INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Multiple Instance Learning for Inference of Child Attachment From Paralinguistic Aspects of Speech

Areej Buker[1], *Huda Alsofyani*[2], *Alessandro Vinciarelli*[2]

[1]CIS Department, CCSIT, Imam Abdulrahman Bin Faisal University, Dammam, SA
[2]University of Glasgow, Glasgow, UK

aanbuker@iau.edu.sa, h.alsofyani.1@research.gla.ac.uk,
Alessandro.Vinciarelli@glasgow.ac.uk

Abstract

Attachment is a psychological construct that accounts for the way children perceive their relationship with their caregivers. Depending on the attachment condition, a child can either be secure or insecure. Identifying as many insecure children as possible is important to mitigate the negative consequences of insecure attachment in adult life. For this reason, this article proposes an attachment recognition approach that, compared to other approaches, increases the Recall, the percentage of insecure children identified as such. The approach is based on Multiple Instance Learning, a body of methodologies dealing with data represented as "bags" of feature vectors. This is suitable for speech recordings because these are typically represented as vector sequences. The experiments involved 104 participants of age 5 to 9. The results show that insecure children can be identified with Recall up to 63.3% (accuracy up to 75%), an improvement with respect to most existing models.

Index Terms: Attachment, Multiple Instance Learning, Child Speech, Multiple Instance SVM

1. Introduction

A few weeks after they are born, children develop attachment relationships with their *caregivers*, the persons providing them with comfort and security (typically the parents) [1]. Attachment relationships are important because they shape the Internal Working Model (IWM) of social relationships, i.e., the way children tend to perceive social interactions for the rest of their life. When caregivers are responsive and supportive, the IWM is positive and children tend to think of themselves as worthy, while in the opposite case, children tend to think of themselves as unworthy and non-valuable [2, 3]. Children with a positive IWM are said to be *secure*, while the others are said to be *insecure*. These latter experience more frequently issues such as lower academic performance, anxiety, problems with body image, sleeping disorders, antisocial behaviour and suicidal tendencies [2, 4].

If identified early enough, insecure children can be helped to avoid the consequences of their attachment condition. Therefore, this article proposes an attachment recognition approach for children of age 5 to 9. The experiments involved 104 children undergoing the *Manchester Child Attachment Story Task* (MCAST) [5], the test psychiatrists use most commonly to identify insecure children. During the MCAST, children have to tell stories about the interaction of a boy with his caregiver (see Section 2 for more details). Besides listening to *what* children say, psychiatrists also pay attention to *how* children say it, i.e., to nonverbal aspects of speech. This is the reason why the proposed approach is based on paralanguage.

Previous speech-based attachment recognition approaches were presented in [6, 7, 8]. The approach proposed in [6] analyzed paralinguistics. The recognition methodology, based on stacked Recurrent Neural Networks, led to Recall up to 53.3% (F1 Scores up to 59.6%). The experiments in [7] take into account language and paralanguage in a multimodal approach (Recall up to 58.7%, F1 Score up to 66.7%). Other experiments were performed in [8], where a multimodal approach took into account paralanguage and facial expressions (Recall up to 56%, F1 Score up to 62.4%). In this case, paralanguage was shown to lead to higher performances. Another attachment-related issue was addressed in [9]. The work proposed an approach to discriminate children affected by Reactive Attachment Disorder [10] from the others. In this case, the best performance was achieved with turn-taking-related features. To the best of our knowledge, the only other two approaches aimed at attachment recognition were presented in [11], where they modelled the gestures of children with a Long Short-Term Memory Network and showed a Recall of 91% and an F1 Score of 77.5% (however, the result was obtained using data from the same child in both training and test set), and in [12], where the experiments involved adults who self-assessed their attachment condition by filling out questionnaires.

An issue common to all speech-based approaches above is that the Recall tends to be low (the highest value is 58.7% in [7]). This is a major issue in medical applications because low Recall means that people affected by an issue are not recognized as such and, therefore, do not get the medical attention they need. The key-assumption of this work is that the main reason behind the low Recall is that attachment does not manifest itself continuously in the data, but it rather tends to leave detectable traces only in a few, possibly sporadic *behaviour slices* [13]. The methodologies used in the works above either learn over time (e.g., Long Short-Term Memory Networks) or map the data into representations of increasingly higher semantic levels (e.g., Convolutional Neural Networks). As a consequence, they might "dilute" the slices, i.e., they might be more sensitive to the average characteristics of the data than to specific properties of a few data slices.

