A domain-agnostic approach for opinion prediction on speech

Pedro Bispo Santos[†] and **Lisa Beinborn**[†] and **Iryna Gurevych**[†‡]
[†]Ubiquitous Knowledge Processing Lab (UKP)
Department of Computer Science, Technische Universität Darmstadt
[‡]Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research
https://www.ukp.tu-darmstadt.de/

Abstract

We explore a domain-agnostic approach for analyzing speech with the goal of opinion prediction. We represent the speech signal by mel-frequency cepstral coefficients and apply long short-term memory neural networks to automatically learn temporal regularities in speech. In contrast to previous work, our approach does not require complex feature engineering and works without textual transcripts. As a consequence, it can easily be applied on various speech analysis tasks for different languages and the results show that it can nevertheless be competitive to the state-of-the-art in opinion prediction. In a detailed error analysis for opinion mining we find that our approach performs well in identifying speaker-specific characteristics, but should be combined with additional information if subtle differences in the linguistic content need to be identified.

1 Introduction

Traditional natural language processing approaches have focused on the analysis of linguistic content and the represented information. With the increasing availability of recorded speech, the interest shifted from pure content processing to analyzing the states and traits of speakers (Schuller et al., 2012). For this purpose, paralinguistic features such as pitch and loudness of voice are playing an important role because they are very predictive social markers (Laver and Trudgill, 1979). They influence our persuasiveness (Burgoon et al., 1990), indicate our emotional state (Scherer, 2003) and correlate with our personality traits (Markel et al., 1972).

The ability to analyze paralinguistic features has led to progress in a multitude of speech processing tasks such as age identification (Metze et al., 2007), personality recognition (Schuller et al., 2012) and emotion recognition (Nwe et al., 2003). A subset of these problems is tackled every year as shared tasks in the *Computational Paralinguistics Challenge* at the INTERSPEECH conference (Schuller et al., 2015; Schuller et al., 2014).[1] For the winning methods of the last editions from these shared tasks, thorough task-specific feature engineering has usually been the key point.

In this paper, we aim at reducing the engineering effort and the dependence on domain-specific knowledge in speech processing tasks for opinion prediction. We approach this goal by applying deep learning methods which have been shown to automatically learn more complex and high-level features from basic features extracted from the signal (Palaz et al., 2015). The main challenge for applying these approaches lies in determining a good representation of the data and choosing a suitable architecture for the task at hand.

For our approach, we use only the speech signal as input, so that expensive textual transcripts are not required. We work on the frame level[2] and choose mel-frequency cepstral coefficients (MFCCs) as our unit of representation because they correspond well to the human auditory system and are very discriminative for speech processing tasks, such as phoneme recognition (Davis and Mermelstein, 1980), speaker identification (Ren et al., 2016) and claim identification in political debates (Lippi and Torroni,

[1] http://emotion-research.net/sigs/speech-sig/is16-compare

[2] Frames are overlapping windows from the signal obtained from short-term analysis.

Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pages 163–172, Osaka, Japan, December 12 2016.

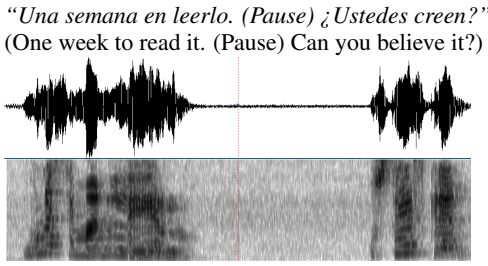

Figure 1: Subject expressing her negative opinion about a book. The dataset contains the textual transcripts and the recorded utterances from the subjects. Here we can visualize the raw signal of her utterance along with the corresponding spectrogram.

2016). Rosen (1992) analyzes that speech perception is strongly influenced by temporal dependencies. We therefore model the speech signal as a time series and use long short-term neural networks as machine learning method. In contrast to previous approaches in computational paralinguistics, we do not need to compute additional task-specific statistics on the features extracted from the frames because LSTM networks are able to learn the temporal regularities automatically from the input signal. This makes it possible to apply our approach to different tasks without additional engineering.

