Prediction of User Emotion and Dialogue Success Using Audio Spectrograms and Convolutional Neural Networks

Athanasios Lykartsis
Audio Communication Group
TU Berlin
Germany
athanasios.lykartsis@tu-berlin.de

Margarita Kotti
Speech Technology Group
Toshiba Research Cambridge
United Kingdom
margarita.kotti@crl.toshiba.co.uk

Abstract

In this paper we aim to predict dialogue success and user satisfaction as well as emotion on a turn level. To achieve this, we investigate the use of spectrogram representations, extracted from audio files, in combination with several types of convolutional neural networks. The experiments were performed on the Let's Go V2 database, comprising 5065 audio files and having labels for subjective and objective dialogue turn success, as well as the emotional state of the user. Results show that by using only audio, it is possible to predict turn success with very high accuracy for all three labels (90%). The best performing input representation were 1s long mel-spectrograms in combination with a CNN with a bottleneck architecture. The resulting system has the potential to be used real-time. Our results significantly surpass the state of the art for dialogue success prediction based only on audio.

1 Introduction

Spoken Statistical Dialogue Systems (SDS) have gained much popularity in the last years, especially due to the widespread need for applications such as assisted living (Portet et al., 2013), phone banking (AbuShawar and Atwell, 2016), intelligent virtual agents (Matsuyama et al., 2016) and health care (Korpusik and Glass, 2017).

An important part of an SDS is spoken language, which is used to communicate directly with the virtual agent in order to pose questions and reply to the agent output. In a modular spoken SDS system, the speech part is converted to text through Automatic Speech Recognition Systems (ASR), which is then analysed using Natural Language Processing (NLP) methods. However, the audio part, which could be of low audio quality, is usually then discarded while the extracted text is fed forward to the SDS. In our view (and this is an important part of our motivation), when looking at dialogue success prediction, this can be seen as a waste of possible resources, since the speech part can contain useful information regarding the emotional state of the user, or verbal cues which can indicate if the user is satisfied with the system performance. The prediction or recognition of such cues can be very helpful for supporting a dialogue management system, which can make better assessments as to what the next steps should be. Taking this thought one step further, we want to assess if it is possible to predict dialogue success based only on the audio, in order to find a light-weight, real-time method to manage the user expectations and, eventually, to build more efficient and user-friendly spoken SDS. A final motivation of this work is that we wanted to experiment with spectrogram input representations and convolutional neural networks (CNNs) as classifiers. Although there have been several examples of such uses for other topics, especially in image processing (Krizhevsky et al., 2012) and music information retrieval (Schlüter and Böck, 2014; Schreiber and Müller, 2018), this approach remains underrepresented in the area of dialogue success prediction. Therefore, our research closes this gap and attempts to evaluate how well such approaches can function for dialogue success prediction.

Considering related works, the use of neural networks in the wider area of modular SDS has been gaining some popularity the last years. For example, neural networks have been utilised for dialogue state tracking. Korpusik and Glass (2018) use CNNs in order to track the user's goal over the whole dialogue without the use of handcrafted semantic dictionaries and achieve high accuracy for their task. Henderson et al. (2014) similarly employ recurrent neural networks to map the results of ASR directly to a dialogue state and also report high performance. Another approach

Proceedings of the SIGDial 2019 Conference, pages 336–344
Stockholm, Sweden, 11-13 September 2019.

(Zhao and Eskenazi, 2016) uses deep reinforcement learning to discover dialogue states and outperform a standard baseline. An additional deep reinforcement learning approach (Su et al., 2015) shows that using both RNNs and CNNs with turn level-features (non-audio) can be useful in predicting dialogue success. Research from Wen et al. (2016) shows that deep learning can be useful in creating more natural conversation task-oriented SDS, whereas Kim et al. (2016) use CNNs and RNNs for dialogue topic tracking.

As the listing of the related previous work shows, the use of neural networks with audio spectrograms or waveforms for the analysis of the audio part of the SDS and its consequent use for tasks such as dialogue success prediction has not been researched adequately. Only a limited number of papers (Papangelis et al., 2017; Kotti et al., 2017; Lykartsis et al., 2018) exist which explore the possibility of dialogue success prediction using audio features extracted from speech paired with standard machine learning techniques such as support vector machines.