For the reasons above, this work proposes an approach based on Multiple Instance Learning (MIL), a body of methodologies designed to deal with data that can be represented with multiple, possibly dependent feature vectors (the *instances*). These are separated into different "bags" that can be classified as positive when the number of positive instances goes above an appropriate, possibly low threshold. This means that a small number of slices compatible with insecure attachment, as long as the number is above an appropriate threshold, is sufficient to identify a child as insecure.

The experiments involved 104 children of age between 5 and 9. The results show that the proposed approach does im-

10.21437/Interspeech.2023-1200

Table 1: *Children distribution across stories based on their attachment style.*

Story	Secure	Insecure	Total
Breakfast	57	43	100
Nightmare	59	45	104
Tummyache	58	44	102
Hopscotch	59	44	103
Shop. Mall	59	44	103
Overall	59	45	

prove the Recall of previous approaches for 80% of the MCAST stories, thus suggesting that the key-assumption underlying this work holds. In particular, the fraction of insecure children classified as secure was reduced by roughly 7% compared to the Recall results observed in previous paralinguistics-based work.

The rest of this article is organized as follows: Section 2 describes the data, Section 3 describes the approach, Section 4 reports on experiments and results, and the final Section 5 draws some conclusions.

2. The Data

The experiments involved 104 children randomly recruited from primary schools in Glasgow, including 57 females and 47 males. The children distribute as follows across the four primary school levels (P1 to P4): 19 children in P1 (age 5-6), 40 in P2 (age 6-7), 29 in P3 (age 7-8) and 16 in P4 (age 8-9). The children were involved only upon authorization of their parents and they were free to stop their participation at any moment. In addition, children and parents were allowed to inspect the data and destroy them, partially or totally, if they considered it necessary. The data collection was performed according to professional codes of conduct and ethical regulations of the British Psychological Society. To the best of our knowledge, there are no publicly available data concerning attachment in children.

Every child was recorded while participating in the MCAST [5], a narrative test based on five story stems corresponding to everyday interactions between children and caregivers (*Breakfast*, *Nightmare*, *Tummyache*, *Hopscotch* and *Shopping Mall*). During the administration, participants are told how the stories begin, in English, and then are asked to continue with the help of two dolls, one representing a child and the other the caregiver. The administration was performed with the help of the School Attachment Monitor (SAM) [11], an interactive system designed to deliver the MCAST automatically and to record the children while they undergo the test. A pool of professional attachment assessors analysed the resulting data, indicating whether a child was secure or insecure. The resulting corpus was used in several previous works [6, 7, 8, 9, 11].

In total, 56.7% of the children were assessed to be secure (see Table 1 for a breakdown distribution over the stories), in line with epidemiological observations showing that this percentage ranges between 36.6% and 75.6%, with an average around 60% [14]. In this respect, the attachment distribution of the participants appears to be representative of the general population. The total length of the recordings is 7 hours, one minute and 24 seconds, with an average length of 204.8 seconds per child.

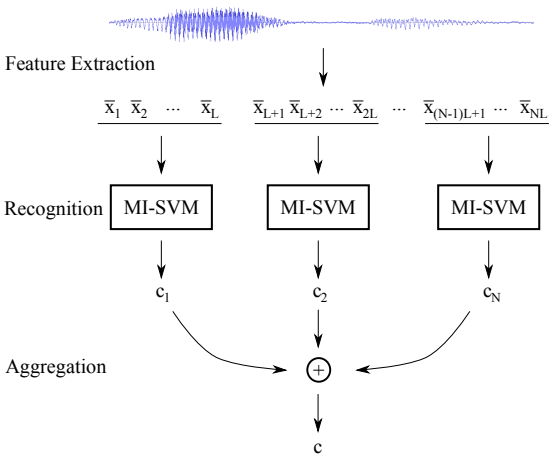

Figure 1: *The scheme shows the attachment recognition approach. The sequence of feature vectors is split into N bags, each including L elements. The expression c_k corresponds to the classification outcome of bag k, while the symbol $\oplus$ corresponds to the aggregation approach. The symbol c corresponds to the final outcome of the process.*

