ACOUSTIC AND LEXICAL SENTIMENT ANALYSIS FOR CUSTOMER SERVICE CALLS

*Bryan Li**

Department of Computer Science
Columbia University, New York, NY
bl2557@columbia.edu

Dimitrios Dimitriadis, Andreas Stolcke

Speech and Dialog Research Group
Microsoft, Bellevue, WA
[didimit, anstolck]@microsoft.com

ABSTRACT

We describe the development of a sentiment analysis system for customer service calls, starting with the data acquisition and labeling, and proceeding to the algorithmic information extraction and modeling process from both spoken words and their acoustic expression. The proposed system is based on the combination of multiple acoustic and lexical models in a late fusion approach. Acoustic aspects of sentiment are captured by utterance-level features based on aggregated openSMILE and raw cepstral features, and further augmented with an energy contour model. Lexical aspects are captured by back-off n-gram language models. These models are found to combine effectively, showing different strengths as pertains to positive and negative sentiment detection.

Index Terms— sentiment analysis, audio feature extraction, acoustic modeling, language modeling, multimodal fusion

1. INTRODUCTION

Sentiment analysis has recently gained much interest in the paralinguistic processing world. It focuses on developing techniques for automatically recognizing the polarity (positive/neutral/negative) of human emotional states or attitudes expressed in natural language. An applicable field for sentiment analysis is in customer support conversations. As a forward-facing function, customer support is critical among other things in maintaining a company's customer relations. These conversations are dyadic, involving a customer and an agent, and typically concern the customer's issues with products or services. Therefore, special focus should be placed on detecting negative emotions, such as frustration, annoyance, or anger. Automatic sentiment analysis can either detect problems in real-time (e.g., leading to escalation), or aid in post-hoc review and analytics of customer-agent interactions. In the context of automated customer dialog, detected sentiments could affect the dialog strategies.

Human oral communication consists of both lexical and acoustic content. The lexical part is the message as conveyed by the words spoken, whereas the acoustic part cues encode how this message is delivered, e.g., by differential use of prosody. Meaning can be altered depending on subtle changes in these cues. Thus, it is expected that emotion and sentiment recognition would benefit from considering both lexical and acoustic features.

In this paper we explore architectures for developing a sentiment recognizer on real-world call center data using acoustic and lexical cues. Such a scenario involves a series of challenges such as the heavily imbalanced data, or sociocultural phenomena such as sarcasm. Most systems described in the literature work with acted, high-quality data, where sentiment analysis is easier. Our work is among the first to address these issues on real-world, telephone quality data. We first develop individual acoustic and lexical models, then perform late fusion to significantly improve results over the individual models.

Some prior work on sentiment analysis for customer support calls was performed on text transcribed from acted calls [1]. As for acoustic sentiment analysis, researchers have worked with frame-level [2, 3] or utterance-level features [4]. Two of the approaches regarding the fusion of multiple knowledge sources are early and late fusion [5, 6, 7]. There is also extensive work on fusion methods for sentiment and emotion analysis [8, 9, 10].

2. DATA COLLECTION AND ANNOTATION

The dataset is created from recorded Microsoft customer support calls, including both B2C (business-to-consumer) and B2B (business-to-business) interactions, and for a range of products and services. It consists of 1957 sessions in total. Each conversation has been automatically segmented into utterances, and separated into agent and customer speech based on channel (although occasional mix-ups occur due to crosstalk or processing glitches). An initial transcription pass is done automatically, followed by human transcription. For the purposes of our task, we will use only the audio data from the customer side.

2.1. Reference label creation

Each utterance is labeled for sentiment by three judges in the Microsoft UHRS crowd-sourcing system. All judges must pass a qualifying test, concurring with a 'gold set' of labeled utterances at least 75% of the time. Judges listen to the entire conversations, one utterance at a time. Additionally, human-transcribed text is presented on-screen for both current and context utterances (three previous and three following). The context displayed includes both agent and customer utterance.

Based on this information, judges label each utterance on a five-point scale: *Clearly Positive, Somewhat Positive, Neutral, Somewhat Negative*, and *Clearly Negative*. Additionally, there are three labels for utterances that are not part of the sentiment task: *Agent Speech*, for agent-spoken utterances; *Can't Label*, for poor-quality or non-speech audio; and *NIS (Not Intended for Service)*, for speech not directed to the agent ("side speech").

