# EXPLORING ATTENTION MECHANISM FOR ACOUSTIC-BASED CLASSIFICATION OF SPEECH UTTERANCES INTO SYSTEM-DIRECTED AND NON-SYSTEM-DIRECTED

*Atta Norouzian*[1], *Bogdan Mazoure*[2*], *Dermot Connolly*[1] *and Daniel Willett*[1]

[1]Nuance Communications
[2]McGill University, Canada
{atta.norouzian, dermot.connolly, daniel.willett}@nuance.com, bogdan.mazoure@mail.mcgill.ca

## ABSTRACT

Voice controlled virtual assistants (VAs) are now available in smartphones, cars, and standalone devices in homes. In most cases, the user needs to first "wake-up" the VA by saying a particular word/phrase every time he/she wants the VA to do something. Eliminating the need for saying the wake-up word for every interaction could improve the user experience. This would require the VA to have the capability of understanding whether the user is talking to it or not. In other words, the challenge is to distinguish between system-directed and non-system-directed speech utterances. In this paper, we present a number of neural network architectures for tackling this classification problem based on using only the acoustic signal. It is shown that a model comprised of convolutional, recurrent, and feed-forward layers can achieve an equal error rate (EER) of below 20% for this task. In addition, we investigate the use of an attention mechanism for helping the model to focus on the more important parts of the signal and to improve handling of variable length inputs sequences. The results show that the proposed attention mechanism significantly improves the model accuracy achieving an EER of 16.25% and 15.62% on two distinct realistic datasets.

***Index Terms*—** Human-machine interaction, spoken utterance classification, wake-up word, attention mechanism

## 1. INTRODUCTION

Thanks to recent advances in speech recognition and natural language understanding, VAs have become part of our daily lives. The VAs are typically activated by a wake-up word/phrase such as *hi Mercedes*, *hey BMW*, *hey Siri*, *Alexa* or *ok Google*. Eliminating such wake-up words in favor of allowing direct requests for assistance from the VA could significantly improve the user experience. This requires the device to have the capability to detect speech directed at it and ignore human-to-human and background speech. The problem of classifying spoken utterances into system-directed and non-system-directed has previously been investigated within the context of virtual assistants [1, 2, 3] and dialogue systems [4, 5].

Both, the spoken words and the way they are spoken provide cues for differentiating between system-directed and non-system-directed speech utterances. Typically, the lexical cues are extracted from a word sequence generated by an automatic speech recognition (ASR) system. One approach is based on computing likelihoods for each class using class-specific language models and then use the likelihood ratio to predict the class [6, 7]. Alternatively, the word sequence can be input to a neural network (NN) model to either directly estimate class probabilities [8] or to generate new features for another model [1]. The non-lexical acoustic cues are often represented as a sequence of feature vectors corresponding to prosodic structure of the speech [2, 6] or the short-time frequency characteristic of the signal [1, 3]. Such feature vectors are typically extracted from a window of 20-25 ms. By sliding this window by 10 ms at a time a sequence of feature vectors is generated. This means the speech utterances of different length will have feature representations of different dimensions. This presents a challenge for training models that take the whole feature sequence as input and output one prediction per utterance and not per frame. The model designed for such tasks needs to be able to deal with inputs of variables lengths. Averaging the features over time [3] or passing the input sequence into a long short-time memory (LSTM) cell and using the last output of the LSTM cell [1] are examples of how others have dealt with this issue. In this paper, we propose a new technique based on attention to address some shortcomings of the other methods.

Similar to systems developed in [1, 2, 3, 6, 9] our plan is to combine information extracted from acoustic features with lexical information for this classification task, however, the focus of this paper is only on acoustic-based classification. The acoustic features explored for this purpose are frame-based log Mel-filterbank coefficients. We favor short-term frame-based acoustic features for this task since they facilitate early detection of user's intent by gradual application of the trained model on the incoming speech. The proposed classification models are based on deep neural networks (DNN) with combination of convolutional, recurrent and feed-forward layers. In addition, the use of an attention mechanism on top of the convolutional layer as well as the recurrent layer is investigated.

