Interspeech 2018
2-6 September 2018, Hyderabad, India

Using Prosodic and Lexical Information for Learning Utterance-level Behaviors in Psychotherapy

Karan Singla[1], *Zhuohao Chen*[1], *Nikolaos Flemotomos*[1], *James Gibson*[1], *Dogan Can*[1], *David C. Atkins*[2], *Shrikanth Narayanan*[1]

[1]Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA, USA
[2]Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA
[1]`sail.usc.edu`, [2]`datkins@u.washington.edu`

Abstract

In this paper, we present an approach for predicting utterance level behaviors in psychotherapy sessions using both speech and lexical features. We train long short term memory (LSTM) networks with an attention mechanism using words, both manually and automatically transcribed, and prosodic features, at the word level, to predict the annotated behaviors. We demonstrate that prosodic features provide discriminative information relevant to the behavior task and show that they improve prediction when fused with automatically derived lexical features. Additionally, we investigate the weights of the attention mechanism to determine words and prosodic patterns which are of importance to the behavior prediction task.

Index Terms: prosody, mutlimodal learning, behavioral signal processing

1. Introduction

Both the words that are spoken and the way in which they are spoken are of fundamental importance in psychotherapy conversations. There are many studies demonstrating the importance of the lexical channel for predicting behaviors in psychotherapy [1, 2, 3], but multimodal information like visual and acoustic cues also carry a wealth of information that is potentially complimentary to the lexical modality [4], and has received less attention in this domain.

In this paper, we focus on data from Motivational Interviewing (MI) sessions, a type of psychotherapy focused on behavior change. Behavior is generally monitored and codified in the form of behavioral coding, which is the process of a human manually observing a session and annotating the behaviors of the participants in that session, as defined by a coding manual. The Motivational Interviewing Skill Code (MISC) manual defines both session and utterance level behaviors that are of interest for understanding therapist efficacy in MI [5]. Several approaches have been proposed for automating the behavioral coding procedure to predict gestalt session level behaviors, especially therapist empathy, using lexical information (both manually and automatically derived) [6], speech rate entrainment [7], and prosody [8]. At the utterance level, automating the behavioral coding process has been entirely focused on linguistic features [3, 9, 10]. In this work, we inspect: if utterance-level behavior codes can be predicted using prosodic cues; and if prosodic information can assist lexical information in making better predictions of participants' behaviors.

We hypothesize prosodic information such as variation in pitch, loudness, pause, etc. will have an important role in predicting behaviors in psychotherapy and can assist lexical features for making improved predictions. Therefore, we propose a multimodal approach for predicting behavior codes that exploits both prosodic and lexical information at the word level. We show that prosodic information can assist lexical information in making a multimodal prediction. Our multimodal architecture is largely inspired from [11] and [3]. Thus, we use Bidirectional long short term memory (LSTM) networks with a self-attention mechanism for predicting MISC Codes at the level of utterances using multimodal information i.e prosodic and lexical features. Our network is different from prior research, as we use a unified architecture, i.e., the same model for predicting therapist/client codes.

2. Related Work

Several computational models have been proposed for predicting MISC behavioral codes at the utterance level [2, 3, 9]. Researchers have addressed this problem by using variety of features, such as word n-grams and linguistic features [1] and recurrent neural networks (RNNs) with word embedding features [3, 11]. Methods using RNNs have shown superior performance to other models (e.g., MaxEnt) for utterance level behavioral code prediction [3]. The success of these RNN based models demonstrates that learning in-domain word representations and parameters in an end-to-end fashion offers better modeling for this task. These models typically use separate models for therapist and client codes, whereas we propose a unified model which still uses utterance level speaker information.

Self-attention mechanisms, which enable models to attend to particular words based on input for predicting output classes, have been used widely in natual language processing [12, 13, 14] and speech processing [15]. Recently [11] extended the work from [3] by using a self-attention mechanism for predicting utterance level MISC codes. They show how attention can improve the interpretability and help in better understanding the decisions made by the model.