Reference labels are created based on the following process:

1. Combine labels: Somewhat and Clearly Positive/Negative to Positive/Negative; Agent Speech, Can't Label, NIS to NA.
2. Discard utterances marked as NA by a majority (at least two) of the judges.

*This work was done during a research internship at Microsoft.

Table 1. Examples of reference label creation

Label 1	Label 2	Label 3	Reference
Neutral	Clearly Positive	Somewhat Positive	Positive
Can't Label	Neutral	Neutral	Neutral
Neutral	Somewhat Negative	Somewhat Positive	Removed
Agent Speech	Agent Speech	Somewhat Positive	Removed

3. Discard utterances without majority agreement on one of {Positive, Neutral, Negative}.

Table 1 provides examples of the filtering process. The final reference label set contains 111,665 utterances out of 122,364 (8.7% removed), labeled on a three-point sentiment scale.

2.2. Analysis of dataset statistics

As is common for linguistic datasets involving marked and unmarked cases, the classes are heavily imbalanced, as shown in Table 2. Metrics and training procedures described later are designed to deal with this imbalance.

Inter-annotator agreement (before filtering, after collapsing to three classes) is shown in Table 3. The metric used is Fleiss' kappa [11], a generalization of Cohen's kappa for multiple labelers in which each pair of labels for an item is treated as a potential agreement or disagreement. As in Cohen's kappa, the statistic expresses the relative difference between observed and chance agreement, ranging from 0 to 1. Judges overwhelmingly prefer labeling utterances as Neutral over Positive or Negative. κ is only 0.48, whereas values over 0.6 are desirable. Human agreement on matching reference labels is 84.45% UAR (unweighted average recall), a measure of accuracy simulating equal class priors that is commonly used for emotion recognition tasks. It is clear that sentiment classification on this dataset is difficult even for the human judges.

Manual inspection of the labeler disagreements reveals that errors can be categorized in a few groups. A common error is missing the tone of voice; for example, sarcasm causes a mismatch between the words said and the sentiment conveyed. Another is marking polite statements, such as "thank you" and "have a nice day" as Positive—these are Neutral unless the customer sounds like they mean it. Sometimes, judges assign Negative to a matter-of-fact description of a frustrating situation. Other cases include interpreting profanity as Negative, laughs as Positive, and failing to consider conversational context. While our task instructions attempted to draw attention to all of these common pitfalls, future labeling efforts could benefit from training judges more thoroughly for problematic cases.

2.3. IEMOCAP Dataset

Although our target task is sentiment, not emotion recognition, we chose an emotion recognition task to validate our acoustic models, since public datasets and prior results are available for the latter task. The IEMOCAP dataset [12] consists of dyadic interactions between actors. There are five sessions, each with a male and a female speaker, for 10 unique speakers total. As in [4], we consider only utterances with majority agreement ground-truth labels, and only those labeled happy, sad, angry or neutral. We combine happy and excited for the happy category. The final dataset contains 5531 utterances (1103 angry, 1708 neutral, 1084 sad, 1636 happy).

Table 2. Reference label distribution

Label	Count	Percent
Neutral	103906	93.1%
Negative	5733	5.1%
Positive	2026	1.8%

Table 3. Inter-annotator agreement (Fleiss' kappa)

	Positive	Neutral	Negative	NA	κ
Positive	6940	9151	185	232	0.42
Neutral	9151	555868	26173	16980	0.91
Negative	185	26173	20266	818	0.43
NA	232	16980	818	34762	0.66

3. SYSTEM DESCRIPTION

In designing the sentiment analysis system, we experimented with both acoustic and lexical models. Our acoustic model based on utterance-level features is compared to a state-of-the-art frame-level system [3]. We trained two lexical models based on backoff n-gram language models, one for automatic (ASR) transcriptions and one for human transcriptions. Additionally, we trained a prosodic model modeling energy dynamics over the utterance. Finally, we performed late fusion of the individual models.