This paper is organized as follows. First, an overview of the models developed for this classification problem is presented in 2. A number of model architectures proposed for frame-based approach and the architecture of the utterance-based model are described in Section 3. The experimental study containing a description of the evaluation data, the model parameters, and the experimental results is provided in Section 4. Summary and conclusion are given in Section 5.

## 2. OVERVIEW

Here, an overview of the two modeling approaches investigated for the system-directed versus non-system-directed classification problem is given. The two approaches are based on using frame-level and utterance-level input features. The models developed based on the frame-level input features are depicted in Figure 1 on the left and the utterance-based model in shown on the right. Several architectures

[*] The author performed the work while an intern at Nuance

978-1-5386-4658-8/18/$31.00 ©2019 IEEE

are explored for the frame-based models as described in Section 3 for dealing with the variable length input sequences. The model developed for the fixed-length utterance-level features is comprised of only dense feed-forward layers.

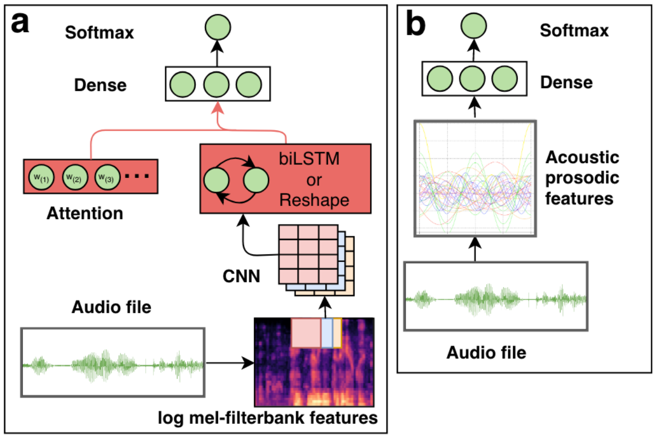

**Fig. 1**. General architecture of frame-based models **(a)** and the utterance-based model **(b)** developed for the classification task at hand.

## 3. MODEL ARCHITECTURES

This section presents a number of model architectures for dealing with the issue of variable-length input feature representation faced in the frame-based approach. Moreover, it describes the input features and the architecture of the utterance-based model in detail.

### 3.1. Frame-based Approach

In this approach the feature vectors input to the models consist of 45 log Mel-filterbank coefficients extracted from 25 ms of acoustic signal with a frame shift of 10 ms. A speech utterance is hence represented by a sequence of feature vectors, $\{\boldsymbol{m}_1^{45}, \ldots, \boldsymbol{m}_T^{45}\}$, where $T$ is the total number of frames in the utterance. All frame-based models developed here use a two-dimensional convolutional layer as input layer which outputs a set of $d$ feature maps denoted by $\{\boldsymbol{E}_1^{j\times l}, \ldots, \boldsymbol{E}_d^{j\times l}\}$. The width of the feature maps, $j$, is proportional to the acoustic feature vector size (i.e., 45) and their length $l$ is proportional to $T$. In a realistic scenario, recorded utterances have different lengths which means $T$ and consequently $l$ vary from one utterance to another. This causes an issue when converting the feature-maps into a vector to pass to feed-forward layers since the input to a feed-forward layer has to have a fixed-dimension for all samples. In the following, three approaches for creating a fixed-length vector from variable-length feature-maps are presented. After creating a fixed-length vector it is input to dense feed-forward layers followed by a softmax layer as shown in Figure 1.

**Global averaging across time:** A simple way of generating a fixed-length representation is to take the average of each feature-map over its length $l$. This will transform every 2-D feature map $\boldsymbol{E}_i^{j\times l}$ to a vector $\boldsymbol{e}_i^j$. The resulting vectors, $\{\boldsymbol{e}_1^j, \ldots, \boldsymbol{e}_d^j\}$, are then concatenated and fed to a feed-forward layer as was done in [10].

**Using a recurrent layer:** One could obtain a fixed-length vector from variable-length feature maps by using a recurrent neural layer. This is done by first concatenating columns of all feature maps to generate $l$ super vectors of dimension $d \times j$. Next, the super vectors are fed to a recurrent layer one by one and the last (i.e., $l$th) output of the recurrent layer is used for the succeeding layer. In addition, one could use a bi-directional recurrent layer and use the last output vector of forward and backward directions to obtain a richer fixed-length representation. In the model explored here, a bi-directional LSTM layer is used for this purpose and the two resulting vectors from both directions are concatenated and used in the feed-forward layer.