In order to test whether our approach can compete with state-of-the-art methods, we focus on two interesting tasks concerning speech: opinion mining and persuasiveness prediction. For both tasks, the goal can be framed as opinion prediction, but the perspective differs. In the first task, our goal is to predict the opinion of a user speaking about a product. In the second task, we aim at predicting the influence of a speaker on the opinion of an audience. Previous approaches to these tasks developed a sophisticated feature set to capture the recognition of emotions for opinion mining (Poria et al., 2015) and the characteristics of voice quality for persuasiveness prediction (Brilman and Scherer, 2015).

We find that the results of our domain-agnostic approach come close to the performance of domain-specific ones that apply thorough feature engineering. As we use the same features for different tasks, we minimize the risk of overfitting to the data. Our error analysis explain in more details the issues with our approach in both datasets, but also highlight how far a generic computational method based solely on speech can go in tasks related to opinion prediction.

2 Tasks

For the evaluation of our approach, we focus on two different speech tasks: opinion mining and persuasion prediction. In both tasks, the goal is to analyze opinions. For opinion mining, we aim at directly predicting the opinion of the speaker and for persuasiveness prediction we aim at indirectly predicting the opinion of an audience based on the persuasiveness of the speaker.

2.1 Opinion mining

In opinion mining, the task is to assign a polarity (*negative*, *neutral*, *positive*) to opinions expressed by users. This task has become increasingly popular with the rise of social platforms which provide valuable information on customers' opinions. As manual analyses cannot scale up to the vast amount of opinionated comments, the application of automatic analyses is required. For our experiments on opinion mining, we use the MOUD dataset.

MOUD Dataset The *Multimodal Opinion Utterance Dataset* (Pérez-Rosas et al., 2013b) is a collection of video blogs extracted from *YouTube*.[3] It consists of videos from 80 Spanish native speakers (15 male, 65 female) who express their opinion about movies, books and cosmetics. Figure 1 shows an example of a review and the corresponding speech signal from the utterance. The speakers' age ranges from 20 to 60 years. Pérez-Rosas et al. (2013b) manually extracted a 30 seconds opinion snippet from each video and segmented it into utterances yielding a total of 498 utterances. Each utterance was then analyzed by two

[3] `https://www.youtube.com/`

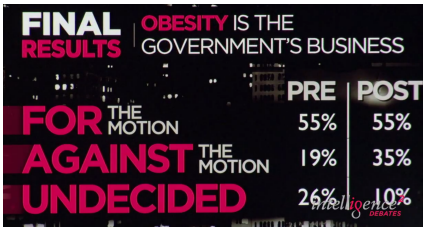

Figure 2: An example poll from the debate dataset. The debate winner is the team which sways more votes; in this case the team that argued *against* the motion.

annotators to determine whether the speaker reveals a *positive*, *neutral* or *negative* sentiment towards the product. They report an inter-annotator agreement of 0.88 and a kappa of 0.81. Conflicting annotations were subsequently resolved by discussions. We use the publicly available dataset and exclude utterances with a neutral label from our experiments to be consistent with previous work (Pérez-Rosas et al., 2013b; Poria et al., 2015).[4]

2.2 Persuasiveness Prediction

The task of persuasiveness prediction in debates has been established by Brilman and Scherer (2015) who worked with videos of debates from the *Intelligence Squared* organization. In these debates, two teams argue about a motion and try to convince the audience of their stance. The team that is able to sway more votes from the audience wins the debate. The goal is to predict the persuasiveness of the teams and the individual debaters.

