TOWARDS AUTOMATING THE ASSESSMENT OF TEACHERS' TEACHING EMOTIONS IN PRESCHOOL CLASSROOMS

Mingyue Huo[1], Yan Tang[1,2]

[1]Department of Linguistics, University of Illinois Urbana-Champaign, USA
[2]Beckman Institute for Advanced Science and Technology, USA
{mhuo5, yty}@illinois.edu

ABSTRACT

In-person assessment of teaching emotion as part of a teacher's working performance evaluation is onerous, and can be intrusive to some teachers in teaching. This study shows preliminary attempts at solving this issue using a machine learning approach with bi-modal artificial neural networks, which made predictions by combining acoustic and textual features extracted from preschool teachers' spontaneous speech captured in real teaching scenarios. In a binary classification task identifying the emotions as "positive" and "negative", the prediction achieved an accuracy of 79.0%, an F1-score of 83.1% in cross-validation, and an accuracy of 68.4%, an F1-score of 67.2% in leave-one-subject-out validation. When further adding speech samples identified as "neutral" to the task, a decreased accuracy of 52.4% and F1-score of 53.6% were received in cross-validation, confirming the difficulties in labelling this type of naturalistic speech data even for human raters.

Keywords: teaching emotion, spontaneous speech, neural network, acoustic features

1. INTRODUCTION

Emotions are prominent aspects of human experiences and play a defining role in social interactions. Ratings of human emotion have been used in various developmental [1], educational [2], and health research studies. These ratings involve observing, interpreting, and coding expressed emotions, which are explicitly or implicitly expressed during interactions between teachers and students in the classroom and invoked in the raters. Emotion ratings are also used in policy decision-making, such as the Emotional Support domain of the Classroom Assessment Scoring System [3], which is a determinant of funding allocations for the largest federally funded preschool program, Head Start, in the United States. However, there is ample evidence that raters frequently disagree [4, 5]. A possible reason for such disagreement is cultural and language diversity. A cross-language and cross-cultural study of a listener's capability to identify emotion using acoustic cues [6] showed that both universal and language-specific cues of emotion expressions affect emotion identification in an unfamiliar language and culture. However, established rating systems have not successfully addressed this issue.

Systems and algorithms using audio information have been developed to assist in identifying and recognising human effects efficiently, such as emotions, sincerity, irony and deception from linguistic and paralinguistic information extracted from physical speech signals [7, 8, 9]. For speech emotion classification, early studies have used Gaussian mixture models and Hidden Markov models to learn the acoustic feature distribution, and used Bayesian classifier and support vector machine (SVM) classifier to categorise emotions [10, 11]. In recent years, deep neural networks have demonstrated better performance on the acted database IEMOCAP, and end-to-end models combining Convolutional Neural Networks with long short-term memory have been reported to outperform the traditional approaches on the naturally elicited database RECOLA [12, 13]. Despite many studies on emotion classification, there is limited work on detecting teachers' teaching emotions, especially from spontaneous speech captured in classrooms. With a Recurrent Neural Network, Liang et al. [14] achieved an accuracy of 85.3% in a teaching emotion recognition task on a Mandarin speech corpus, for which teaching scenarios were simulated using professional voice actors under specific instructions. Cen et al. [15] trained an SVM model for detecting the emotions of students during online learning. This model combined the probabilities of finer emotion categories to predict "positive" ("happy") and "negative" ("sad" and "anger") emotions in one speech utterance.

Developing a machine-aided rating system for teachers' teaching emotions is significantly meaningful, because it may provide more objective and consistent assessment for teachers from different cultural and racial backgrounds, reducing biased feedback and subsequent unfair treatment due to insular individual opinions. It could alleviate the mental stress on some teachers, who are sensitive to their working ambience and audience, due to the presence of their performance evaluator(s). When relying on computer coding of teachers' basic emotions, raters can focus on higher-level constructs (e.g. responding and instructing).

In this study, we examine the feasibility of applying ML techniques to recognising teaching emotion from teachers' *spontaneous* speech as an essential step towards investigating how culturally-specific acoustic signatures are associated with teaching emotions. As opposed to elicited emotional speech used in many previous studies, we work on natural utterances captured when teachers were conducting ordinary teaching activities in preschool classrooms in the United States, in order to obtain an ecologically-valid assessment of machine performance. Instead of proposing novel ML algorithms or using sophisticated ML techniques to address the problem, we

focus on obtaining a benchmark to inform future studies.