While using multimodal information is rather unexplored in predicting utterance level MISC codes, it has been an exciting venue of research in some other related domains such as multimodal parsing [16], prediction of psychological disorders [17], and audio-visual applications like speech recognition [18]. Our proposed multimodal approach is similar to [16] in the sense that we also concatenate prosodic features obtained from audio signals with lexical features to get word-level representations.

3. Data

In this paper, we use data from Motivational Intervewing sessions, presented in [19, 20], for behavior prediction. Table 1 shows statistics of utterance level MISC data used in this paper after removing utterances where there is a speaker overlap.

10.21437/Interspeech.2018-2551

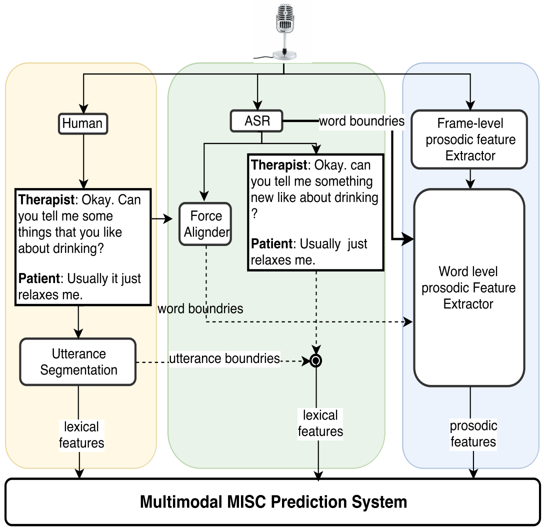
Human
ASR
word boundries
Frame-level prosodic feature Extractor
Therapist: Okay. Can you tell me some things that you like about drinking?
Patient: Usually it just relaxes me.
Force Alignder
Therapist: Okay. can you tell me something new like about drinking ?
Patient: Usually just relaxes me.
Word level prosodic Feature Extractor
word boundries
Utterance Segmentation
utterance boundries
lexical features
lexical features
prosodic features
Multimodal MISC Prediction System

Figure 1: *Data Pipeline. Left part shows data flow to get human transcribed utterance level data. Middle part shows the data flow to create similar data using automatic transcription. Right, shows pipeline for extracting prosodic features at the word level.*

For Therapist codes, a reflection (REF) is a reflective listening statement made by the counselor in response to a client statement. Question (QES) is either an open or a close question asked by the therapist. Other (OTH) can include Advise, Affirm, Confront, Facilitate etc. among other therapist behaviors. Client behavior is observed from three dimensions. In a follow-neutral turn (FN), there is no indication of client inclination either toward or away from the target behavior change. Client behavior is otherwise marked with a positive (POS) or negative (NEG) valence, depending on whether it reflects inclination toward (POS) or away from (NEG) the target behavior change. Figure 1 shows a snapshot for our data pipeline.

Our data flow pipeline handles three main types of data:

- **Human transcribed text data:** Audio signals are first transcribed at speaker turn level by humans. Each turn is then segmented by humans experts to get utterances. Human experts then annotate these Therapist and Patient utterances with MISC codes.
- **Automatically transcribed text data:** Middle part of Figure 1 shows the pipeline where we use utterance text generated using ASR. We use the ASR system presented in [21] to get automatic transcripts from audio signals. ASR does use speaker turn boundaries marked by human transcribers. Reported word error rate (WER) is 44.1 % [21], where a major chunk of errors is because of substitution (27.9 %). We use utterances segmentation information and MISC labeling done by human experts for generating this data.
- **Word level Prosodic feature Extractor :** Using audio signals we first extract prosodic features at frame level. As our multimodal approach uses word-level features, for human transcribed training data we use a force aligner [21] to align human transcribed transcripts to get word boundaries. For automatic generated text data, ASR directly gives word boundaries to extract word-level prosodic features. Prosodic features extraction is described in more detail in section 4.1.