Intelligence Squared Dataset *Intelligence Squared* is an organization which promotes debates about controversial motions between topic experts. The debates are all recorded and available online.[5] Each debate team is composed of two debaters and the debates are split into three rounds: opening statements, question round and closing statements. The debates are performed in Oxford-style which means that the opinion of the audience is measured by two polls. The first poll is conducted before the start of the debate, and the second one after the closing statements. The audience can vote *for* or *against* the motion or choose to remain *undecided*. In Figure 2, we see an example for a motion stating that *obesity is the government's business*. In this case, the team *against* the motion won because they achieved a higher relative gain of votes (16%). It should be noted that the team *for* the motion represents the opinion of the majority here, but could not convince the remaining audience to change their opinion during the debate.

We implemented a crawler to obtain the debates from the organization's website. For our experiments, we used the same setup as Brilman and Scherer (2015). This means that debates which had a voting difference equal to or smaller than six are excluded and the prediction is only based on the opening and closing statements of each debate. This procedure yields 30 debates in total and includes 120 debaters (19 female, 101 male). We publish the code for the crawler and the list of seed urls.[6]

3 Related Work

The task of opinion mining is quite established in natural language processing, but most approaches have been developed for textual data (Pang and Lee, 2008). In this work, we focus on opinion mining in speech. Persuasiveness prediction is a relatively new task in the area of debating technologies.

3.1 Opinion Mining

Scherer (2003) shows that paralinguistic features are particularly informative for identifying the speakers' emotional state, and they have been used extensively for the task of detecting principal emotions such as fear or anger (Batliner et al., 2011). However, the subtler task of analyzing the opinion of a speaker

[4] http://web.eecs.umich.edu/~mihalcea/downloads.html#MOUD
[5] http://intelligencesquaredus.org/
[6] https://github.com/UKPLab/coling-peoples2016-opinion-prediction

towards a product has not yet received much attention. Mairesse et al. (2012) compare models built on textual features with models built on paralinguistic features to predict the opinion expressed in short spoken reviews. They found that the results improve if the features calculated on transcripts are combined with paralinguistic features. Morency et al. (2011) examine three modalities and extract visual, audio and textual features to predict the opinion expressed in videos. They find that combining the three modalities produces the best outcome. The approach was then extended to other languages (Pérez-Rosas et al., 2013a) and to more fine-grained analyses on the utterance level leading to the *MOUD* dataset (Pérez-Rosas et al., 2013b). Poria et al. (2015) improve the results for the *MOUD* dataset by applying a deep learning approach that builds a representation for the transcripts with convolutional neural networks. Both approaches use a wide range of thoroughly engineered features including acoustic-prosodic features like pitch and speaking rate for emotion recognition, textual features for the detection of sentiment words, and visual features such as facial landmarks for capturing emotional states.

To account for the importance of temporal aspects for speech perception (Rosen, 1992), we model the speech signal as a time series. In previous work on opinion mining in speech, complex functions had been calculated over the features extracted at the frame level to account for the temporal dependencies. Recent progress in modeling time series data has been achieved with long short term memory networks. They have obtained good results for audio processing tasks such as music composition (Coca et al., 2013) and phoneme classification (Graves and Schmidhuber, 2005). They have also been applied in opinion mining on text (Wang et al., 2016), but have not yet been explored for opinion mining on speech.

3.2 Debating Technologies

The field of debating technologies is a newly developing research area that focuses on computational methods to support human argumentation and debating (Gurevych et al., 2016). In recent work, claim identification for controversial topics (Roitman et al., 2016), evidence detection (Rinott et al., 2015) and argument convincingness prediction (Habernal and Gurevych, 2016) have been tackled.

