Lecture video highlights detection from speech

Meishu Song, Ilhan Aslan, Emilia Parada-Cabaleiro, Zijiang Yang, Elisabeth André, Yoshiharu Yamamoto, Björn W. Schuller

Angaben zur Veröffentlichung / Publication details:

Song, Meishu, Ilhan Aslan, Emilia Parada-Cabaleiro, Zijiang Yang, Elisabeth André, Yoshiharu Yamamoto, and Björn W. Schuller. 2024. "Lecture video highlights detection from speech." In *2024 32nd European Signal Processing Conference (EUSIPCO), Lyon, France, 26-30 August 2024*, edited by Patrice Abry, Maria Sabrina Greco, Claire Goursaud, and Adrien Meynard, 361–65. Piscataway, NJ: IEEE. https://doi.org/10.23919/eusipco63174.2024.10715058.

Lecture Video Highlights Detection from Speech

Meishu Song[1,2], Ilhan Aslan[3], Emilia Parada-Cabaleiro[4], Zijiang Yang[1,2], Elisabeth André[5], Yoshiharu Yamamoto[2], Björn W. Schuller[1,6]
[1]*Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany*
[2]*Educational Physiology Laboratory, University of Tokyo, Japan*
[3]*Device Software Lab, Munich Research Center, Huawei Technologies, Germany*
[4]*Institute of Computational Perception, Johannes Kepler University Linz, Austria*
[5]*Group on Language, Audio, & Music, Imperial College London, UK*
[6]*Human Centered Multimedia Lab, University of Augsburg, Germany*
meishu@p-u.tokyo.ac.jp

***Abstract*—In interpersonal co-located and online teaching, lecturers highlight words and sentences in their speech in order to implicitly communicate that particular content is important. This social behaviour aimed to capture students' attention becomes crucial in distance learning, where the teacher's voice is an essential instrument to maximise students' attention. To enable intelligent systems, such as smart tutors, to understand and replicate this social behaviour, the ability to automatically recognise speech-based highlighting is needed. To this end, we introduce a public corpus for automatic detection of speech-based highlighting in learning context. With "Highlighting" we refer to the emphasised content, i.e., the important content which lecturers try to emphasise (highlight) by attracting the listeners attention. The dataset is derived from YouTube tutorial videos featuring 104 different English speakers who cover different disciplines. In sum, the dataset, which will be made freely available to the community. In addition, to establish an analysis for the corpus, we report on a series of experiments with the best results being achieved with a combination of a VGG net and transformer architectures. Our initial results of 78.2 % Accuracy and 78.8 % Unweighted Average Recall (UAR), encourage us to believe that this new dataset will facilitate progress in speech processing research for education.**

Index Terms: Transformer, VGG, Speech, Highlighting Content

I. Introduction

The availability of datasets is essential for the development of machine learning, deep learning and the related research fields, such as the automatic recognition of linguistic and paralinguistic information [1]–[3], intelligent educational systems [4] and intelligent health care systems [5] . The proliferation of speech processing techniques is not surprising considering that speech is embedded in many forms of shared media, such as social networks. The automatic analysis of information from the media is indeed an essential step to filter an always increasing amount of information, otherwise barely manageable. In particular, meaningfully labelling media content becomes especially important in order to promote contextual access in consumer applications and services. Various datasets have been collected and made available to enable the development of socially intelligent systems, which typically observe users and adapt content depending on their social signals [6].

Throughout the Covid-19 pandemic, the need for alternative forms of learning and teaching became eminent when families, students, and teachers struggled to implement efficient forms of (digital) distance-learning by themselves. While there is no lack of learning material available online, such as video lectures on YouTube and similar video sharing sites, these materials are rarely annotated in terms of highlighting, , the emphasis (given , by using a particular intonation) which implicitly indicates that a given content is particularly relevant. In fact, highlights are common in books and e-books (e. g., [7]), where single sentences of importance have been marked. Early research already found that teachers emphasise the key points of their learning material while teaching [8]. Content with highlight annotations would be particularly relevant to develop systems able to recognise and emulate this form of emphasis typical of teaching. For instance, such a technology would enable a smart tutor (already capable to recognise student engagement levels), to also estimate when it might be worthy to highlight particular content in order to request students' attention.