2. METHOD

2.1. Speech data

The speech recordings used in this study were drawn from the EMOtion TEaching Rating Scale (EMOTERS) dataset [16]. This corpus is primarily used to assess teachers' emotion-focused teaching practices, which support children's social and emotional development and school readiness. It contains 1,606 10-minute (approx. 268 hours) realistic spontaneous speech recordings with human-coded emotion labels. The original recordings include both video and audio recordings; only audio signals were extracted for analyses in this study. Due to the original audio being saved in MP3 format, some acoustic information had no longer been preserved in the files. Constrained by the data sharing protocol, out of the 1606 recordings in the original dataset, we could only access 31 recordings (approximately 5 hours) from 11 teachers (9 females and 2 males) for this study.

During data preparation, the over 5-hour recordings were segmented into small excerpts based on the teacher's conversational turns, as most speech emotion recognition tasks are developed and evaluated on utterance-level data [17]. This process resulted in 1570 speech excerpts in total, among which 540 excerpts were used as the samples for the subsequent experiments. Other excerpts were discarded due to the target voice being mixed with other unintended sounds (e.g. children's voices and background noise) to different degrees. All the excerpts were primarily coded into "positive" (POS) or "negative" (NEG) individually by three raters: two native American English speakers and one Chinese English speaker. If the emotion in a sample was not obvious to the rater, this sample was labelled as "neutral" (NEU). Of the 540 samples, 261 (48%) were labelled consistently across the raters; 134 (25%) were labelled the same by any two raters, while there was no consensus at all (i.e. the three raters had totally different opinions) for 145 samples (27%). The final label for a sample was the one rated by the majority of the raters (i.e. rated the same by at least two raters); the 145 samples without agreement were merged to NEU, leading to 206, 147 and 187 samples for POS, NEG and NEU, respectively as in Table 1. The labelling results suggested the human raters were indecisive between POS and NEG for over 1/3 of the samples.

Table 1: Teacher sex and rated emotion ("POS", "NEG", "NEU") distribution of the samples

Teacher	Gender	No. of Samples	**POS**	**NEG**	**NEU**
1	Female	15	9	2	4
2	Female	100	50	17	33
3	Female	29	15	4	10
4	Female	85	34	23	28
6	Female	49	18	14	17
5	Female	27	6	11	10
7	Female	79	27	10	42
8	Female	54	21	15	18
9	Female	18	12	4	2
10	Male	21	8	6	7
11	Male	63	6	41	16
total:		540	206	147	187

2.2. Feature preparation

Previous studies [18, 19, 20] suggested that humans can encode and show their emotions using both linguistic (via word choice) and acoustic (via vocal configuration) means. To exploit both acoustic and textual properties from teachers' vocalisation, acoustic measurements and bag-of-words representation were extracted from the speech samples.

There is a wide range of acoustic feature sets available for speech-related identification and analysis tasks, e.g. eGeMAPS [21], ComParE [9], YAAFE [22]. Most of these feature sets calculate different dimensions of low-level temporal and spectral descriptors (LLDs) from 10-20 ms speech frames. A non-exhaustive list of the LLDs is shown below:

- Energy: intensity, loudness
- Pitch: F0, voicing probability, F0 contour
- Voice quality: jitter, shimmer and logarithmic harmonic-to-noise ratio
- Spectral shape: Mel-Frequency Cepstral Coefficients, Perceptual Linear Predictive coefficients.
- Temporal dynamics: 1st-order differential and 2nd-order acceleration coefficients

High-level statistics functions are then derived based on the LLDs for the global utterance, including max, min, mean, the slope of linear approximation, skewness, kurtosis and more. We chose the ComParE feature set, which was designed for the Computational Paralinguistics Challenge and has been used in many speech emotion recognition tasks, such as [7]. Therefore, it was considered a reasonable starting point to obtain a baseline performance on the current speech data. Since ComParE extracts many possible features for a general purpose, a total of 6373 acoustic features was calculated from each speech sample. We assumed that many features in the ComParE set might overlap in terms of useful information they provide for classification tasks. Initial experiments suggested that reducing the feature dimensionality from 6737 to 162 explaining 90% of the variance in the data using Principal Component Analysis led to a similar machine performance as when all 6373 features were used. Therefore, only the first 162 principal components were supplied as acoustic features to the downstream ML algorithm to expedite training and prediction.

As for textual features, the transcriptions of the 540 samples were processed into 2447 unique words. The number was further reduced to 378 by removing the low-frequency words (which occurred only once in the dataset) to avoid noise interference. As a result, the textual model took the 378-dimension bag-of-words representation as its input features.