These works focus on analyzing the content, but Hosman et al. (2002) showed that paralinguistic features are very informative to detect credibility and persuasiveness of speakers. This observation has been used in the work by Lippi and Torroni (2016) who combine paralinguistic features with textual features to detect claims in political debates. They represent the input signal by mel-frequency cepstral coefficients and find that the combination of text and audio modalities yields the best results. Brilman and Scherer (2015) also apply a multi-modal approach and combine textual, acoustic and visual information to predict the persuasiveness of speakers in the *Intelligence Squared* dataset. They represent the speech data by features related to voice perception such as pitch, formants and voice quality. Park et al. (2014) did a very similar approach to Brilman and Scherer (2015), although not working with data from debates, but with movie reviewers from *ExpoTV*.[7] They used even more features: MFCCs, pitch, formants and all the voice quality features used by Brilman and Scherer (2015). All approaches extract speech features on the frame level, calculate statistics such as average and standard deviation over the sequential data and feed them to support vector machines. Unfortunately, statistical functions computed over static representations of frames cannot capture temporal dependencies in the speech sequence. Chung et al. (2016) have shown that LSTMs can overcome this issue and model the speech signal more adequately.

4 Methodology

Our domain-agnostic approach is based on two aspects: a simple but informative paralinguistic feature set which can be easily extracted for speech signals from different domains and a deep learning approach which can discover temporal regularities in the data.

4.1 Features

Creating textual transcripts of speech recordings is an expensive and time-consuming task. It requires either thorough manual work or a sophisticated acoustic model trained on large corpora for automatic

[7] `http://www.expotv.com`

speech recognition (Xiong et al., 2016). In contrast to previous work, we rely only on the basic speech signal in order to evaluate whether satisfactory prediction quality can be reached even without transcripts.

Since Hosman et al. (2002) find that powerful speeches are more persuasive and Pérez-Rosas et al. (2013b) analyze that the energy level of the voice is predictive for opinion mining, we aim at representing the speech signal by paralinguistic features from the power spectrum. Our auditory system is very sensitive to changes in the frequency of an acoustic wave when the frequency is low, but more robust to changes in higher frequency ranges. The mel-scale is a scale which corresponds to our perception on frequency changes (Stevens et al., 1937). We use mel-frequency cepstral coefficients (MFCCs) from 13 different frequency ranges as our representation unit because they are a good approximation of the human auditory perception (Davis and Mermelstein, 1980). The MFCCs are obtained by dividing the speech signal into frames and applying a discrete fourier transform. Based on a filter-bank analysis with mel-scaled frequency bins, the cepstral coefficients can then be determined with a cosine discrete transform. Using only one basic operationalization for speech that can be calculated automatically, it keeps our feature extraction effort small and allows us to apply our approach to different domains. These coefficients are usually interpreted as a good generic indicator for different tasks in speech processing, such as speaker identification (Ren et al., 2016) and claim identification in debates (Lippi and Torroni, 2016).

4.2 Learning Architecture

Deep learning architectures have the power to learn high-level abstractions from raw features and are strongly used in vision, language and speech (Bengio, 2009). To account for the sequential nature of speech signals, we apply an LSTM architecture which has been developed for processing time series (Hochreiter and Schmidhuber, 1997). LSTM networks are based on recurrent neural networks and use memory cells to keep track of long-term dependencies by the usage of gate units. The network directly processes the extracted features from each frame and automatically learns high-level abstractions. Using this architecture, we avoid the effort of manually defining task-specific statistics over the frame level features which has usually been necessary for speech labeling tasks.

4.3 Experimental Setup

The MFCCs were extracted using the python library *python_speech_features*.[8] The window size was 25 ms with a sliding window of 10 ms. The *Keras* framework[9] was used for implementing the LSTMs. The code from both experiments is available on GitHub.[10]

Opinion Mining The audio files from this dataset have a sampling rate of 44,100 Hz. We have implemented a bi-directional LSTM with 128 nodes at each hidden layer. The batch size is 128 and the dataset is divided into 10 folds in order to perform cross-validation. Each utterance is preprocessed, and sequences with a length greater than 236 were truncated. Adam is used as optimizer and binary cross-entropy is used as loss function. We use hyperbolic tangent as activation function for all hidden layers and for the merging layer. The last fully connected layer which assigns the binary label to the sequence uses sigmoid as activation function. All hyperparameters were set based on empirical evidence obtained from experiments on a single fold.