3.1. Metrics Used

There are two formulations of this sentiment analysis problem. Straightforward is a three-way classification: given an input utterance, predict an output label from {Negative, Neutral, Positive}. To address heavy class imbalance, the metric used is UAR, which is equivalent to class-balanced label accuracy. We also applied a 2-class detection framework and evaluated two models: Positive versus non-Positive and Negative versus non-Negative. The class-prior-invariant metric here is equal error rate (EER), the point at which false detection rate equals miss rate based on thresholding.

3.2. Acoustic Model

The acoustic model is concerned with paralinguistic features– how something is said. Audio files are standardized to a 16kHz sampling rate. For each utterance, we extract 988 features using the emobase feature set from openSMILE [13]. This consists of low-level descriptors such as intensity, loudness, Mel-frequency cepstral coefficients, and pitch. For each low-level descriptor, functionals such as max/min value, mean, standard deviation, kurtosis, and skewness are computed. Finally, global mean and variance normalization are applied to each feature, using training set statistics. These features serve as inputs to a deep neural network classifier implemented in PyTorch [14]. The model thus captures acoustic-prosodic features aggregated over the utterance.

3.3. Cepstral Model

We compare the above model to a previously developed state-of-the-art system which we term the cepstral model [3]. This system uses each utterance as a mini-batch of frames, then mean-pools the hidden activation layers to obtain an utterance-level representation. Silence frames are filtered out, so that training is done using only frames with voice activity. Twenty-five frames are selected for each utterance, with 58 features per frame, including log-mel spectrum, pitch, and voice, for 1450 features per utterance (vs. 988 features in our acoustic model).

3.4. Energy Contour Model

This acoustic-prosodic model specifically captures the shape of the speech energy contour over the duration of the utterance [15]. Zeroth and first-order Mel frequency cepstral coefficients are computed every 10 milliseconds, and the contours of these values over windows of 200 ms are characterized by computing a discrete cosine transform (DCT) in the temporal domain. The first 5 DCT values for cepstral coefficient c0 are retained, as are the first 2 DCT values for c1, resulting in a 7-dimensional feature vector for every 200 ms window. For classification purposes we train Gaussian mixture models for each target class, and use length-normalized likelihoods as discriminant scores [15] .

3.5. Lexical Classifier

The lexical content of utterances gives important cues to sentiment, and word-based classifiers exploit this fact. To obtain a sentiment score that may be combined statistically with other knowledge sources we trained statistical language models for each target class. The models are backoff trigrams with Witten-Bell smoothing, estimated over the full training set vocabulary. The class log likelihoods divided by the utterance lengths serve as discriminant scores [16].

Our automatic transcripts came from a generic speech recognition service that was not adapted to the customer service domain. For comparison purposes we also built a lexical classifier on the human utterance transcripts, simulating close-to-perfect speech recognition.

3.6. Model Fusion

Humans use both acoustic and lexical modalities to make informed decisions about sentiment of statements. A statistical model combining both modalities should likewise make better decisions. Two approaches are early (feature-level) or late (decision-level) fusion. In early fusion, joint feature vectors are created by combining the features from multiple modalities and fed into a unified model. In late-fusion, models are independently built for each modality, and outputs of these models are used as features to train a downstream combined model.

Early fusion assumes a level of temporal synchrony between the individual modalities, which may not be easy achieved. In contrast, the individual models in late fusion consider features from only one modality, obscuring time-varying properties but alleviating the assumption of time synchrony.

For our setting, it is unwieldly to combine n-gram textual features with extracted audio features at the feature level. Poria et al. [8, 17] achieved similar results with early and late fusion of the audio, textual, and visual modalities of IEMOCAP: 73.22% and 73.25% weighted accuracy, respectively. Therefore, we work exclusively with late fusion.

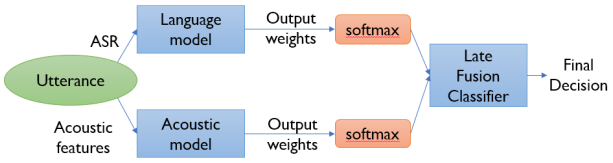
The features for late fusion are the output scores of each class from each model. For each set of scores we apply a softmax to obtain the posterior probabilities. We then train a machine learning classifier, such as logistic regression or a support vector machine (SVM) as the fusion model. Fig. 1 depicts the fusion framework.