2.3. Model Structure and evaluation

We trained four uni-modal models using acoustic features or textual features separately to construct a baseline for

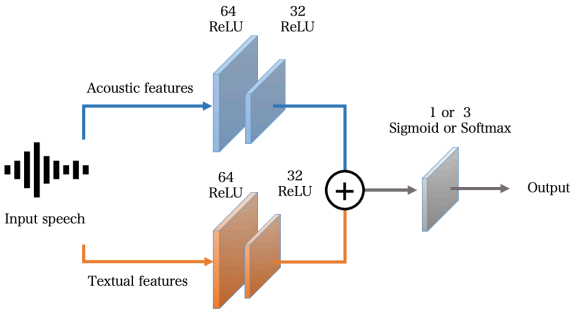

Figure 1: The structure of the bi-modal ANN model: acoustic and textual features were fed into two linear layers separately, then the outputs were concatenated before being fed into the output layer. The number of neurons in each layer is noted.

comparison. The first two classifiers were trained by a kernelised SVM with the radial basis function. The other models were trained as standard feedforward Artificial Neural Networks (ANN) with two hidden layers of 64 and 32 neurons towards the output layer. The Rectified Linear Unit (ReLU) was used as the activation function at each hidden layer, and a sigmoid or softmax function was used for the output layer depending on whether the task was binary or three-class classification. For a given ML algorithm, all the model specifications were the same for both the acoustic and textural features, except that the input size of the former was 162 (principal components) while that of the latter was 378 (words).

A bi-modal ANN was further trained to combine the textural and acoustic information. As shown in Figure 1, the 162 acoustic features were fed through two hidden layers with ReLU as the activation function, resulting in a 32-dimension vector for each sample as the output of the second hidden layer. Meanwhile, the textual features were fed through another two hidden layers, also resulting in a 32-dimension vector. Then, the vectors from the two modalities were concatenated and fed into the output layer. We expected the model to be informed of more information by combining the two kinds of features, leading to better predictive performance than the uni-modal models. To obtain a benchmark performance on this data set, no hyperparameter-tuning was performed on the models; all model parameters were empirically chosen.

Having observed the difficulties the human raters encountered during the rating process, we evaluated the model performance in two tasks: a binary (POS vs NEG) classification and a three-class classification including all three emotional categories (POS, NEG and NEU). Model performance was evaluated as accuracy and F1-score. Since the dataset was relatively small, a 5-fold cross-validation (CV, teacher-dependent) was employed to assess the overall predictive power of each model, along with leave-one-subject-out (LOSO, teacher-independent) CV for testing the model generalisability across teachers. During LOSO CV, the model each time was trained on the samples from 10 teachers and the samples from the remaining teacher were reserved for testing.

3. RESULTS

3.1. Uni-modal baselines

Table 2 presents the average performance of the 5-fold CV on the uni-modal models trained using SVM and ANN with different types of features in the binary classification task. Overall, the ANN models outperformed SVM models in both modalities, and the models trained using the acoustic features outperformed the models trained using the textural features.

Table 3 shows the performance of LOSO CV of the uni-modal models trained on acoustic or textural features. For the 11 teachers, the number of samples considerably varied across individuals from 11 to 57, with an average of 31 samples. Since the dataset was created from a spontaneous speech corpus, the imbalanced distributions in sample number and emotion class across teachers were inevitable at the preliminary stage. Compared to the performance in the early teacher-dependent evaluation, both the model using the acoustic features and the model using textual features were seen to decrease by approximately 10.6 and 12.2 percentage points (ppts) respectively in accuracy, and 15.4 and 14.6 ppts in F1-score. Nevertheless, the acoustic model (64.3%) still exhibited a somewhat more robust performance than the textural model (60.7%) in terms of accuracy, but otherwise similar measured as F-score. Since the sample distribution was imbalanced, the weighted average scores (weighted by the sample number for each teacher) are also shown in Table 3.

3.2. Bi-modal model

As illustrated in Figure 1, the bi-model uses both the acoustic features and textural features for making predictions. The last row of Table 2 shows the CV performance of the bi-modal ANN model when identifying POS and NEG samples in the teacher-dependent task. As anticipated, the bi-modal ANN achieved a higher accuracy of 79% and F1-score of 83.1% than the best uni-modal model, i.e. "ANN-acoustic" with an accuracy of 74.9% and F1-score of 79.5%. The last two columns of Table 3 display the LOSO CV performance of the bi-modal ANN model. While teacher-independent models exhibited decreased performance compared to the uni-modal models, the bi-modal model still outperformed the textual and acoustic models in weighted accuracy, and showed a better balance between the two performance metrics, suggesting a better generalisability than the uni-modal models.