These approaches have shown promising results, especially for creating a way to reliably estimate the user satisfaction. Additionally, they are able to do so in real-time or near-real-time and subsequently enable suitable next steps for the dialogue policy. Moreover, the recent success of deep learning approaches for audio tasks suggests that using these can bring an advantage: By exploiting input representations such as spectrograms, the estimation of task success can take place even at an ever finer time resolution level (e.g., very short audio frames), providing the possibility for even faster processing and reaction. Furthermore, data augmentation methods can provide a possibility to achieve higher accuracy rates.

Since CNNs combined with audio spectrograms as input have been shown to provide very good results in a multitude of tasks (for example for tempo estimation (Schreiber and Müller, 2018) and beat tracking (Schlüter and Böck, 2014)), we choose to employ them for the creation of an experimental setup for dialogue success prediction. In that sense, we frame our task as an emotion recognition one: As dialogue success is expected to show a high correlation with user satisfaction, which in turn is closely related with the user's emotional state, we investigated similar works using neural networks for speech emotion recognition.

Such works include those of Tzirakis et al. (2017) who use a Long-Short-Term-Memory (LSTM) network on top of a CNN in order to extract information and consider contextual information from raw audio data (waveforms), outperforming existing systems for speech emotion recognition. Similar work has been performed by Trigeorgis et al. (2016), where audio waveforms are used in combination with a CNN followed by an LSTM for speech emotion recognition, achieving high results for arousal and valence. In the work of Lee et al. (2017), a CNN is used to predict emotions based on speech spectrograms for a virtual elderly companion agent with very good results. Gu et al. (2018) create a multimodal framework with text and speech for emotion recognition. For the audio part, besides hand-crafted features, spectrograms with CNNs and LSTMs are used and fused with text features to predict 5 emotions and achieve better results than all other methods. Another interesting method comes from Yenigalla et al. (2018), where spectrograms of different sizes are used as an input for a CNN, achieving very good results for 4 emotional states. Neumann and Vu (2017) study the impacts of input features, signal length and speech type, using spectrogram or raw waveform input and CNNs, achieving state of the art results and reaching very useful conclusions for speech emotion recognition: input representation is not as important as the model architecture, which in turn is task and speech type specific. Fayek et al. (2015) also achieve very good results in speech emotion recognition using a simple deep neural network and spectrograms as input. A similar strategy is employed by Wang and Tashev (2017) for successful prediction of emotion, as well as gender and age on an utterance level, showing that even simple deep architectures can provide good results for speech emotion recognition. CNNs have also been used with success for general audio classification (Lee et al., 2009), which is a broader task, hinting at the suitability of this architecture for the task at hand in this paper. For this paper, we decided - for the sake of simplicity and due to the not enormous size of the dataset - to resort to only CNNs and determine which architectures, input representation forms and parameters provide good classification results for this task. Another reason for the use of CNNs is not only their aforementioned success in many tasks, but also the possibility to establish a

better understanding of the suitability of this approach for the task of dialogue success prediction. The latter is slightly different than speech emotion recognition per se because the user's emotional state is not the only factor that affects the final success label. Finally, this way we can establish a very fast and simple pipeline, which can also be used in a real-time setting to provide useful auxiliary information about the dialogue success, so as to inform the dialogue manager. This approach is compared to a baseline, involving hand-crafted audio features as in (Lykartsis et al., 2018), which have been shown to provide satisfactory results. Experiments are performed on the publicly available Let's Go V2 Database (Schmitt et al., 2012), which contains three kind of labels (for objective and subjective dialogue success and the emotional state of the user, for more information see 2.3).

This paper is structured as follows: In the next section the used methods are presented in depth, whereas in section 3, the results of the classification are shown and discussed. We close with conclusions and suggestions for future work.