4. EXPERIMENTS

4.1. Validation on IEMOCAP

Adhering to the methodology of [4, 18, 19], we perform a ten-fold cross-validation scheme, stratified by speaker. For each evaluation,

Fig. 1. Decision-level fusion framework



we use eight folds for training, one fold for early-stopping and hyperparameter tuning, and one fold for reporting test results. Acoustic features and the DNN model are similar to Sec. 3.2, except that five emotion labels are estimated, rather than three sentiment labels.

The UAR of our acoustic model is 61.74%, which may be compared to the state-of-the-art of 65.70% [4]. We also tried adding contextual features using a window of three feature vectors before and after the current utterance, for a total of 7×988 features per utterance. This brings our UAR to 64.96%. The results show that our acoustic model, while computationally simple, yields reasonable results for emotion-related classification.

4.2. Sentiment Experiment Setup

The customer support dataset is split into training, validation, and test sets randomly by session, with 1557, 200, and 200 (about 80/20/20%) sessions, respectively. We use the training set for acoustic and language model training, the validation set for hyperparameter tuning, early stopping and fusion model estimation, and the test set for reporting results.

There are two approaches to handling class imbalance. First, we can resample the training set, preserving all Negative labels, oversampling Positives and undersampling Neutrals to the number of Negatives. Second, we can re-weight the objective function for each sample such that each class has the same aggregate weight. For example, missing one Negative sentiment incurs the same loss as missing 93/5 Neutral sentiments, in inverse proportion to the class priors.

For our acoustic and lexical models, we choose the second approach, as using more data gave better results. For the cepstral model of [3] a resampling approach was taken. 5238 utterances from the training split are selected, with roughly equal class proportions. The entire development and test sets are used in both approaches.

4.3. Sentiment Classification Results

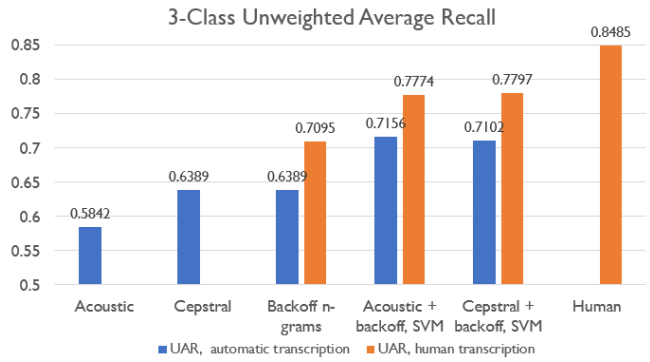
Hyperparameters for our DNN on acoustic features are hidden size=1024, hidden layers=2, initial learning rate=0.01, learning anneal=0.75, batch size=32, dropout=0.5, activation=ReLU, optimizer=SGD with momentum.

Separate lexical models are trained on both ASR and human transcripts. Fig. 2 shows the results. (We also tried logistic regression and DNNs for late fusion, but report only SVM results for brevity.) Contextual features did not help for this dataset, likely due to the heavy imbalance (93% Neutral) adding little distinguishing information for individual utterances.

The cepstral model outperforms our acoustic model by 5.47% absolutely (63.89% vs. 58.42%). The lexical model based on ASR also achieves 63.89% UAR. A better UAR of 71.56% is achieved by a fusion of acoustic and lexical models, slightly better (by 0.54%) than a fusion of cepstral and lexical models. We thus find that fusion

[0]Chen et al. [2] propose a frame-based 3DCRNN, and report results of 64.96% using only improvised segments of IEMOCAP.

Fig. 2. Results for sentiment classification task



mutes differences between the two acoustics-based models. Regardless which acoustic model is used, it is evident that acoustic and lexical information complement each other.

Considering results from language models trained on human transcripts, the errors incurred by ASR do affect sentiment classification. There is a 6-7% absolute increase in UAR for both fusion models compared to their ASR-based counterparts. Cepstral fusion now performs 0.23% better than acoustic, but the observation about diminished differences between acoustic models after fusion holds.