Table 2: Uni-modal and bi-modal model accuracy and F1-score on binary classification (POS vs NEG). Metrics in percentage are calculated across 5-fold cross-validation

Model	Accuracy (%)	F1-score (%)
SVM-text	67.3	76.9
SVM-acoustic	70.4	78.8
ANN-text	72.9	78.1
ANN-acoustic	74.9	79.5
Bi-modal ANN	**79.0**	**83.1**

Table 3: Model accuracy and F1-score in LOSO CV for binary classification (POS vs NEG). Both mean performance across teachers and mean further weighted by the number of samples are also presented. Teachers marked by "*" are males.

Teacher	No. of Samples	ANN-acoustic		ANN-text		**Bi-modal ANN**	
		Acc. (%)	F1 (%)	Acc. (%)	F1 (%)	Acc. (%)	F1 (%)
1	11	63.6	70.2	54.5	66.7	72.7	80.1
2	67	61.3	66.2	64.5	68.4	69.4	74.0
3	19	63.2	53.0	57.9	56.4	68.4	55.9
4	57	64.9	72.3	56.1	57.4	64.9	61.4
5	32	75.0	80.3	62.5	71.3	78.1	77.3
6	17	76.5	74.0	64.7	69.5	70.6	76.8
7	37	71.9	68.7	68.8	60.0	78.1	67.7
8	36	47.2	61.1	52.8	69.0	50.0	56.9
9	16	62.5	65.0	75.0	79.9	81.3	74.0
10*	14	64.3	70.6	64.3	71.0	57.1	66.7
11*	47	57.4	24.0	46.8	29.4	61.7	48.6
Mean	31	**64.3**	**64.1**	**60.7**	**63.5**	**68.4**	**67.2**
Weighted mean		**63.5**	**62.6**	**59.8**	**60.9**	**67.7**	**65.6**

3.3. Three-class classification

When further adding speech samples identified as neutral (NEU) to the classification task, the accuracy of the acoustic and textual uni-modal ANNs, and bi-model ANN decreased by 27.9, 24.4 and 25.4 ppts; the F1 shrunk by 33.1, 29.0 and 29.6 ppts, respectively. The predictive power of the textual uni-modal ANN appeared to deteriorate more severely than the acoustic-only ANN. Despite the lower performance compared to the binary classification, the bi-modal ANN held its marginal lead in the three-class classification over the uni-modal models. Figure 2 compares the 5-fold teacher-dependent CV performance of all models tested above. Unless specified as "3-class", the model is for the binary classification task. As shown in the figure, the bi-modal ANN outperformed the uni-modal models in both the binary and the three-class classification tasks.

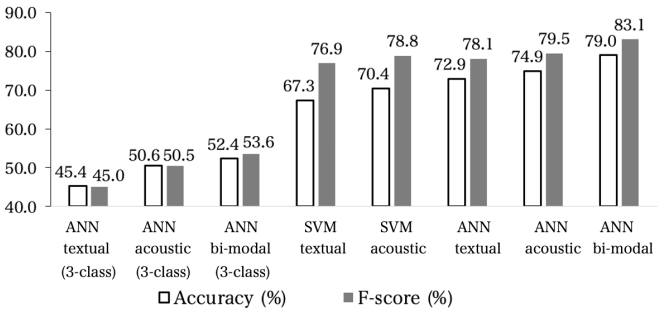

Figure 2: Teacher-dependent CV comparison among models

4. DISCUSSION AND CONCLUSION

This study used a bi-modal ANN that combines acoustic and textual features extracted from spontaneous speech to recognise teachers' teaching emotions in preschool classrooms in the United States. In the binary classification task, the bi-modal ANN achieved an accuracy of 79.0% and an F1-score of 83.1%. In all the tasks including three-class classification (POS, NEG and NEU) and LOSO CV, the bi-modal model outperformed the uni-modal models, which only used acoustic features or textural features. The more robust performance of the uni-modal model using acoustic features, especially in the three-class classification task, could suggest that some emotions are better encoded in teachers' vocalisation than in the vocabulary they use while talking. The reduced predictive power in the three-class classification task shows that the ML models employed in this study cannot account for the uncertainties in ratings due to human raters' different perceptions and cultural backgrounds. Further using fine-tuned models and more advanced ANN architectures could better capture the subtle nuances encoded in teachers' speech and can help differentiate emotions, which may explain the better machine performance reported in the literature. Though the performance falls behind Liang et al. [14], it should be noted that the database in their study was recorded by six male and six female teachers on five balanced emotions, resulting in a potentially higher probability for more robust classification. Other factors such as sample size, audio quality and features used may have also affected the machine performance in this study. Gent et al. [23] compared the machine performance in irony recognition from speech when using a set of refined acoustic features to using the entire CompPare set, and argued that a robust performance comes from the relevance of the features, not the quantity. Further detailed acoustic analyses on this type of spontaneous speech could help identify a set of more representative features for recognising teaching emotions.