Table 1: *Frequency of samples for each class label in the pre-processed MISC data.*

Code	Description	#Train	#Test
Therapist			
REF	Reflection	6577	3456
QES	Question	6546	3348
OTH	Other	13112	7625
Total		26235	14429
Client			
FN	Follow/Neutral	22020	12229
NEG	Sustain Talk	4019	1660
POS	Change Talk	3151	1272
Total		29190	15161

4. Method

4.1. Features

Our multimodal system can exploit two types of features; namely features exploiting prosodic information and lexical text.

4.1.1. Prosodic

- **Prosodic** (a)**:** We use pitch, loudness and jitter as prosodic features. We extract frame-level pitch using pyaudioanalysis [22] & loudness and jitter using *Praat* [23], where frame size is 50ms and step size is 25ms. We then calculate the mean and standard deviation (std) across frames within a word to represent 6 (3 mean, 3 std) word level features.
- **Pause** (p)**:** We also encode word-level pause information i.e pause taken before and after a word. For each word, pause is quantized into a 10 bit vector (5 for pause before and 5 for pause after) depending upon if pause time lies before, between and after {0.01, 0.06, 0.2, 1.0} seconds. These boundaries are selected so that the words are approximately uniformly distributed in those bins.
- **Average word length** (wl)**:** We keep an additional feature for marking word length i.e number of frames used to speak a word. We normalize word length by average word length over the entire train and test dataset separately.

We concatenate a_i, p_i and wl_i to get a 17-dimensional prosodic feature representation A_i for each word W_i.

Speaker normalization: There are various different studies collected across different settings that are part of this dataset with different speakers, both in terms of therapists and patients. Therefore, we do a two-fold speaker normalization. First, we do a z-normalization for each audio feature for each study type and second, we normalize each audio feature for each speaker (Therapist and Patient) for each audio session.

4.1.2. Lexical

For textual features we remove all punctuations and lower case all words. We also replace any words having frequency less than 5 with the <unk> symbol. Our final vocabulary has 11219 unique words. Each word W_i is represented by a 100-dimensional vector T_i, initialized using a uniformly distributed

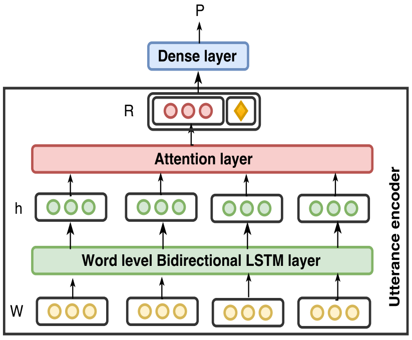

Figure 2: *Architecture for Utterance Encoder.* ♦ *can be 1 or 0 for Therapist and Patient utterance respectively.*

random word embedding layer which we learn as a part of the encoder described in the following section.

4.2. Utterance Encoder

We assume that each utterance is represented by a word sequence $W = \{W_0, W_1, \cdots, W_{L-1}\}$, where L is the number of words in the utterance. Each word can be represented either by prosodic features, or by lexical text, or both. We then assume there exists a function $c = f(W)$ that maps W to a behavioral code $c \in 1, 2, \cdots, C$, with C being the number of defined code types. Our goal is to find the function $f*$ minimizing the error between the predicted and expert-annotated codes.

We use a parametric composition model to construct utterance-level embeddings from word-level embeddings. We process word-level embeddings with an LSTM [24, 25] and then take a weighted average of the LSTM outputs using a task-specific attention model [13]. There are various implementations of LSTMs available; in this work we use an implementation based on [26]. The LSTM outputs (hidden states) h_i contextualize input word embeddings W_i by encoding the history of each word into its representation. The attention layer can be seen as a mechanism for accessing internal memory of the system, i.e. the hidden states of the LSTM. It can learn what to retrieve from the memory while constructing an utterance representation. For example, it can learn to ignore stop-words or downweight words that are not essential for predicting behavioral codes. We use an attention layer (equations 1-3) with an internal context vector [13].

$$k_i = \tanh(Wh_i + b) \tag{1}$$

$$\alpha_i = \mathrm{softmax}(k_i^T a) \tag{2}$$

$$R = \sum_i \alpha_i h_i \tag{3}$$

The attention layer first applies a one-layer MLP to its inputs h_i to derive the keys k_i. Then it computes the attention weights by applying a softmax nonlinearity to the inner products between the keys k_i and the internal context vector a . Finally it computes the sentence representation R by taking a weighted average of its inputs. The context vector a is a model parameter that is initialized with uniform weights so that it behaves like an averaging operation at the beginning of the training.