3. The Approach

The proposed approach includes three main steps, namely *feature extraction*, *recognition* and *aggregation* (see Figure 1). The first step uses OpenSmile [15] to convert the audio recordings into sequences of feature vectors extracted from non-overlapping windows spanning the entire duration of every recording, where the length of these windows was set to 33 ms. These choices were made to limit the number of differences between the approach proposed in this paper and [6]; the other work that focused solely on paralinguistics. Every vector includes Root Mean Square of the Energy (RMSE), 12 Mel-Frequency Cepstral Coefficients (MFCC), Zero-Crossing Rate (ZCR), Voicing Probability (VP) and Fundamental Frequency (F0). The difference between feature values extracted from two consecutive windows was added to capture temporal dynamics, thus leading to a dimensionality $D = 32$. The feature set was developed for the *Emotion Recognition Challenge* at Interspeech [16] and it was shown to account for a wide spectrum of social and psychological phenomena in speech.

At the end of the feature extraction process, every recording corresponds to a set of feature vectors $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$, where T depends on the length of the recording. The X sets are the input to the recognition step and this naturally leads to the application of *Multiple Instance Learning* (MIL), "a form of weakly supervised learning where training instances are arranged in sets, called bags, and a label is provided for the entire bag" [17]. In other words, bags are sets of feature vectors sharing a common label even if it is not possible to ensure that they all actually belong to the same class. In the experiments of this work, this means that all vectors extracted from a given recording are assigned to class *secure* or *insecure* depending on how the corresponding child was assessed (see Section 2).

The MIL approach selected for the experiments is the *Multiple-Instance Support Vector Machine* (MI-SVM) [18]. In the case of binary classification problems like the one addressed here, the key-idea of MI-SVM is that an SVM can be trained over bags by maximizing the margin of the "most positive" instance, i.e., the one with the highest distance from the hyper-

plane separating positive and negative instances. At the recognition step, this makes it possible to identify positive bags as those in which at least one instance, typically called the *witness* instance, is assigned to the positive class (*insecure* in this case).

Compared to Deep Learning methodologies, MIL approaches tend to require less data to be trained effectively. This is a major advantage in the case of this work because the number of children involved in the experiments is limited. However, the computational cost of training increases with the size of the bags L. For this reason, the sequence of the feature vectors is split into sub-sequences of length L and these act as bags that are classified individually (see Figure 1). This makes it necessary to add an aggregation step aimed at combining the classification outcomes corresponding to the multiple bags. The aggregation was performed through a majority vote (it assigns the recording to the class its bags are most frequently assigned to).

4. Experiments and Results

The experiments were performed according to a k-fold protocol ($k = 4$). The reason behind the choice of the k value is that it allows one to have a similar class distribution in all folds. The MI-SVM implementation used in the experiments is described in [19]. The approach involves two hyperparameters, the cost C of the linear kernel in the SVMs (no other kernels were tried) and the size L of the bags. Parameter C was set through nested cross-validation applied on data collected from *Breakfast*, with values 1, 25, 50, 75 and 100. The Breakfast story was chosen to fine-tune C because, unlike the other stories, it was not designed to cause distress. Therefore, traces of insecure attachment are expected to be limited. A value that can lead the algorithm to recognise limited traces in *Breakfast* should result in the creation of models that perform well when there are more traces generated from the distress-inducing stories.

The selection criterion was to choose a value that maximises Recall while maintaining high accuracy. As for parameter L, it was set to 200 because this is the largest possible value compatible with the computational resources at disposition (greater values of L tend to be more desirable, but require more computational power). On the other hand, such a value corresponds to a bag covering roughly 6.6 seconds, a length comparable with thins slices likely to account for attachment behaviour. In this respect, the value is the best possible tradeoff between the need to have a bag size large enough to benefit from MIL, but small enough to be used with the model without the need for a higher computational power.

All data belonging to a child were included in one fold to make the experiments person independent, meaning that the same child is never represented in both the training and test sets. This ensures that the approach detects attachment and does not simply recognize speakers. The final experiments were performed using a linear kernel SVM with the *cost* parameter C set to 50 and the bag size L set to 200. The feature vector sequences extracted from the data were segmented into non-overlapping subsequences of length L (see Figure 1). The vectors belonging to each subsequence were then used as bags. The length of the recordings belonging to a given story were capped to the length of the shortest recording for the same story. In this way the number of bags is the same for all recordings corresponding to a story (5 for Breakfast, 7 for Nightmare, 8 for Tummyache and Hopscotch, and 15 for Shopping Mall). Following the approach proposed in [18], the SVM was trained to maximize the margin of the "most positive" instance in every training bag (see Section 3 for more details). Since the algorithm does not involve a random initialization step, no repetitions were needed. Training the model took around 22 hours, on a MacBook Pro with the Apple M1 Pro chip, a 16-core GPU, and a 10-core CPU.