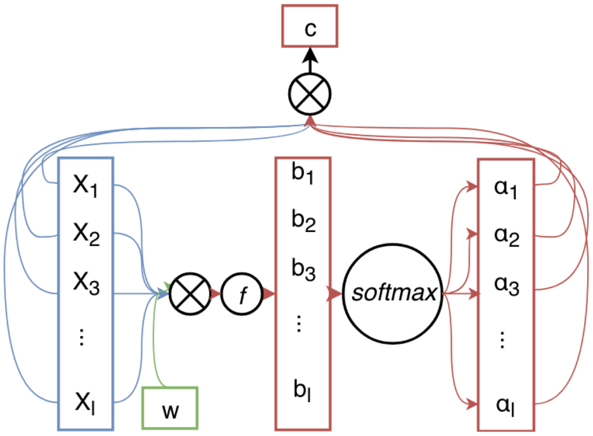

**Fig. 2**. Attention mechanism applied to the input sequence $\boldsymbol{X}$ to generate the context vector $\boldsymbol{c}$.

**Using attention mechanism:** Simple averaging of feature maps or passing them through a recurrent layer and using only its last output could result in losing important information. An attention mechanism could retain most of the relevant information while resolving the variable-length issue. The attention mechanism explored here is somewhat different from the traditional encode-decoder based attention introduced in [11]. It is in essence a weighted average of sequence of vectors where the weights are learned through back-propagation. This mechanism was first explored for emotion recognition in [12] and is similar to the idea of self-attention in [13]. Denoting a sequence of $l$ vectors of dimension $s$ by the matrix $\boldsymbol{X}^{s\times l}$, attention is computed as

$$\begin{aligned} \boldsymbol{b}^{l\times 1} &= f(\boldsymbol{w}^{1\times s}\boldsymbol{X}^{s\times l}), \\ \alpha^i &= \frac{\exp(b_i)}{\sum_{j=1}^{l}\exp(b_j)}, \quad i = 1, \ldots, l, \end{aligned} \tag{1}$$

where $\boldsymbol{w}$ is the weight vector learned through back-propagation, $f$ is a non-linear function (here $tanh$), $b$ is the attention vector, and $\alpha$ is the normalized attention vector. Applying attention to the input sequence results in a vector known as context vector given by

$$\boldsymbol{c}^{s\times 1} = \boldsymbol{X}^{s\times l}\boldsymbol{\alpha}^{l\times 1}. \tag{2}$$

As can be seen in Equation 2, the context vector dimension is independent of the length of the input sequence $l$. Furthermore, the attention vector $\boldsymbol{\alpha}$ helps to put more emphasis on the parts of the input sequence $\boldsymbol{X}$ that carry the most relevant information for distinguishing the two classes. The process of computing the attention and applying it to the input sequence is shown in Figure 2. The input sequence in this case could be the flattened feature maps or the output of the recurrent layer.