Persuasion Prediction We extracted the speech signal for each debater with *FFmpeg*.[11] The audio segments have a sampling rate of 48,000 Hz. In contrast to the input sequences from the *MOUD* dataset which were split into utterances and lasted only a few seconds, the segments in the *Intelligence Squared* dataset last a few minutes resulting in up to 25,000 frames. We apply padding to the shorter sequences.

We implemented an LSTM network with hidden layers containing 64 nodes in the *Keras* framework. We use hyperbolic tangent as activation function and a dropout of 0.2 for both the matrix and the recurrent weights. The last layer is a fully connected layer with a single node and a sigmoid activation function which assigns the label to the sequence. The label indicates whether the debater belongs to the winning or

[8] http://python-speech-features.readthedocs.io/en/latest/
[9] https://keras.io/
[10] https://github.com/UKPLab/coling-peoples2016-opinion-prediction
[11] https://ffmpeg.org/

the losing team. We use binary cross-entropy as loss function, RMSProp as optimizer, and a batch size of 1. The hyperparameters were set based on empirical evidence from experiments on a single fold. Like Brilman and Scherer (2015), we perform a leave-one-debate-out cross-validation to avoid a topic-specific bias. The data is split into 30 different folds, each using 29 debates for training and the remaining debate for testing.

5 Results

We evaluate our domain-agnostic approach on two tasks with different languages and compare the results to the state-of-the-art in each task.

Opinion Mining For opinion mining, we compare our approach to a majority baseline and to the results obtained by the speech features from the domain-specific approaches by Pérez-Rosas et al. (2013b) and Poria et al. (2015) in Table 1. It can be seen that our approach outperforms the majority baseline and the method by Pérez-Rosas et al. (2013b). As expected, the approach by Poria et al. (2015) which uses carefully engineered features for emotion recognition performs better on the task. It should be noted that the results of our approach even get close to the results obtained by content-specific textual features calculated over the transcripts, where the textual features are only 4.1% better than our approach. This shows that a generic speech feature set processed by a bi-directional LSTM can approximate the results of domain-specific approaches for opinion mining without further engineering.

System	Modality	Accuracy
Majority baseline	-	.559
Our approach	Audio	.668
Pérez-Rosas et al. (2013b)	Audio	.648
Poria et al. (2015)	Audio	**.742**
Pérez-Rosas et al. (2013b)	Text	.709
Poria et al. (2015)	Text	**.797**

Table 1: Accuracy results for opinion mining

Persuasion Prediction For persuasion prediction, we use the same evaluation setup as Brilman and Scherer (2015). They evaluate the accuracy for the opening and closing statements separately and distinguish between the accuracy on the individual level and on the debate level. The classifier predicts for each debater individually whether she belongs to the winning or the losing debate team. This can lead to a tied prediction for a team as each team consists of two debaters. To account for this, the debate-level accuracy measure combines the two individual labels by computing an accuracy of 1 if both individual labels match the team label, 0 for a complete mismatch and 0.5 for a tied prediction. Both accuracy measures – individual and debate level – are averaged over all folds. As the dataset is balanced for winning and losing teams, the majority baseline obtains an accuracy of 0.5.

Level	System	Opening	Closing	Modality
Individual	Majority baseline	.500	.500	-
	Our approach	**.683**	.642	Audio
	Brilman and Scherer (2015)	.675	**.650**	Audio
	Brilman and Scherer (2015)	.550	.600	Text
Debate	Majority baseline	.500	.500	-
	Our approach	**.767**	.683	Audio
	Brilman and Scherer (2015)	.717	**.733**	Audio
	Brilman and Scherer (2015)	.533	.700	Text

Table 2: Accuracy results for persuasion prediction at the individual level and the debate level.