Table 2: *Attachment recognition performance. The results are reported in terms of Accuracy, Precision, Recall and F1 Score. Given that MI-SVM does not require a random initialization of its parameters, the performance metrics are not reported in terms of average and standard deviation across multiple repetitions of the experiments. Row "Random" reports the result of the random baseline classifier.*

Story	**Acc (%)**	**Prec (%)**	**Rec (%)**	**F1 (%)**
Breakfast	71.0	68.4	60.5	64.2
Nightmare	71.2	71.4	55.6	62.5
Tummyache	71.6	68.3	63.6	65.9
Hopscotch	72.8	70.0	63.6	66.7
Shop. Mall	67.0	67.9	43.2	52.8
All (SMV-M)	75.0	77.1	60.0	67.5
All (SMV-W)	67.3	65.7	51.1	57.5
All (SMV)	73.1	74.3	57.8	65.0
Random	51.0	43.0	43.0	43.0

In Table 2, the upper part shows the results for every story individually, while the lower part shows the performance achieved by using all data available for every child. There were three approaches to move from results obtained at the level of individual stories to results obtained at the level of all data. The first is a majority vote over individual story outcomes (SMV-M), the second one is a weighted majority vote over the story outcomes, where the weight is the fraction of bags assigned to a given class in every story (SMV-W), and the third one is a majority vote applied to all bags irrespectively of the story they belong to (SMV).

All approaches were compared with a random baseline classifier that assigns a child to class c with a probability equal to the prior $p(c)$ of c. The accuracy of such a classifier is $\alpha = \sum_{c \in \mathcal{C}} p(c)^2$, where $\mathcal{C}$ is the set of all classes (*secure* and *insecure* in the experiments of this work). Precision, Recall and F1 Score of the random classifier are all equal to the prior of the positive class (*insecure* in the experiments of this work). According to a binomial test, all approaches outperform the random classifier to a statistically significant extent ($p < 0.001$ in all cases after application of Bonferroni correction), except for the Recall of Shopping Mall.

The highest Recall values were obtained in correspondence of two stories, namely Tummyache and Hopscotch. This seems to suggest that such stories stimulate detectable attachment-relevant behaviours in a larger number of children. One possible reason is that the two stories involve the experience of physical pain and, therefore, they are more likely to induce mild stress in children. This, according to the theory underlying the MCAST, should make the behavioural differences between secure and insecure participants more evident [5]. Although the data is not linearly separable, the linear kernel yielded high results. One possible reason is that MI-SVM focuses on the most "extreme" instances of the positive class (insecure in this case), which are probably linearly separable from the rest of the instances.

Figure 2 shows a comparison between the results obtained with MI-SVM in this work and those obtained in [6] using stacked Recurrent Neural Networks (S-RNN) [20]. The main reason for performing this comparison is that this approach also

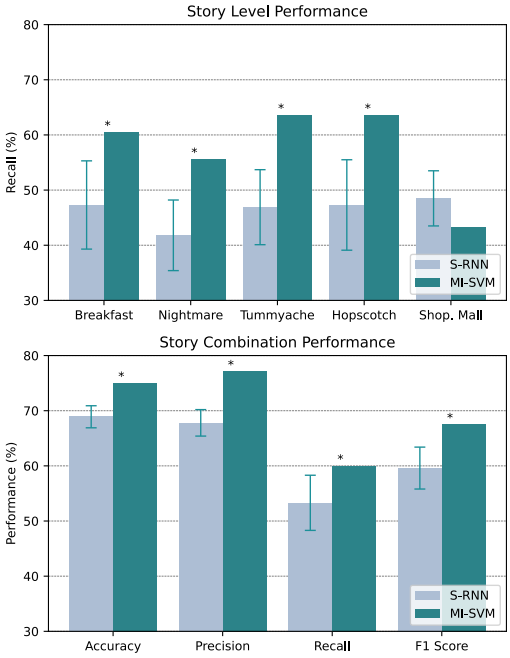

Figure 2: *The upper chart shows the comparison between S-RNNs and MI-SVM at the story level, and the lower chart shows the comparison between the two best approaches using all data. The asterisk accounts for statistically significant differences ($p < 0.05$ after Bonferroni correction). For S-RNNs the performance is presented in terms of average and standard deviation over $R = 10$ repetitions of the experiments. In the case of MI-SVM, there is no need to repeat the experiments because no random initialization is required.*