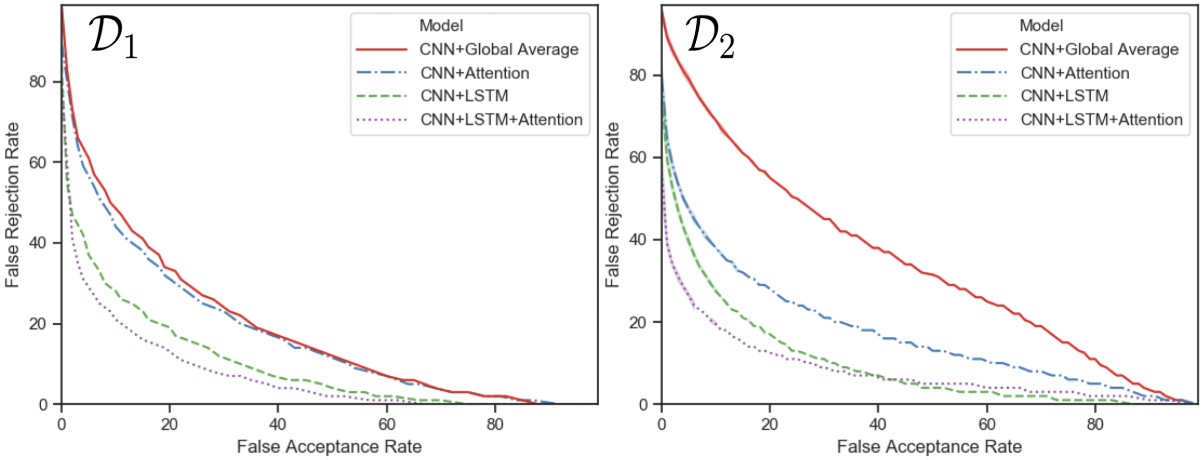

**Fig. 3**. Detection error trade-off curves of the four frame-based models on dataset $\mathcal{D}_1$ (left) and dataset $\mathcal{D}_2$ (right).

### 3.2. Utterance-based Approach

As an alternative to the frame-based approach one could represent every utterance with a fixed-length feature representation prior to any modeling. This can be done by computing some functions over the frame-based features. The feature set used here was developed for INTERSPEECH ComParE emotion recognition sub-challenge [14]. It contains 6373 acoustic-features described in [15]. We used the openSmile toolkit [16] for extracting these features from the speech utterances in our corpus. Although, this feature set was originally developed for an emotion recognition task, it contains a variety of acoustic-prosodic features (F0, energy, zero crossing, mfccs) many of which are relevant for this classification problem as well. A three-layer dense feed-forward model is trained and evaluated for these features.

## 4. EXPERIMENTAL STUDY

This section provides a description of the two data sets used for training and evaluation of the classifier. Afterwards, the model parameter are given and finally experimental results are presented and analyzed.

### 4.1. Datasets

Two datasets are used for validation of the proposed techniques. Both contain recordings of single-user interactions with a virtual assistant. The language of all recordings is English and many of the user's are not native English speakers. The first dataset denoted by $\mathcal{D}_1$ contains roughly 105 thousand speech utterances from 74 thousand speakers. Every uttarence is categorized as system-directed or non-system-directed based on the user's intent as predicted by an NLU model. The system-directed category covers all questions and commands while all dictation and background speech are assigned to the non-system-directed category. The second dataset denoted by $\mathcal{D}_2$ is constructed in a similar fashion to $\mathcal{D}_1$ but it also includes open microphone recordings and recordings of non-speech noise in the non-system-directed category. The models are trained only on the training subset of $\mathcal{D}_1$ and the dataset $\mathcal{D}_2$ was used just for testing. Table 1 shows the breakdown of both datasets by class and training/validation/test. The training subset is balanced in terms of the number of samples from the two classes to prevent any bias towards a specific class.

| | $\mathcal{D}_1$ | | | $\mathcal{D}_2$ |
|---|---|---|---|---|
| | Training | Validation | Test | Test |
| System | 35k | 12k | 14k | 7k |
| Non-system | 35k | 5k | 4k | 127k |

**Table 1**. Number of utterances per training, validation and test splits across both classes and datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ rounded to the nearest thousand.

### 4.2. Model Parameter

Prior to training the models, the frame-based and utterance-based feature vectors are normalized to have zero mean and unit standard deviation along each dimension to facilitate model convergence. Moreover, Adam optimizer and early stopping are used for training the models. For the frame-based models, the convolutional layer has a depth of 50 with a kernel height of 20 and width of 9 for the models without the LSTM layer and width of 5 for the models with LSTM layer. A stride of 5 is used along the time access and 3 along the log Mel-filterbank coefficients. The LSTM layer is bi-directional with 128 units in each direction. The three feed-forward layers in frame-based models each have 128 units. The best classification accuracy was obtained from the utterance-based model when three feed-forward layers of 128 units were used. All the models were trained using tensorflow toolkit [17].