2 Methods

2.1 Input Features

The input features chosen to be used for the CNN classifier in our case were mel-spectrograms (which can be seen as images summarizing the frequency content of a turn over time), extracted with the librosa python library (McFee et al., 2015). Mel-spectrograms have been used in a multitude of tasks for music information retrieval (Schlüter and Böck, 2014; Lidy and Schindler, 2016; Choi et al., 2017), as they are relatively simpler to calculate (in contrast to other transforms), while also providing a connection to human auditory perception through the use of the mel-scaling. Therefore, we reasoned that they could be a good basis for the task of dialogue success and speech emotion recognition. For an 1s long audio file we acquired a resulting 32 bins x 16 frames array (using the default librosa settings for spectrogram extraction that is a frame size of 92ms, a hann window and an overlap of 75% between consecutive frames). These settings are fairly standard for audio processing, as they allow a good temporal resolution but also a fair enough frequency resolution. We used this window size could mean that the speech segment is not necessarily stationary, but since we are looking for larger structure in the spectrogram (probably spanning several frames), this should not constitute a problem for the further processing (as it was also shown by our results). Using a shorter time window might produce even more temporally accurate spectrograms, but it would also require more computational resources. After conducting preliminary experiments with a window of 46ms, we could see that results were not improved, while at the same time requiring much more computational power for the spectrogram extraction. Therefore, we retained the window size of 92ms for all the further experiments. We also experimented with a length of 2s in order to see if longer (in the time domain) spectrograms would give better results - which can be seen as a trade-off between speed of processing (and therefore a close to real-time behavior of the classification/prediction system) and the accuracy of the prediction itself. This resulted to a 32 bins x 32 frames input array. We did not experiment with longer files, since most files in the Let's Go V2 database are not much longer than 2s (the average user turn duration is 1.5s with a standard deviation of 1.9s (Schmitt et al., 2012)). If the file is shorter than the selected analysis length, it is zero-padded at its end. All the files were of 8000 kHz sampling rate, no further preprocessing was performed, leading to a very lightweight pipeline, which is very close to a real-time processing. The goal of using these input features was to determine if a short spectrogram could suffice for providing good classification results.

2.2 Neural Nets/Classifiers

As mentioned in Section 1, we employ CNNs in this paper. The theory and inspiration for using CNNs can be found in Section 1. Specifically, we utilized Keras, which is based on the tensorflow library in python (Abadi et al., 2016). Keras has many advantages, such as that it is very effective, allowing for fast prototyping and training, even just by using CPUs (instead of GPUs). Inspired by similar experiments in other areas, we wished to test two different types of architectures:

- A standard **bottleneck architecture**, with 4 convolutional layers with 2-by-2 rectangular filters and a decreasing number of nodes (100-75-50-25), 2-by-2 max pooling and all activation functions being ReLU. This was followed by a batch normalization and 2 fully connected (FCN) layers (also with a decreas-

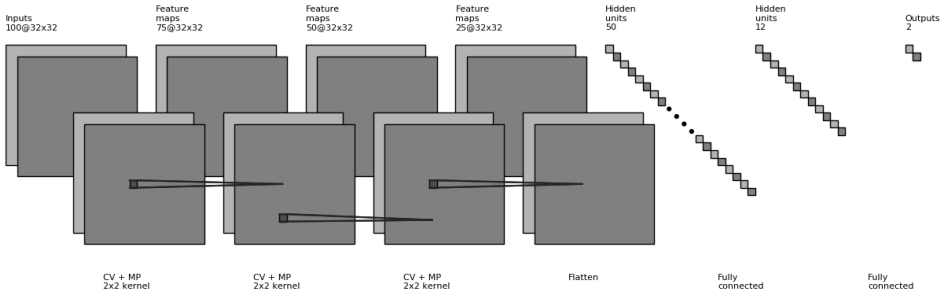

Figure 1: Bottleneck architecture flowchart diagram. For the details of the CNN, see the detailed architecture description in section 2.2

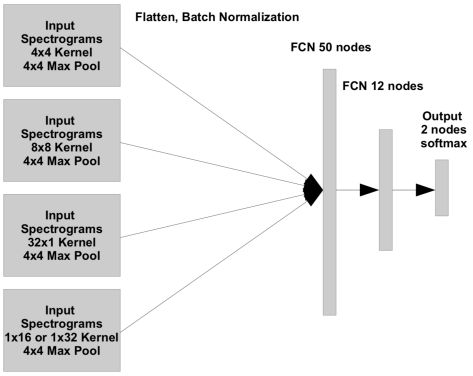

Figure 2: Parallel architecture flowchart diagram. For the details of the CNN, see the detailed architecture description in section 2.2

ing number of nodes, namely 50 and 12 and a dropout of 50%) and an output layer with softmax activation. The stride is always one, padding is always set at "same", so as that the output has the same length as the original input. As an optimizer, ADAM was used with a learning rate of 0.001, whereas a categorical cross entropy was utilized as the loss function. For this architecture we were inspired from (Tzirakis et al., 2017; Trigeorgis et al., 2016). The above architecture is depicted in Figure 1.