As the preliminary study, this work tested the feasibility of using a machine learning-based approach to aid in the emotion-based assessment of preschool teachers' teaching performance. It can also provide a benchmark for automatic emotion detection from speakers' vocalisation acquired in realistic situations.

Acknowledgements We thank Drs Rachel Gordon, Katherine Zinsser and Timothy Curby for sharing the EMOTER database. We also thank Dr Chilin Shih for the early discussion of the idea and her support. We thank the reviewer for their thoughtful comments.

5. REFERENCES

[1] D. S. Messinger, "Positive and negative: Infant facial expressions and emotions current directions in psychological science," *Psychological Science*, vol. 11, pp. 1–6, 2002.

[2] S. A. Denham and H. H. Bassett, "Early childhood teachers' socialization of children's emotional competence," *Journal of Research in Innovative Teaching and Learning*, vol. 12, pp. 133–150, 2019.

[3] R. C. Pianta, K. M. La Paro, and B. K. Hamre, *Classroom Assessment Scoring System (CLASS) Pre-K version*, Paul H Brookes Publishing, Baltimore, MD, 2008.

[4] B. K. Hamre, R. C. Pianta, J. T. Downer, J. DeCoster, A. J. Mashburn, S. M. Jones, J. L. Brown, E. Cappella, M. Atkins, S. E. Rivers, M. A. Brackett, and A. Hamagami, "Teaching through interactions: Testing a developmental framework of teacher effectiveness in over 4,000 classrooms," *The Elementary School Journal*, vol. 113, pp. 461–487, 2013.

[5] R. A. Gordon, F. Peng, T. W. Curby, and K. M. Zinsser, "Using the many-facet Rasch model to improve observational quality measures: An introduction and application to measuring the teaching of emotion skills," *Early Childhood Research Quarterly*, vol. 55, pp. 149–164, 2021.

[6] K. R. Scherer, R. Banse, and H. G. Wallbott, "Emotion inferences from vocal expression correlate across languages and cultures," *Journal of Cross-cultural psychology*, vol. 32, no. 1, pp. 76–92, 2001.

[7] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen *et al.*, "The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. ISCA, 2020, pp. 2042–2046.

[8] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *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.

[9] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity and native language," in *Proceedings of Interspeech 2016*, 2016, pp. 2001–2005.

[10] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 2. Ieee, 2003, pp. II–1.

[11] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9-10, pp. 1162–1171, 2011.

[12] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Interspeech 2014*, 2014.

[13] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 5200–5204.

[14] L. Jie, Z. Xiaoyan, and Z. Zhaohui, "Speech emotion recognition of teachers in classroom teaching," in *2020 Chinese Control And Decision Conference (CCDC)*. IEEE, 2020, pp. 5045–5050.

[15] L. Cen, F. Wu, Z. L. Yu, and F. Hu, "A real-time speech emotion recognition system and its application in online learning," in *Emotions, technology, design, and learning*. Elsevier, 2016, pp. 27–46.

[16] K. Zinsser, T. Curby, and R. Gordon, "Emotion teaching rating scale," www.emoters.org, 2016.

[17] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review," *International Journal of Speech Technology*, vol. 21, no. 1, pp. 93–120, 2018.

[18] A. Ortony, G. L. Clore, and A. Collins, *The cognitive structure of emotions*. Cambridge university press, 1990.

[19] J. M. Wilce and J. M. Wilce, *Language and emotion*. Cambridge University Press, 2009, no. 25.

[20] A. Majid, "Current emotion research in the language sciences," *Emotion Review*, vol. 4, no. 4, pp. 432–443, 2012.

[21] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[22] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard, "Yaafe, an easy to use and efficient audio feature extraction software," in *Proceedings of the 11th International Society for Music Information Retrieval Conference*, Utrecht, The Netherlands, August 9-13 2010, pp. 441–446, http://ismir2010.ismir.net/proceedings/ismir2010-75.pdf.

[23] H. Gent, C. Adams, C. Shih, and Y. Tang, "Deep Learning for Prosody-Based Irony Classification in Spontaneous Speech," in *Proc. Interspeech 2022*, 2022, pp. 3993–3997.