### 4.3. Results

In this section the classification models described in Section 2 are evaluated on $\mathcal{D}_1$ and $\mathcal{D}_2$ datasets defined in Section 4.1. Figure 3 shows the performance of the models in terms of detection error tradeoff (DET) curves. A number of observations can be made from these plots. First, using an LSTM significantly improves the model performance compared to global averaging. Moreover, adding attention to the mix yields an additional boost to the performance with and without the LSTM. To have a single point of reference to compare the models, the equal error rate metric which corresponds to equal Type I and Type II errors is measured and shown in Table 2. The main question here is what the model is actually learning. This is not easy to answer especially when it comes to neural network models. However, the attention mechanism could help shedding some light on this matter. Aligning the attention vector $\boldsymbol{\alpha}$ with the original speech utterance, one could find out where the model is putting the most emphasis. This is done in Figure 4 for two utterances from the two classes. The word sequences associated with the utterances are also shown in the figure to identify possible correlations between the spoken words and where the model is mostly focusing on. The vertical lines correspond to start and end time of the words. The plot shows that for the system-directed utterance the attention is on both "*play*" and "*music*" while for the non-system-directed example the attention is mostly on the words "*that*", "*fine*", and "*period*". It is interesting to note that in the training dataset the word "*period*" is spoken only when the users are dictating a phrase. In other words, this word is a strong indication that the speech utterance is of dictation style which belongs to non-system-directed class. This led us to think that maybe the model is just learning keywords and is not learning any para-linguistic information. To answer this question we looked at a number of system-directed utterances such as "*you didn't catch that*" and "*one more run after that*" that were not part of the training data and did not contain any word highly correlated with system-directed class. The model classified both of these utterances correctly with high confidence. This suggests that the model is not just learning keywords or para-linguistic information but rather a combination of both. Adding more training data from different domains would make the model less sensitive to words and more sensitive to para-linguistic information. In Figure 5, the proposed frame-based approach is compared to the utterance-based approach described in Section 3.2 on the $\mathcal{D}_1$ dataset. It should be noted that the utterance-based acoustic-prosodic feature set was designed for emotion recognition and contains several features that may not be relevant for this task. Nevertheless, the gap between the two curves indicates that even without using hand-crafted features and only relying on frame-based log Mel-filterbank features very good classification performance can be achieved with the proposed attention-based modeling technique.

| | $\mathcal{D}_1$ | $\mathcal{D}_2$ |
|---|---|---|
| CNN+Global Average | 26.99 | 39.25 |
| CNN+Attention | 26.21 | 24.90 |
| CNN+LSTM | 19.46 | 18.79 |
| CNN+LSTM+Attention | **16.25** | **15.62** |

**Table 2**. Equal error rates of the four frame-based models measured on $\mathcal{D}_1$ and $\mathcal{D}_2$ test sets.

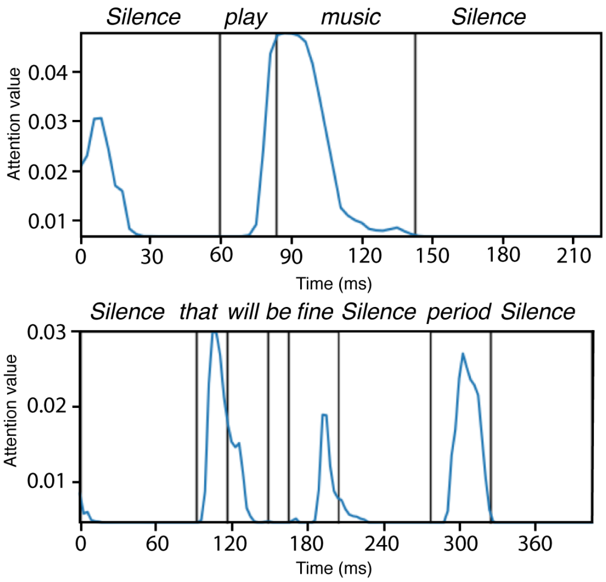

**Fig. 4**. Attention vector spread across time for a system-directed utterance (top) and a non-system-directed utterance (bottom) from $\mathcal{D}_1$ testset.