Inspired by related works on emphasis recognition [9] and automatic highlighting of video lectures [10], we presented ARLH (Automatically Recognising Lecture Highlighting), a novel corpus of speech content extracted from education resources, fully annotated according to highlighting information.

TABLE I
Comparison between available research in [9] and ours. Datascale(Data) in Segments, language, access, methods, and accuracy(Acc)%, are reported. Note, statistical here is Sta.

	Data	Language	Access	Methods	Acc
[9]	348	Chinese	Private	Sta	70.2
Ours	1670	English	Public	CNNs	78.2

Compared with the previous work [9], there are several differences. Firstly, we have bigger amount of data samples. Secondly, the languages in these corpus are different, one is English and the other is Chinese. In addition, our dataset can be freely access. Further, we utilize different mythologies for analysis as shown in Table I.

Since the outbreak of the Covid-19 pandemic, the fragility of the currently available distance-learning technology has become evident. We strongly believe that AI can offer powerful solutions for the improvement of such a technology. Hence, we make ARLH freely available to the research community, by this promoting further improvements in the e-learning domain and aiming at the development of a more inclusive education for the future generations.

II. ARLH Corpus

The ARLH (Automatically Recognising Lecture Highlighting) corpus is aimed at facilitating highlighting speech recognition in learning scenarios, thereby bridging the gap between the state of the English tutorial highlighting research and what is required in real tutoring systems. The ARLH Corpus(cf. Table II) is made up of 1670 audio clips (mean length 3.8 seconds, std dev 2.6 seconds, and a total duration of 1 hour and 46 minutes) retrieved from YouTube.

The content of ARLH is based on tutorials, , courses aimed to teach a variety of disciplines, including English language, Python programming, or Social Psychology amongst others. The database contains clear speech produced by 104 adult speakers (52 are female and 52 are male). On average, each speaker has produced 16 spoken utterances binary annotated: highlighted speech (, an utterances in which particular content was emphasised); and no-highlighted speech (, speech not highlighted). For each speaker, both highlighted and no-highlighted utterances were collected. For exemplary reasons, Figure 1 shows two samples of spectrogram and intensity from a female speaker with a highlighted utterance and no highlighted one. Due to the purpose for which the corpus was collected, ARLH contains only audio files; yet, the necessary information to retrieve the original videos is released along the corpus.

TABLE II
Speaker information for HIGHLIGHTED (Hi) and NO-HIGHLIGHTED (No-Hi) segments in the corpus. Overall duration per class, total number of utterances (# Utts), number of speakers (# Spkrs), and the mean and standard deviation (Std) of samples across speakers for each class, are reported.

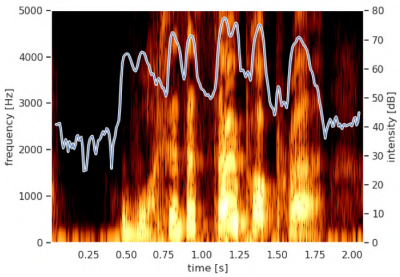
Classes	Duration	# Utts	#Spkrs	Mean	Std
Hi	1 h 6 min	799	52	4.95	2.96
No-Hi	40 min	871	52	2.75	1.70

A. Data Acquisition

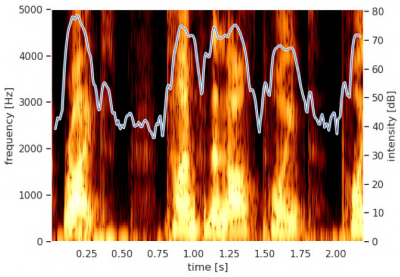
The first step to collect the corpus was to identify appropriate videos from which to retrieve the audio samples. To ensure a common standard across samples, this process was carried out manually, while following three guiding principles:

(i) Audio content presenting background music/noise, recorded at low quality, or presenting more than one speaker simultaneously was dismissed;

(ii) to prevent a linguistic bias, only recordings produced by English speakers were considered;



Highlighted Speech: “What is Soap Opera?”



No-Highlighted Speech: “Right, time for the last point.”

Fig. 1. Two samples of Spectrogram and Intensity from a female speaker in our corpus.