We then concatenate oracle speaker information (Therapist (1) vs Client (0)) to R before it's passed through a dense layer to get P. P is a C-dimensional vector on which we take softmax to predict MISC label. We also show experiments with the model where the utterance encoder doesn't use the attention mechanism. Instead, it just uses the last hidden state from the LSTM layer (*LSTM-l*). We will refer to our model with attention as *LSTM-a*.

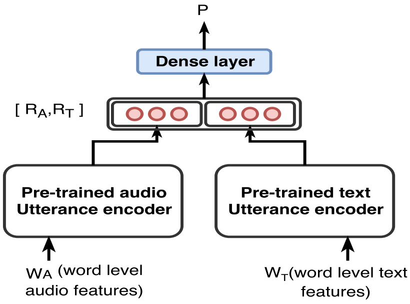
P
Dense layer
[R_A,R_T]
Pre-trained audio Utterance encoder
Pre-trained text Utterance encoder
W_A (word level audio features)
W_T(word level text features)

Figure 3: *Multimodal system architecture using Comb-LF approach.*

4.3. Multimodal Approach

Prosodic feature vector A_i for each word W_i is first processed through a dense layer to get a high dimensional representation, which matches the lexical representation of the word in terms of dimensionality before it's fed into the LSTM layer.

We do a multimodal combination by two methods :

- **Comb-WL** : Word-level lexical features T and prosodic features A are word-wise concatenated to make input W before feeding it to the utterance encoder for predicting MISC labels.
- **Comb-LF** : As show in Figure 3, we first train utterance encoder using lexical features and a separate encoder using prosodic features. For fusion, word-level audio sequence W_A is processed through a pretrained utterance encoder trained on audio data and similarly W_T is processed separately to get R_A and R_T respectively. R_A and R_T are then concatenated and then passed through another dense layer to get the C-dimensional output P. This allows us to tune the entire system for multimodal information in an end-to-end fashion

Training Routine : The batch size is 40 utterances. LSTM hidden state dimension is 100 (50 forward, 50 backward). We use dropout at the embedding layer with drop probability 0.3. Dense layer is of 100 dimensions. The model is trained using the Adam optimizer [27] with a learning rate of 0.01 and an exponential decay of 0.98 after 10K steps (1 step = 40 utterances). We weight each sample using class weights derived using class frequencies. Formally, the weight given to a sample belonging to class i is

$$w_i = \frac{\tilde{w_i}}{\sum_i \tilde{w_i}}, \text{where } \tilde{w_i} = \frac{\text{total \#samples}}{\text{\#samples}_i}$$

5. Experiments & Results

5.1. Behavior (MISC) Code Prediction

5.1.1. Baselines

We train models with just lexical features (*Text*) and just prosodic information (*Prosodic*). The first two rows in Ta-

Table 2: *Results for single modality (Text, Prosodic) and multi-modal approach for human generated test data.*

Features	Avg. F1-score	
	LSTM-l	LSTM-a
Text	0.54	0.57
Prosodic	0.42	0.42
Comb-WL	0.56	0.58
Comb-LF	0.58	0.60

Table 3: *Results for using automatically generated transcripts from ASR*

Features	Avg. F1-score
ASR text	0.47
Comb-WL	0.52
Comb-LF	0.53

ble 2 show class averaged f-scores for our baseline systems where we only use one modality (lexical-features or prosodic-features). Model with just lexical features (*Text*) performs better than the model which only uses word-level prosodic information (*Prosodic*). *Prosodic* model performs better than majority class baseline, which is 0.33, since we report class averaged f-scores. This shows that prosodic information alone is quite informative about predicting behavior codes.

