015. Time-frequency convolutional networks for robust speech recognition.. In *2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 317–323.

[20] Alexandros Papangelis, Margarita Kotti, and Yannis Stylianou. 2017. Predicting dialogue success, naturalness, and length with acoustic features. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[21] Siddharth Sigtia, Emmanouil Benetos, and Simon Dixon. 2016. An End-to-end Neural Network for Polyphonic Piano Music Transcription. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.* 24, 5 (May 2016), 927–939. http://dl.acm.org/citation.cfm?id=2992480.2992488

[22] Malcolm Slaney. 1998. Auditory Toolbox, version 2. (1998).

[23] Stefan Steidl, Christian Hacker, Christine Ruff, Anton Batliner, Elmar Nöth, and Jürgen Haas. 2004. *Looking at the Last Two Turns, I'd Say This Dialogue Is Doomed – Measuring Dialogue Success*. 629–636.

[24] Pei-Hao Su, Vandyke David, Dongho Gašić, Kim, Nikola Mrkšić, Tsung-Hsien Wen, and Steve Young. 2015. Learning from real users: Rating dialogue success with neural networks for reinforcement learning in spoken dialogue systems. In *INTERSPEECH '15*. 2007–2011.

[25] Pei-Hao Su, Milica Gašić, Nikola Mrkšić, Lina M. Rojas Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. On-line Active Reward Learning for Policy Optimisation in Spoken Dialogue Systems. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 2431–2441. http://www.aclweb.org/anthology/P16-1230

[26] Joseph Tepperman, David Traum, and Shrikanth Narayanan. 2006. "Yeah right": Sarcasm recognition for spoken dialogue systems. In *INTERSPEECH '06*. 1838–1841.

[27] Aaron van den Oord, Sandler Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Anderw Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. *ArXiv e-prints* (Sept. 2016). arXiv:cs.SD/1609.03499

[28] Aäron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. 2016. Conditional Image Generation with PixelCNN Decoders. In *NIPS '16*.

[29] David Vandyke, Pei-Hao Su, Milica Gašić, Nikola Mrksic, Tsung-Hsien Wen, and Steve Young. 2015. Multi-domain dialogue success classifiers for policy training. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 763–770.

[30] Jure Žbontar and Yann LeCun. 2016. Stereo Matching by Training a Convolutional Neural Network to Compare Image Patches. *J. Mach. Learn. Res.* 17, 1 (Jan. 2016), 2287–2318.

[31] Marilyn Walker, Irene Langkilde, Jerry Wright, Allen Gorin, and Diane Litman. 2000. Learning to Predict Problematic Situations in a Spoken Dialogue System: Experiments with How May I Help You?. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference (NAACL 2000)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 210–217. http://dl.acm.org/citation.cfm?id=974305.974333

[32] Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 271–280.

[33] Marilyn A. Walker, Rebecca Passonneau, and Julie E. Boland. 2001. Quantitative and qualitative evaluation of DARPA Communicator spoken dialogue systems. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 515–522.