The results in Table 2 show that our approach outperforms the majority baseline by at least 14.2% for each setting and performs on par with the results obtained for the speech features by Brilman and

Scherer (2015) (slightly better for the opening statements and slightly worse for the closing statements). It is particularly interesting to note that the results for the speech features are even stronger than the results obtained by content-specific textual features. This indicates that voice quality aspects have a strong influence on the persuasiveness of a speaker independent of the actual content of his arguments.

For our experiments, we only operated on the speech level to evaluate the predictive power in the absence of textual transcripts. Obviously, better results can be obtained by combining information from multiple modalities and by using domain-specific features. Nevertheless, the results show that our approach can provide a competitive start when switching to new domains.

6 Error Analysis

In order to better identify the strength and weaknesses of our naïve approach for opinion prediction, we perform a more detailed analysis of the results.

Opinion Mining After a first round of qualitative analyses, we noticed that many speakers express mixed opinions towards a product as in the following example: *The thing is: when you use it, it may hurt your eye a little bit, (**negative**) so after using it for the first time, I thought: "Oh no, I am not going to use it anymore, that is not possible!" (**negative**)[...] However, it is super easy to be washed.(**positive**).*

In the *MOUD* dataset, this opinion is segmented into three utterances with the polarity labels indicated in brackets. We noticed that from the subjective perception only minor changes in the voice could be observed for these three utterances because the speaker kept a rather neutral tone. As the dataset contained many similar examples, we were puzzled by the fact that the classifier was still able to predict the correct opinion label for the majority of utterances based on the voice features alone and started a deeper investigation.

We observe that most speakers have a tendency towards expressing either mostly positive or mostly negative utterances. In the current evaluation setup established in previous work, utterances by the same speaker are distributed over the training and test set which might lead to a speaker bias. A speaker-majority classifier, i.e. a classifier which learns to assign the majority label for a particular speaker to all her utterances, would obtain 87.7% of accuracy for this dataset and strongly outperform all results in Table 1. This indicates that the underlying task of this dataset is not necessarily opinion mining, but rather speaker identification which explains the acceptable performance of our domain-agnostic approach. [12] This observation should be considered when evaluating the findings for opinion mining obtained on this dataset in previous work. Cepstral coefficients are an important indicator for speaker identification and the recognition of extreme emotions. In order to capture the subtle sentiment differences expressed in rather neutral speech, content-specific features are likely to be more predictive. Unfortunately, these aspects cannot be disentangled for the current dataset and we consider our analysis an important contribution that should be considered for future work on the *MOUD* dataset.

Phase	System	Correct	Tie	Wrong
Opening	**Our Approach**	**19**	8	**3**
	Brilman and Scherer (2015)	18	**7**	5
Closing	**Our Approach**	13	15	2
	Brilman and Scherer (2015)	**15**	**14**	**1**

Table 3: Number of corrected predictions, ties and wrong predictions for the debate-level.

Persuasiveness Prediction As described above, the debate level accuracy for the persuasiveness prediction tasks is composed by correct, wrong and tied predictions for the two debaters of each team. In Table 3, we see that our approach completely misclassifies only 10% of the debates, but often yields a tied prediction for the two debaters. Unfortunately, information about the persuasiveness of the individual speakers cannot be obtained because they are evaluated as a team. For future work, it might be reasonable

[12] If we perform Leave-One-Speaker-Out cross-validation, the accuracy of our approach drops by 5.1%.

to add an additional layer to the network that learns how to merge the labels for the individuals into a team label. It should be noticed that there exists of course a wide range of additional factors influencing the persuasiveness of the debaters (Hunter, 2016) such as the previous opinion of the audience, the arguments used during the debate, the appearance and the non-verbal behavior of the speakers. Our approach has shown that cepstral coefficients form a very important indicator for persuasiveness that seems to be at least equally predictive as the actual content of the arguments.