Note that the "human performance" of 84.85% is a biased measure as it was calculated with the set of judgments used to create reference (majority) labels; we can expect independent human relabeling performance to be lower.

4.4. Sentiment Detection Results

As discussed in Section 3.1, for sentiment detection we generate two models, ones for positive sentiment, and one for negative sentiment, with EER as the metric in both cases. On these two-class problems we train models using the same setup as above. In addition to the two acoustic models and the lexical model, we add the energy contour model to the model fusion. Logistic regression is used for fusion. Fig. 3 shows detection error trade-off (DET) plots and EERs.

For Negative detection, our acoustic model (29.7% EER) performs better than the Cepstral (31.7%) and lexical (33.1%) models. Fusion of acoustic and lexical models decreases EER to 24.8%, and including energy contours sees a further improvement to 24.6%. The fusion results with human transcripts ("refwords") is 21.7%.

For Positive detection, the lexical model (18.5% EER) performs far better than the acoustic models (23.7%, 31.3%). Still, combining acoustics with the lexical model gives the best result, as 16.4% EER. Energy contours did not help for Positive detection. System fusion with human transcripts gives 10.3% EER.

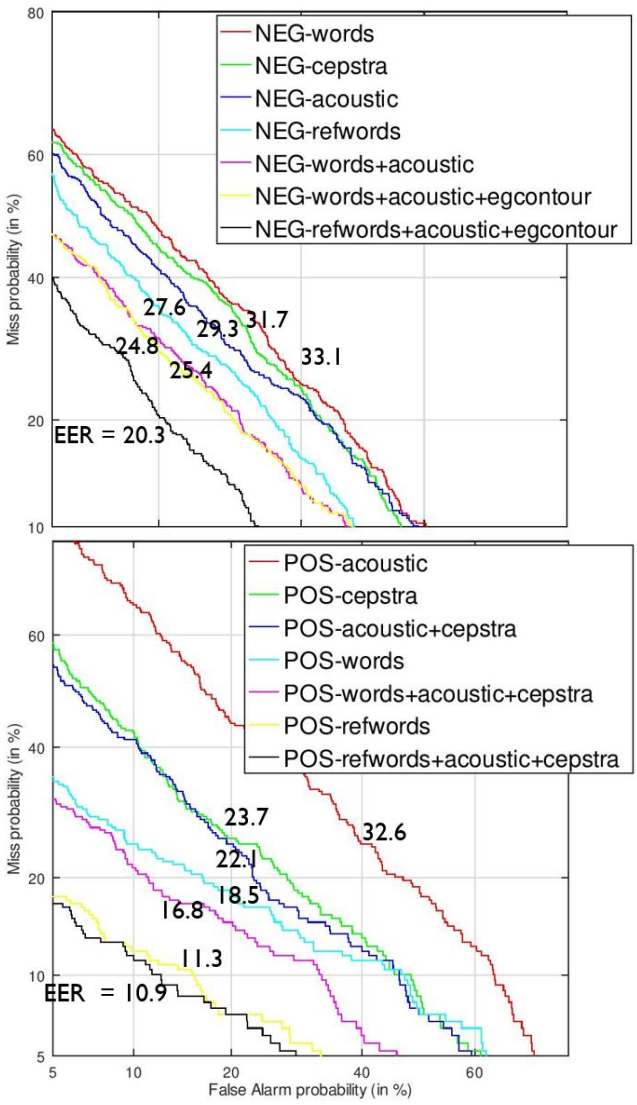
These results indicate that people express positivity more in their word choices, whereas negativity is expressed more through tone of voice. This is consistent with politeness: customers are happy to explicitly praise helpful agents, but express displeasure indirectly.

5. CONCLUSIONS

We studied sentiment extraction from spoken utterances in a large corpus of customer interactions with support agents. The task is hard even for humans: individual labelers have a residual error rate of at least 15% relative to the reference (majority) labels.