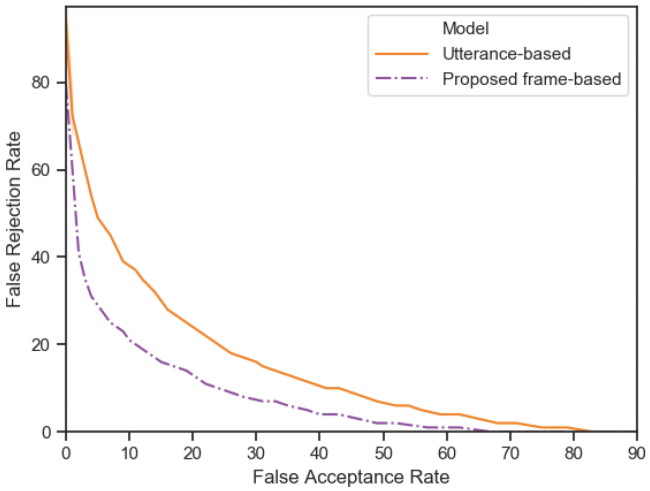

**Fig. 5**. Detection error trade-off curves for the utterance-based approach and the proposed frame-based approach measured on $\mathcal{D}_1$ testset.

## 5. CONCLUSION

In this paper, the problem of classifying speech utterances into system-directed and non-system-directed was addressed. A number of neural network architectures were explored for this task. It was shown that having an attention mechanism improves the classification performance whether applied directly to the output of the convolutional layer or to the output of the recurrent layer. The best performing model was built by stacking a convolutional layer, a recurrent layer and three feed-forward layers with attention applied to the output of the recurrent layer. This model achieved an EER rate of 16.25% on one test set and 15.62% on the second test set. As continuation of this work we are looking into combining direct audio classification with ASR-output based text classification for improved accuracy.

## 6. ACKNOWLEDGEMENTS

The authors would like to thank two colleagues from Nuance Communications, Raymond Brueckner for helping out with the openSmile toolkit and Yasser Hifny for very insightful discussions.

## 7. REFERENCES

[1] Sri Harish Mallidi, Roland Maas, Kyle Goehner, Ariya Rastrow, Spyros Matsoukas, and Björn Hoffmeister, "Device-directed utterance detection," in *INTERSPEECH*, 2018, pp. 1225–1228.

[2] Daniel Reich, Felix Putze, Dominic Heger, Joris Ijsselmuiden, Rainer Stiefelhagen, and Tanja Schultz, "A real-time speech command detector for a smart control room," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[3] Tomoyuki Yamagata, Tetsuya Takiguchi, and Yasuo Ariki, "System request detection in human conversation based on multi-resolution gabor wavelet features," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[4] Dong Wang, Dilek Hakkani-Tür, and Gokhan Tur, "Understanding computer-directed utterances in multi-user dialog systems," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8377–8381.

[5] John Dowding, Richard Alena, William J Clancey, Maarten Sierhuis, and Jeffrey Graham, "Are you talking to me? dialogue systems supporting mixed teams of humans and robots," in *AAAI Fall Symposium on Aurally Informed Performance: Integrating Machine Listening and Auditory Presentation in Robotic Systems, Arlington, Virginia*, 2006.

[6] Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, and Larry Heck, "Learning when to listen: Detecting system-addressed speech in human-human-computer dialog," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[7] Suman V Ravuri and Andreas Stolcke, "Neural network models for lexical addressee detection," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[8] Suman Ravuri and Andreas Stolcke, "Recurrent neural network and lstm models for lexical utterance classification," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[9] Elizabeth Shriberg, Andreas Stolcke, and Suman V Ravuri, "Addressee detection for dialog systems using temporal and spectral dimensions of speaking style.," in *INTERSPEECH*, 2013, pp. 2559–2563.

[10] Suwon Shon, Ahmed Ali, and James R. Glass, "Convolutional neural networks and language embeddings for end-to-end dialect recognition," *CoRR*, vol. abs/1803.04567, 2018.

[11] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[12] Michael Neumann and Ngoc Thang Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," in *INTERSPEECH*, 2017.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[14] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, et al., "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[15] Felix Weninger, Florian Eyben, Björn W Schuller, Marcello Mortillaro, and Klaus R Scherer, "On the acoustics of emotion in audio: what speech, music, and sound have in common," *Frontiers in psychology*, vol. 4, pp. 292, 2013.

[16] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[17] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, Software available from tensorflow.org.