7 Conclusions

We implemented a novel domain-agnostic approach for opinion prediction on speech using MFCCs as input representation and a bidirectional LSTM architecture. We evaluated our approach on opinion mining and persuasiveness prediction and found that our results come close to the performance of domain-specific approaches that apply task-specific feature engineering. In a thorough error analysis, we have shown that our approach performs well in identifying speaker-specific characteristics, but should be combined with additional information if subtle differences in the linguistic content need to be identified. Our publicly available implementation can serve as a starting point for more complex domain-specific approaches for a wide range of speech processing tasks. In addition, our analyses have revealed important characteristics of the two datasets that should be taken into account in future work.

Acknowledgements

This work has been supported by the FAZIT-STIFTUNG and by the German Research Foundation (DFG) as part of the Research Training Group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES) under grant No. GRK 1994/1. We would like to thank Kunal Saxena for preprocessing and organizing the Intelligence Squared dataset.

References

Anton Batliner, Björn Schuller, Dino Seppi, Stefan Steidl, Laurence Devillers, Laurence Vidrascu, Thurid Vogt, Vered Aharonson, and Noam Amir. 2011. The automatic recognition of emotions in speech. In Roddy Cowie, Catherine Pelachaud, and Paolo Petta, editors, *Emotion-Oriented Systems*, pages 71–99. Springer Berlin Heidelberg.

Yoshua Bengio. 2009. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127.

Maarten Brilman and Stefan Scherer. 2015. A multimodal predictive model of successful debaters or how I learned to sway votes. In *ACM International Conference on Multimedia*, pages 149–158, New York, NY, USA.

Judee Burgoon, Thomas Birk, and Michael Pfau. 1990. Nonverbal behaviors, persuasion, and credibility. *Human Communication Research*, 17(1):140–169.

Yu-An Chung, Chao-Chung Wu, Chia-Hao Shen, Hung-Yi Lee, and Lin-Shan Lee. 2016. Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder. In *Annual Conference of the International Speech Communication Association*, pages 765–769, San Francisco, CA, USA.

Andrés Coca, Débora Corrêa, and Liang Zhao. 2013. Computer-aided music composition with lstm neural network and chaotic inspiration. In *International Joint Conference on Neural Networks*, pages 1–7, Dallas, TX, USA.

Steven Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610.

Iryna Gurevych, Eduard Hovy, Noam Slonim, and Benno Stein. 2016. Debating Technologies (Dagstuhl Seminar 15512). *Dagstuhl Reports*, 5(12):18–46.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? Analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the ACL*, pages 1589–1599, Berlin, Germany.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1–32.

Lawrence Hosman, Thomas Huebner, and Susan Siltanen. 2002. The impact of power-of-speech style, argument strength, and need for cognition on impression formation, cognitive responses, and persuasion. *Journal of Language and Social Psychology*, 21(4):361–379.

Anthony Hunter. 2016. Computational persuasion with applications in behaviour change. In *Computational Models of Argument*, pages 5–18, Potsdam, Germany.

John Laver and Peter Trudgill. 1979. Phonetic and linguistic markers in speech. In Klaus Scherer and Howard Giles, editors, *Social Markers in Speech*. Cambridge University Press.

Marco Lippi and Paolo Torroni. 2016. Argument mining from speech: Detecting claims in political debates. In *AAAI Conference on Artificial Intelligence*, pages 2979–2985, Phoenix, AZ, USA.

François Mairesse, Joseph Polifroni, and Giuseppe Di Fabbrizio. 2012. Can prosody inform sentiment analysis? Experiments on short spoken reviews. In *International Conference on Acoustics, Speech and Signal Processing*, pages 5093–5096, Kyoto, Japan.

Norman Markel, Judith Phillis, Robert Vargas, and Kenneth Howard. 1972. Personality traits associated with voice types. *Journal of Psycholinguistic Research*, 1(3):249–255.