- A **parallel** CNN architecture: In this case, 4 input layers with 32 nodes and with different kernel sizes (a quadratic 4-by-4 kernel, a quadratic 8-by-8 kernel, a 1-dimensional 32-by-1 filter (for the mel-spectrogram frequency bins) and a 1-dimensional 1-by-16 or 1-by-32 filter (for the time frames, corresponding to the file length of 1s or 2s, respectively)) are processed in parallel and their output is combined (concatenated and flattened). In this case, the max-pooling is done in a 4-by-4 manner and the activations are also all ReLU. The combined output of the four parallel layers is batch-normalized fed into 2 FCN layers with 50 and 12 nodes with a dropout of 0.5 between them, followed by a 2-node output softmax layer. Same as before, the stride is always one and padding is set at "same". Also in this case, an ADAM optimizer was used (with a learning rate of 0.001), and a categorical cross entropy as a loss function. For this architecture we were inspired from the implementation in (Yenigalla et al., 2018) (using parallel layers) and from the one in (Schreiber and Müller, 2018), using one-dimensional filters. Our reasoning was that combining these two features, a powerful network could be constructed which would be able to learn features pertaining to emotional states of the user, as well as more specific signal features inherent in the spectrogram (such as the tempo of the utterance). The above architecture is depicted in Figure 2.

Finally, we implemented a baseline following the scheme in (Lykartsis et al., 2018), comprising 5 hand-crafted spectral and rhythmic features (the standard deviation of the three MFCCs and the tempo and mean of the RMS-based beat histograms) and featuring an SVM classifier with

$C = 2, \gamma = 1/N_{features}$ and an RBF kernel using the scikit learn python module. These parameters were kept the same as in the aforementioned publication, since they resulted via a grid search there for a dataset of similar audio quality, and for ensuring comparability between the studies.

The whole pipeline was developed and tested using python 3.6 on a Windows 7 OS with 8 GB of RAM and an Intel i5 quad core processor running at 3.2 GHz. Using this, extracting the spectrogram of a turn with librosa is achieved in under 1s, whereas training of the model for one epoch takes around 5s. After the model has been created, prediction, that is running the validation spectrogram via the trained CNN, is taking 0.1-0.2s (depending on the length of the turn), resulting into a near-real-time system. We refrained from using a development set due to the small size of the dataset and because our averaged results over the 3 validation folds should provide sufficient validity.

2.3 Dataset

The spoken dialogue corpus used in this study is based on the the CMU Let's Go Bus Information System (Schmitt et al., 2012) (from this point on referred to as the *Let's Go V2 dataset*). This has been developed by the university of Ulm in order to evaluate dialogue quality, user emotion and task success for an SDS which was used as an information system for bus itinerary search. The database contains 9083 system-user exchanges (to which we will refer as *interactions* in the following). For our experiments, we kept a total of 5065 audio files for the interactions, for which all labels where available, so as to be able to compare between the results using the different label sets.

Each interaction has been rated with three labels. The first is an **emotional label**, signifying the emotional state of the user. The label has four levels, ranging from non-angry to very angry. This label was assigned from the users themselves. Another label shows the **subjective** dialogue success, dubbed IQ (Interaction Quality) in the corpus annotation (Schmitt et al., 2012), indicating whether the user was satisfied with the interaction. This label ranges from satisfied to extremely unsatisfied and has five levels and was agreed on by three individual external raters. We refer to it here as *subjective label*. Finally, the **objective labels** indicate whether the goal of the dialogue was reached, i.e., the information looked for was actually provided by the system. This label also exists on an interaction level and has two levels (successful or not).

In order to simplify the classification, we choose to create a binary model which results from taking the most highly ranked result of each label set as the positive label, and all the other results pooled together as the negative label. In that way, it was possible to create an almost balanced dataset for the subjective labels (53% negative and 47% negative ones), but not for the other two labels sets (having correspondingly a distribution of 65% positive/35% negative for the emotional labels and 85% positive/15% negative for the objective samples). Therefore, we then created a balanced version of the dataset for the emotional and the objective labels by taking the smaller class and randomly choosing as many examples for the other class. The balanced subjective set contained 5065 samples, the balanced objective one 1146 and the balanced emotional one 3660 samples.