5.1.2. Human vs Automatically transcribed Data

Bottom part of Table 2 shows that multimodal information can in fact help in making a better prediction of behavior codes compared to single modality models (*Text* and *Prosodic*). We get best results for *Comb-LF* where we do late fusion of utterances. Scores where we use attention are better, therefore, we use model with attention (*LSTM-a*) for further experiments.

Using automatically generated lexical features, results in Table 3 show high gains for multimodal systems. Comb-LF outperforms other models with automatically transcribed lexical features. This number is a bit worse than the model which uses human transcribed lexical features *Text*. Moreover, the Comb-WL model which fuses word level lexical and prosodic information also shows improvements.

5.2. Attention Weight Analysis

Prosodic model generally gives high weight to utterance endings, indicating it's important to attend to the last part of the utterance for predicting behavior. It reinforces the hypothesis that pitch rises at the end of questions which makes it an important marker for discrimination. It also always gives some weight to start of the utterance, along with attending a bit to word *Did* for example in Figure 4. It can also be seen that *Text* model attends to lexical words that are necessary to mark a question. (high weights to words: did, you, say).

5.3. Evaluating on Utterances $>$ 15 words

Ablation experiments where we only choose utterances longer than 15 words (4824 and 5313 samples for Therapist and Patient codes respectively), suggest that *Prosodic* model shows improved performance for longer utterances. Table 4 shows results for this. Scores for *Prosodic* features improve only evaluated for longer utterances. ASR text follows a similar trend. Results for combination experiments are also slightly better when evaluated for longer utterances.

Table 4: *Ablation experiments results when evaluated on utterances longer than 15 words.*

Features	Avg. F1-score
Prosodic	0.48
ASR text	0.50
Comb-WL	0.54
Comb-LF	0.55

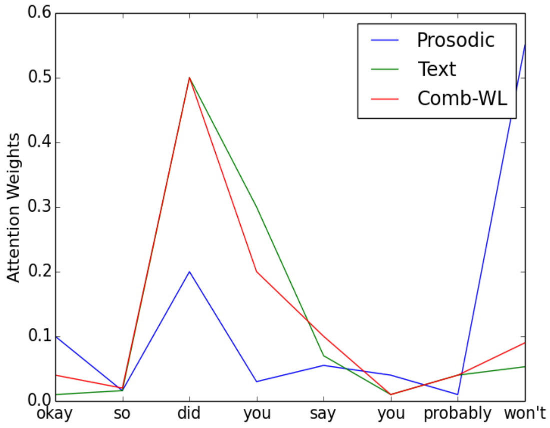

Figure 4: *Comparison of attention weights for one question sample (QES)*

These results validate our hypothesis that as prosodic features (pitch, loudness and jitter) are continuous values, what we essentially measure is the variation in them as we pass over words. When the utterance has very few time stamps (less words), the model with prosodic information performs badly as it is not able to cover the variation in them.

6. Conclusions and Future Work

In this paper, we demonstrated that using prosodic features in addition to lexical features aid in the prediction of certain utterance-level behaviors in psychotherapy sessions. We employed bi-directional LSTMs with an attention mechanism with both word-level and utterance-level fusion of prosodic and lexical modalities. We also presented an analysis with examples of the types of words and prosodic patterns that are attended to by the attention mechanism. Additionally, we discussed how the length of utterances influences performance of the prosodic modality. Ablation experiments suggest our encoder architecture relies on variation between prosodic features over words; thus, we plan to investigate using discrete representation of prosodic features in the future. We also plan to use more complicated compositional models to represent word-level prosodic information instead of using just the mean and standard deviation.

7. Acknowledgements

This work was supported by the NIH.