We found that automatic sentiment extraction requires both acoustic-prosodic and lexical modeling for best results, both for

Fig. 3. Negative (top) and positive (bottom) sentiment DET plots



three-way classification and binary sentiment detection. Lexical cues dominate acoustic ones for positive sentiment detection, but the reverse is true for negative sentiments. A score-level combination of both cue types is always helpful, giving about 40% relative error reduction over a single cue type. Smaller additional gains can be obtained by combining different types of acoustic features and models, such as openSMILE-based utterance-level features with frame-level cepstral and energy-contour models. Word-based modeling can be effective even with high word-error rates, using matched training and test conditions, but classification error still increases considerably (27% relative) compared to near error-free (human) transcripts.

Plenty of future work awaits, such as neural lexical models, or joint acoustic/lexical models, as well as comparisons with sentiment classification in other domains.

6. ACKNOWLEDGMENTS

We would like to thank our colleagues Ivan Tashev, David Johnston, Dimitra Emmanouilidou, and Ashley Chang.

7. REFERENCES

[1] Souraya Ezzat, Neamat El Gayar, and Moustafa M Ghanem, "Sentiment analysis of call centre audio conversations using text classification," *International Journal of Computer Information Systems and Industrial Management Applications*, vol. 4, no. 1, pp. 619–627, 2012.

[2] Mingyi Chen, Xuanji He, Jing Yang, and Han Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[3] Zhong-Qiu Wang and Ivan Tashev, "Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks," in *Proc. IEEE ICASSP*, 2017, pp. 5150–5154.

[4] John Gideon, Soheil Khorram, Zakaria Aldeneh, Dimitrios Dimitriadis, and Emily Mower Provost, "Progressive neural networks for transfer learning in emotion recognition," *arXiv preprint arXiv:1706.03256*, 2017.

[5] J. Wagner, E. Andre, F. Lingenfelser, and J. Kim, "Exploring fusion methods for multimodal emotion recognition with missing data," *Affective Computing, IEEE Transactions on*, vol. 2, pp. 206 – 218, Jan. 2012.

[6] Florian Lingenfelser, Johannes Wagner, Jun Deng, Raymond Brueckner, Björn Schuller, and Elisabeth André, "Asynchronous and Event-based Fusion Systems for Affect Recognition on Naturalistic Data in Comparison to Conventional Approaches," *IEEE Transactions on Affective Computing*, vol. 7, 2016, 13 pages, to appear (IF: 3.466, 5-year IF: 3.871 (2013)).

[7] Björn Schuller, "Speech Emotion Recognition: 20 Years in a Nutshell, Benchmarks, and Ongoing Trends," *Communications of the ACM*, 2016.

[8] Soujanya Poria, Haiyun Peng, Amir Hussain, Newton Howard, and Erik Cambria, "Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis," *Neurocomputing*, vol. 261, pp. 217–230, 2017.

[9] S. Motamed, S. Setayeshi, A. Rabiee, and A. Sharifi, "Speech emotion recognition based on fusion method," *Journal of Information Systems and Telecommunication*, vol. 5, pp. 50–56, Dec. 2017.

[10] J. Yana, W. Zhenga, Z. Cuib, C. Tanga, T. Zhangc, and Y. Zonga, "Multi-cue fusion for emotion recognition in the wild," *Neurocomputing*, vol. 309, pp. 27–35, 2018.

[11] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[12] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.

[13] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, New York, NY, USA, 2010, MM '10, pp. 1459–1462, ACM.

[14] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in PyTorch," in *NIPS-W*, 2017.

[15] Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, and Larry Heck, "Learning when to listen: Detecting system-addressed speech in human-human-computer dialog," in *Proc. Interspeech*, Portland, Oregon, Sept. 2012, pp. 334–337.

[16] Heeyoung Lee, Andreas Stolcke, and Elizabeth Shriberg, "Using out-of-domain data for lexical addressee detection in human-human-computer dialog," in *Proceedings North American ACL/Human Language Technology Conference*, Atlanta, GA, June 2013, pp. 221–229.

[17] Samarth Tripathi and Homayoon S. M. Beigi, "Multi-modal emotion recognition on IEMOCAP dataset using deep learning," *CoRR*, vol. abs/1804.05788, 2018.

[18] Aharon Satt, Shai Rozenberg, and Ron Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *Proc. Interspeech*, 2017.

[19] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. IEEE ICASSP*, March 2017.