3 Results and Discussion

The results of the classification for all 3 labels can be seen in Tables 1 and 2 for the training and the validation set, respectively. The respective results for the baseline system can be seen in Table 3. The results reported here are the average accuracy over the three folds, followed by the loss of the network. The standard deviation of the accuracy over the folds is not reported, since it ranges from 0.5% to 1.5%, and can therefore be considered negligible, showing that the system is robust. It must be mentioned here again, that the basic unit of classification was the audio of the user turn, for which the labels are also available. The accuracy reported refers to the amount of correctly predicted labels for the user turns as a ratio of all turn classifications.

Concerning the effect of different parameters for the CNNs, the best parameter set was determined by 3 fold cross-validated grid search. The aforementioned cross-validation lead to the results reported in tables 1 and 2. We experimented with several values for the learning rate, the optimizer and the batch size. We observed an effect for better results with a learning rate of 0.001, a batch size of 8 and by using the ADAM optimizer. Finally, the results shown here were the result of 500 epochs long training procedure. We did not observe any improvement when training for longer time, and this is definitely an amount of training time which

Setting	1s		2s	
	Accuracy	Loss	Accuracy	Loss
Bottleneck Architecture, subjective labels	0.95	0.12	0.97	0.07
Bottleneck Architecture, objective labels	0.97	0.07	0.98	0.05
Bottleneck Architecture, emotional labels	0.98	0.06	0.98	0.05
Parallel Architecture, subjective labels	0.81	0.31	0.97	0.07
Parallel Architecture, objective labels	0.91	0.24	0.97	0.1
Parallel Architecture, emotional labels	0.92	0.17	0.97	0.07

Table 1: *Classification results, training set, average accuracy over 3 folds and corresponding loss for 1 and 2 s segments. All datasets are balanced, the prior is 0.5.*

Setting	1s		2s	
	Accuracy	Loss	Accuracy	Loss
Bottleneck Architecture, subjective labels	0.78	0.57	**0.9**	0.3
Bottleneck Architecture, objective labels	**0.9**	0.48	0.86	0.5
Bottleneck Architecture, emotional labels	**0.9**	0.33	0.86	0.5
Parallel Architecture, subjective labels	0.7	0.93	0.7	0.93
Parallel Architecture, objective labels	0.88	0.46	0.88	0.78
Parallel Architecture, emotional labels	0.82	0.69	0.74	1.25

Table 2: *Classification results, validation set, average accuracy over 3 folds and corresponding loss for 1 and 2 s segments. All datasets are balanced, the prior is 0.5.*

Setting	Whole turn
Baseline (SVM), subjective labels	0.59
Baseline (SVM), objective labels	0.57
Baseline (SVM), emotional labels	0.75

Table 3: *Classification results, baseline system, average accuracy over 3 folds. Features are extracted over the whole turn and aggregated. All datasets are balanced, the prior is 0.5.*

is very manageable on reasonably strong computation systems (see 2.2). With regards to the effect of the spectrograms' length, this did not seem to have a large effect on classification accuracy. In general, results were somewhat better for then 1s case. We therefore assume that in the case of less data, the length of the segment can be kept to a minimum value. These findings corroborate the results from Neumann and Vu (2017), which mention that the NN architecture is more important than the input representation form, at least in the context of speech emotion recognition.

Comparing the two architectures, the first architecture with the sequential layers has shown slightly better results. This might be due to the parallel models lacking the information to extract useful patterns, probably due to possible data deprivation. In total, the results are much higher than the ones produced from the baseline. We observed some important trends (with regards to the validation set results). The first architecture using the bottleneck structure has proven to be useful for all labels. This might be due to phonetic features in the spectrogram indicating task success being very concrete (such as "thank you", or the user's voice melody sinking) and therefore rendering a simpler structure to extract the features more suitable. Between the different label types, the emotional and objective label sets show somewhat better results when using smaller lengths, showing that for the subjective labels, a greater length is essential for the CNN to extracting more relevant information. The parallel layer architecture has shown to be useful for the objective labels. This is probably due to the higher complexity of predicting an objective task success from purely sound data. Additionally, the turn length does not seem to play an important role, which might mean that for more complex architectures, less information length can be sufficient to achieve good accuracy. All in all, the parallel architecture was somewhat less performant than the bottleneck one, which shows that for these data, simpler structures are more useful.