8. References

[1] D. Can, P. G. Georgiou, D. C. Atkins, and S. S. Narayanan, "A case study: Detecting counselor reflections in psychotherapy for addictions using linguistic features," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[2] D. Can, D. C. Atkins, and S. S. Narayanan, "A dialog act tagging approach to behavioral coding: A case study of addiction counseling conversations," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[3] B. Xiao, J. Gibson, D. Can, Z. E. Imel, D. C. Atkins, P. Georgiou, and S. S. Narayanan, "Behavioral coding of therapist language in addiction counseling using recurrent neural networks," in *Proceedings of Interspeech*, 2016.

[4] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[5] W. R. Miller, T. B. Moyers, D. Ernst, and P. Amrhein, "Manual for the motivational interviewing skill code (misc)," *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico*, 2003.

[6] B. Xiao, Z. E. Imel, P. G. Georgiou, D. C. Atkins, and S. S. Narayanan, "" rate my therapist": Automated detection of empathy in drug and alcohol counseling via speech and language processing," *PloS one*, vol. 10, no. 12, p. e0143055, 2015.

[7] B. Xiao, Z. E. Imel, D. Atkins, P. Georgiou, and S. S. Narayanan, "Analyzing speech rate entrainment and its relation to therapist empathy in drug addiction counseling," in *Proceedings of Interspeech*, sep 2015.

[8] B. Xiao, D. Bone, M. Van Segbroeck, Z. E. Imel, D. Atkins, P. Georgiou, and S. Narayanan, "Modeling therapist empathy through prosody in drug addiction counseling," in *Proceedings of Interspeech*, sep 2014.

[9] M. Tanana, K. A. Hallgren, Z. E. Imel, D. C. Atkins, and V. Srikumar, "A comparison of natural language processing methods for automated coding of motivational interviewing," *Journal of substance abuse treatment*, vol. 65, pp. 43–50, 2016.

[10] V. Pérez-Rosas, R. Mihalcea, K. Resnicow, S. Singh, L. Ann, K. J. Goggin, and D. Catley, "Predicting counselor behaviors in motivational interviewing encounters," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, vol. 1, 2017, pp. 1128–1137.

[11] J. Gibson, D. Can, P. Georgiou, D. Atkins, and S. Narayanan, "Attention networks for modeling behavior in addiction counseling," in *In Proceedings of Interspeech*, August 2017.

[12] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[13] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1480–1489.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 6000–6010.

[15] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[16] T. Tran, S. Toshniwal, M. Bansal, K. Gimpel, K. Livescu, and M. Ostendorf, "Joint modeling of text and acoustic-prosodic cues for neural parsing," *arXiv preprint arXiv:1704.07287*, 2017.

[17] Z. Yu, S. Scherer, D. Devault, J. Gratch, G. Stratou, L.-P. Morency, and J. Cassell, "Multimodal prediction of psychological disorders: Learning verbal and nonverbal commonalities in adjacency pairs," in *Semdial 2013 DialDam: Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue*, 2013, pp. 160–169.

[18] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2130–2134.

[19] D. C. Atkins, M. Steyvers, Z. E. Imel, and P. Smyth, "Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification," *Implementation Science*, vol. 9, no. 1, p. 49, 2014.

[20] J. S. Baer, E. A. Wells, D. B. Rosengren, B. Hartzler, B. Beadnell, and C. Dunn, "Agency context and tailored training in technology transfer: A pilot evaluation of motivational interviewing training for community counselors," *Journal of Substance Abuse Treatment*, vol. 37, no. 2, pp. 191–202, 2009.

[21] B. Xiao, C. W. Huang, Z. E. Imel, D. C. Atkins, P. Georgiou, and S. S. Narayanan, "A technology prototype system for rating therapist empathy from audio recordings in addiction counseling," *PeerJ Computer Science*, vol. 2, no. e59, apr 2016.

[22] T. Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis," *PloS one*, vol. 10, no. 12, p. e0144610, 2015.

[23] P. Boersma, "Praat: doing phonetics by computer," *http://www.praat. org/*, 2006.

[24] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," 2001.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[26] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," *arXiv preprint arXiv:1409.2329*, 2014.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.