Florian Metze, Jitendra Ajmera, Roman Englert, Udo Bub, Felix Burkhardt, Joachim Stegmann, Christian Müller, Richard Huber, Bernt Andrassy, Josef Bauer, and Bernhard Littel. 2007. Comparison of four approaches to age and gender recognition for telephone applications. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages IV–1089 – IV–1092, Honolulu, HI, USA.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis. In *International Conference on Multimodal Interfaces*, pages 169–176, Alicante, Spain. ACM Press.

Tin Lay Nwe, Say Wei Foo, and Liyanage Chandratilak De Silva. 2003. Speech emotion recognition using hidden markov models. *Speech Communication*, 41(4):603–623.

Dimitri Palaz, Mathew Magimai-Doss, and Ronan Collobert. 2015. Analysis of cnn-based speech recognition system using raw speech as input. In *Annual Conference of the International Speech Communication Association*, pages 11–15, Dresden, Germany.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2014. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *International Conference on Multimodal Interaction*, pages 50–57, New York, NY, USA.

Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013a. Multimodal sentiment analysis of Spanish online videos. *IEEE Intelligent Systems*, 28(3):38–45.

Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013b. Utterance-level multimodal sentiment analysis. In *Proceedings of the ACL*, pages 973–982, Sofia, Bulgaria.

Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Conference on Empirical Methods in Natural Language Processing*, pages 2539–2544, Lisbon, Portugal.

Jimmy Ren, Yongtao Hu, Yu-Wing Tai, Chuan Wang, Li Xu, Wenxiu Sun, and Qiong Yan. 2016. Look, listen and learn - a multimodal lstm for speaker identification. In *AAAI Conference on Artificial Intelligence*, pages 1–7, Phoenix, AZ, USA.

Ruty Rinott, Lena Dankin, Carlos Alzate, Mitesh Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Empirical Methods on Natural Language Processing*, pages 441–450, Lisbon, Portugal.

Haggai Roitman, Shay Hummel, Ella Rabinovich, Benjamin Sznajder, Noam Slonim, and Ehud Aharoni. 2016. On the retrieval of wikipedia articles containing claims on controversial topics. In *International Conference Companion on World Wide Web*, pages 991–996, Geneva, Switzerland.

Stuart Rosen. 1992. Temporal information in speech: Acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 336(1278):367–373.

Klaus Scherer. 2003. Vocal communication of emotion: a review of research paradigms. *Speech Communication*, 40(1-2):227–256.

Björn Schuller, Stefan Steidl, Anton Batliner, Elmar Nöth, Alessandro Vinciarelli, Felix Burkhardt, Rob van Son, Felix Weninger, Florian Eyben, Tobias Bocklet, Gelareh Mohammadi, and Benjamin Weiss. 2012. The INTERSPEECH 2012 speaker trait challenge. In *Annual Conference of the International Speech Communication Association*, pages 254–257, Portland, OR, USA.

Björn Schuller, Stefan Steidl, Anton Batliner, Julien Epps, Florian Eyben, Fabien Ringeval, Erik Marchi, and Yue Zhang. 2014. The INTERSPEECH 2014 computational paralinguistics challenge: cognitive & physical load. In *Annual Conference of the International Speech Communication Association*, pages 427–431.

Björn Schuller, Stefan Steidl, Anton Batliner, Simone Hantke, Florian Hönig, Juan Rafael Orozco-Arroyave, Elmar Nöth, Yue Zhang, and Felix Weninger. 2015. The INTERSPEECH 2015 computational paralinguistics challenge: nativeness, parkinson's & eating condition. In *Annual Conference of the International Speech Communication Association*, pages 478–482, Dresden, Germany.

Stanley Stevens, John Volkmann, and Edwin Newman. 1937. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190.

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016. Dimensional sentiment analysis using a regional cnn-lstm model. In *Proceedings of the ACL*, pages 225–230, Berlin, Germany.

Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. 2016. Achieving human parity in conversational speech recognition. *CoRR*, abs/1610.05256.