uses paralinguistics as a basis for attachment classification. At the story level, the results seem to suggest that MI-SVM tends to outperform stacked RNNs (see the upper chart of Figure 2). According to one-sample t-tests, the performance of MI-SVM is higher to a statistically significant extent ($p < 0.05$ after application of Bonferroni correction) for all stories except Shopping Mall. One possible explanation for the improvement of Recall is that as the S-RNN trains, earlier traces of insecure attachment are lost or diluted. Whereas the MI-SVM model utilises the most positive instances (insecure instances) in the training bags to train a classifier, and it needs no more than a few positive instances in a test bag to classify it as positive. Therefore, the insecure traces detected by the model are preserved. Furthermore, it is possible that children manifest attachment through speech patterns short enough to be captured with a few feature vectors (corresponding to a few analysis windows) rather than through long-term patterns (between 4 and 5 seconds) like those captured with S-RNNs.

The lower chart of Figure 2 shows the result of the same comparison above, but after moving from results obtained at the story level to results obtained over all data. In this case as well, MI-SVMs seem to outperform S-RNNs. Although the highest Recall value achieved by MI-SVM is 63.6%, the fusion of the stories reduced the score by 3.6%, a statistically non significant margin according to a binomial test. One possible explanation is that the positive classifications get diluted by the majority vote combination method, especially when a child leaves a few insecure attachment traces in the different stories. The majority vote method, although yielding the highest results, is not in line with what psychologists use to diagnose attachment using MCAST. In [5], if a child is "insecure" according to 33% of the stories (unlike the 50% used here), the child is diagnosed with insecure attachment. By using the same threshold for the data of this work the results are accuracy 73.1%, Precision 68.1%, Recall 71.1% and F1 Score 69.6%. This corresponds to an improvement of the Recall, but to a decrease in Precision.

5. Conclusions

The experiments of this work focused on attachment recognition based on Multiple Instance Learning. The key-result is that MI-SVM outperforms, in most cases, S-RNNs. This can be a major advantage because training MI-SVM is easier than training Neural Networks. In fact, there are only two hyperparameters in MI-SVM (cost C for the SVM and bags' size), while there are many for S-RNNs and NN-based models in general (size of the layers, learning rate, number of training epochs, activation functions, etc.). Furthermore, S-RNNs (and NN-based models in general) can change their performance significantly depending on the initialization of their parameters, a problem that MI-SVM does not have.

The main limitation of the approach is that the Recall does not improve when combining the results obtained over multiple stories. This probably happens because the MCAST is not designed to make an assessment according to a majority vote over stories, but to create as many chances as possible to elicit attachment relevant behaviours. In this respect, even if a child shows insecure attachment in one story only, the overall assessments is that the child is insecure. In this respect, one possible direction for future work is to avoid the classification of stories to rather target the recognition of insecure behaviours.

The question that remains open is why a model relatively basic as MI-SVM outperforms S-RNNs. One possible answer is that MI-SVM separates the data into bags, and looks for a set of the most positive instances to be used for classification. This ensures that informative features are always utilised by the model, unlike S-RNNs where the longer the model trains, the more the previously learned patterns become vague. Furthermore, the approach proposed breaks the data of each recording into sub-sequences of 200 instances. This means that MI-SVM uses only around 6.6 seconds of each recording to fit the model, which possibly limits the influence of noise in some of these bags, resulting in a more accurate classification.

The improved performance observed from training MI-SVM on smaller sub-sequences might also be explained by the *Thin Slices Theory* [13], a model stating that behavioural traces of social and psychological phenomena tend to take place over short time intervals. One of the main differences between the models compared in this work is that S-RNNs work over relatively long speech segments (between 4 and 5 seconds). While MI-SVM classifies every feature vector individually, and it is sufficient to find one positive instance to classify an entire bag as positive. Given that feature vectors are extracted from analysis windows of 33 ms, it is possible that they correspond to thin slices of attachment-relevant behaviour. In other words, it is possible that attachment manifests itself through short-term behavioural cues rather than through long-term behaviours. If this is the case, attachment recognition is easier for MI-SVM than for S-RNNs. Future work will focus on attention mechanisms as an alternative approach to identify attachment-relevant behavioural slices.