In general, the results were very positive and surpass results on similar datasets which are state-of-the-art: The maximum accuracy on the validation set, for the subjective labels, achieved using only sound files was 90%, which surpasses

the best results in (Lykartsis et al., 2018) by 16%. However, it must be noted, that the datasets used are slightly different, in the sense that the task is a different one (finding the right laptop vs. finding the right itinerary while interacting with an SDS). Also, in (Lykartsis et al., 2018), both the subjective and the objective labels were provided by the user, but in the Let's Go V2 system, those were provided by external raters, as well as having a different resolution (labeled turns instead of full dialogues). Therefore, the results not directly comparable, but the research question is the same. Furthermore, the length and audio quality of the recordings is very similar, so that it can be claimed that using mel-spectrograms as input and CNNs as classifiers provides a successful and computationally not too intensive way to achieve emotion detection and dialogue success prediction only from audio. We are therefore optimistic, that with more training data, we could build sounder models which can generalize better and build on the tendencies observed here, achieving even better results. Finally, the results achieved in (Lykartsis et al., 2018), that smaller files are more suitable for higher accuracy, are also observable here.

In comparison to other studies which used the Let's Go V2 dataset, two works have been found in the literature, that of Schmitt et al. (2011) and that of Stoyanchev et al. (2019), both of which resort to linguistic features, among others. In (Schmitt et al., 2011), the best achieved result was 61.6% unweighted average recall for predicting quality of interaction (i.e., the subjective label as mentioned in this paper) using a multitude of automatically extracted hand-crafted features (linguistic and dialogue state ones) and support vector machines. Our baseline system achieves a close result (59% average accuracy). Also, by using ASR and linguistic features alone in combination with support vector machines, Stoyanchev et al. (2019) manage to achieve 50% unweighted average recall. It must be mentioned, that a direct comparison is not possible due to the different nature of the features and the different categories (both papers mentioned here predicted 5 categories of interaction quality), however we can see that our system can predict dialogue success with very high performance. Another interesting observation in that context is the fact that although in our study the best results were achieved with the objective label set, in (Lykartsis et al., 2018), the better results were achieved with the subjective labels, which in our case provide the least good results - but still better than the baseline. This might be a consequence of a different definition of what constitutes subjective success between the two datasets: For the laptop dataset of (Lykartsis et al., 2018), subjective success means that the users found all the information they were looking for (when asked at the end of the dialogue), whereas for the Let's Go V2 system, subjective success meant that external raters were judging the interaction to be successful or not, probably leading to different label distributions. The different results might also be a consequence of the different tasks involved.

4 Conclusion

In this paper, we have shown that classification of user emotion, and prediction of objective and subjective task success of a spoken SDS using only audio in the form of spectrograms is not only possible, but also can be achieved to a high standard using CNNs with small computational effort, resulting in an almost real-time system. Our results greatly surpassed those of other similar studies and can be used to train models which can - on a turn level, i.e., with audio information of limited duration - predict the direction a dialogue takes and can therefore act to change the dialogue course.

We are optimistic that if our features are combined with other non-sound features (such as linguistic features), we will have the possibility to raise classification accuracy even more. However, this falls outside the aim of the current study and will be part of our future work. Furthermore, a possibility would be to perform system fusion at the classifier level, combining for example different CNN architectures (like the ones shown in this paper) and other classifiers with hand-crafted features, as in the approach from (Lykartsis et al., 2018). Such a system could benefit from the multiple different input representation and could potentially provide very good results, as in (Gu et al., 2018). As additional future work, we plan to conduct experiments with more architectures and parameters, and also employ other neural network classifiers such as Temporal Convolutional Networks (TCNs), which combine the merits of both CNNs and RNNs/LSTMs. Finally, we will also experiment with data preprocessing methods, such as denoising and data augmentation methods such as transformations in time and frequency.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283.

Bayan AbuShawar and Eric Atwell. 2016. Usefulness, localizability, humanness, and language-benefit: additional evaluation criteria for natural language dialogue systems. *International Journal of Speech Technology*, 19(2):373–383.

Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. 2017. Transfer learning for music classification and regression tasks. *arXiv preprint arXiv:1703.09179*.

Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. 2015. Towards real-time speech emotion recognition using deep neural networks. In *2015 9th international conference on signal processing and communication systems (ICSPCS)*, pages 1–5. IEEE.

Yue Gu, Shuhong Chen, and Ivan Marsic. 2018. Deep mul timodal learning for emotion recognition in spoken language. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5079–5083. IEEE.

Matthew Henderson, Blaise Thomson, and Steve Young. 2014. Word-based dialog state tracking with recurrent neural networks. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 292–299.

Seokhwan Kim, Rafael Banchs, and Haizhou Li. 2016. Exploring convolutional and recurrent neural networks in sequential labelling for dialogue topic tracking. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 963–973.

Mandy Korpusik and James Glass. 2017. Spoken language understanding for a nutrition dialogue system. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(7):1450–1461.

Mandy Korpusik and James Glass. 2018. Convolutional neural networks for dialogue state tracking without pre-trained word vectors or semantic dictionaries. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 884–891. IEEE.

Margarita Kotti, Alexandros Papangelis, and Yannis Stylianou. 2017. Will this dialogue be unsuccessful? prediction using audio features. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR)*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Honglak Lee, Peter Pham, Yan Largman, and Andrew Y Ng. 2009. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Advances in neural information processing systems*, pages 1096–1104.

Ming Che Lee, Sheng Cheng Yeh, Sheng Yu Chiu, and Jia Wei Chang. 2017. A deep convolutional neural network based virtual elderly companion agent. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 235–238. ACM.

Thomas Lidy and Alexander Schindler. 2016. Parallel convolutional neural networks for music genre and mood classification. *MIREX2016*.

Athanasios Lykartsis, M Kotti, A Papangelis, and Y Stylianou. 2018. Prediction of dialogue success with spectral and rhythm acoustic features using dnns and svms. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 838–845. IEEE.

Yoichi Matsuyama, Arjun Bhardwaj, Ran Zhao, Oscar Romeo, Sushma Akoju, and Justine Cassell. 2016. Socially-aware animated intelligent personal assistant agent. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 224–227.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science conference (SciPy)*, pages 18–25.

Michael Neumann and Ngoc Thang Vu. 2017. Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech. *arXiv preprint arXiv:1706.00612*.

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)*, pages 5010–5014. IEEE.

François Portet, Michel Vacher, Caroline Golanski, Camille Roux, and Brigitte Meillon. 2013. Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects. *Personal and Ubiquitous Computing*, 17(1):127–144.

Jan Schlüter and Sebastian Böck. 2014. Improved musical onset detection with convolutional neural networks. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6979–6983. IEEE.

Alexander Schmitt, Benjamin Schatz, and Wolfgang Minker. 2011. Modeling and predicting quality in spoken human-computer interaction. In *Proceedings of the SIGDIAL 2011 Conference*, pages 173–184. Association for Computational Linguistics.

Alexander Schmitt, Stefan Ultes, and Wolfgang Minker. 2012. A parameterized and annotated spoken dialog corpus of the cmu lets go bus information system. In *LREC*, pages 3369–3373.

Hendrik Schreiber and M Müller. 2018. A single-step approach to musical tempo estimation using a convolutional neural network. In *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR), Paris, France*.

Svetlana Stoyanchev, Soumi Maiti, and Srinivas Bangalore. 2019. Predicting interaction quality in customer service dialogs. In *Advanced Social Interaction with Agents*, pages 149–159. Springer.

Pei-Hao Su, David Vandyke, Milica Gasic, Dongho Kim, Nikola Mrksic, 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 *Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2016. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5200–5204. IEEE.

Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou. 2017. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1301–1309.

Zhong-Qiu Wang and Ivan Tashev. 2017. Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5150–5154. IEEE.

Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2016. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*.

Promod Yenigalla, Abhay Kumar, Suraj Tripathi, Chirag Singh, Sibsambhu Kar, and Jithendra Vepa. 2018. Speech emotion recognition using spectrogram & phoneme embedding. *Nineteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3688–3692.

Tiancheng Zhao and Maxine Eskenazi. 2016. Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, page 1.