6. References

[1] J. Bowlby, "Attachment and loss: Retrospect and prospect," *American Journal of Orthopsychiatry*, vol. 52, pp. 664–678, 1982.

[2] M. Esposito, L. Parisi, B. Gallai, R. Marotta, A. D. Dona, S. M. Lavano, M. Roccella, and M. Carotenuto, "Attachment styles in children affected by migraine without aura," *Neuropsychiatric Disease and Treatment*, vol. 9, pp. 1513–1519, 2013.

[3] G. Bosmans and J. Borelli, "Attachment and the development of psychopathology: Introduction to the special issue," *Brain Sciences*, vol. 12, February 2022.

[4] J. A. Feeney, "Implications of attachment style for patterns of health and illness," *Child Care Health Development*, vol. 26, pp. 277–288, 2000.

[5] J. Green, C. Stanley, V. Smith, and R. Goldwyn, "A new method of evaluating attachment representations in young school-age children: The Manchester Child Attachment Story Task," *Attachment and Human Development*, vol. 2, pp. 48–70, 2000.

[6] H. Alsofyani and A. Vinciarelli, "Stacked recurrent neural networks for speech-based inference of attachment condition in school age children," in *Proceedings of Interspeech*, 2021, pp. 2491–2495.

[7] ——, "Attachment recognition in school-age children: A multimodal approach based on language and paralanguage analysis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 8172–8176.

[8] ——, "Attachment recognition in school age children based on automatic analysis of facial expressions and nonverbal vocal behaviour," in *Proceedings of the ACM International Conference on Multimodal Interactions*, 2021, pp. 221–228.

[9] A. Bîrlădeanu, H. Minnis, and A. Vinciarelli, "Automatic detection of reactive attachment disorder through turn-taking analysis in clinical child-caregiver sessions," in *Proceedings of Interspeech*. ISCA, September 2022, pp. 1407–1410.

[10] R. F. Hanson and E. G. Spratt, "Reactive attachment disorder: What we know about the disorder and implications for treatment," *Child Maltreatment*, vol. 5, no. 2, p. 137–145, May 2000.

[11] G. Roffo, D.-B. Vo, M. Tayarani, M. Rooksby, A. Sorrentino, S. Di Folco, H. Minnis, S. Brewster, and A. Vinciarelli, "Automating the administration and analysis of psychiatric tests: The case of attachment in school age children," in *Proceedings of CHI*, 2019, pp. 1–12.

[12] F. Parra, S. Scherer, Y. Benezeth, P. Tsvetanova, and S. Tereno, "Development and cross-cultural evaluation of a scoring algorithm for the biometric attachment test: Overcoming the challenges of multimodal fusion with" small data"," *IEEE Transactions on Affective Computing*, 2021.

[13] N. Ambady, F. Bernieri, and J. Richeson, "Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream," in *Advances in Experimental Social Psychology*. Elsevier, 2000, vol. 32, pp. 201–271.

[14] M. Schröder, J. Lüdtke, E. Fux, Y. Izat, M. Bolten, G. Gloger-Tippelt, G. Suess, and M. Schmid, "Attachment disorder and attachment theory–Two sides of one medal or two different coins?" *Comprehensive Psychiatry*, vol. 95, no. 152139, pp. 1–9, 2019.

[15] F. Eyben, M. Wöllmer, and B. Schuller, "opensmile - the munich versatile and fast open-source audio feature extractor," in *Proceedings of the ACM Multimedia (MM)*. ACM, October 2010, pp. 1459–1462.

[16] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 Emotion Challenge," in *Proceedings of Interspeech*, 2009.

[17] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple Instance Learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329–353, 2018.

[18] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in Neural Information Processing Systems*, vol. 15. MIT Press, 2002.

[19] G. Doran and S. Ray, "A theoretical and empirical analysis of support vector machine methods for multiple-instance classification," *Machine Learning*, vol. 97, no. 1–2, p. 79–102, Oct 2014.

[20] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, "How to construct deep recurrent neural networks," *arXiv preprint arXiv:1312.6026*, 2013.