(iii) only content associated to a Creative Commons License was take into account.

All these criteria were applied subjectively by two auditors (authors of the presented work), and only samples targeted as valid by both of them were considered as part of the database. Note that all the considered videos were recorded in-the-wild, , in realistic scenarios without any predefined instruction/requirement. With this, we avoid any bias towards unrealistic recordings collected in-the-lab; hence, encouraging the development of robust models applicable in a variety of real-life settings. Similarly, a clear definition of ‘highlighted sample’ was stated according to two criteria:

Content level: Utterances containing important information, , tips, that the lecturer wants to transmit, and are typically emphasised by the use of particular words.

Acoustic level: Utterances emphasised by the use of a specific prosody, typically produced with a strong pronunciation and slow speed. After the videos were identified, these were retrieved through the YouTube API[1] and the audio layer was subsequently extracted in WAV format encoded in single channel 16k Hz 16 bit PCM format. In order to segment the selected videos in the target categories, highlighted and no-highlighted, we used the annotation tool NoVA [11]. Through this, we labelled each utterance’s beginning and end position and retrieved the corresponding timestamps. Afterwards, the validity of the gold standard was guaranteed via Cohen’s Kappa Coefficient(CKC) [12]. With CKC we measured the agreement over the collected binary annotations, applying it as a function for interval-scale ratings. CKC yielded 0.8

[1]https://yt5s.com/en11

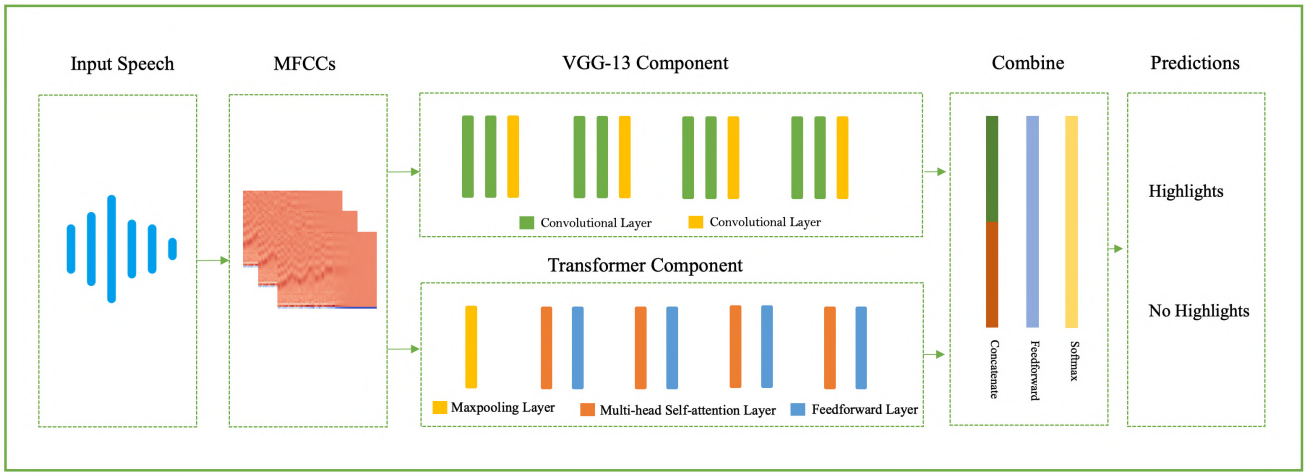


Fig. 2. A graphical overview of our framework.

for highlighted and no-highlighted, which indicates that the annotations are correlated. In total we collected 1670 audio samples balanced between the two categories: 799 of highlighted speech, 871 of no-highlighted speech.

III. Highlighting Recognition Framework

Applications of VGG net, a very deep Convolutional Neural Network (CNN) poposed in the ImageNet 2014 competition [13], have achieved excellent results in recent studies on speech recognition [14]. Despite the immense success, CNN-based methods have also shown lack of efficiency in capturing long time information [15].

Differently, transformers, a sequence-to-sequence architecture originally proposed for neural machine translation in Natural Language Processing (NLP), are characterised by a strong ability of long-range modelling, reason why they have been successfully used in speech processing, too [15]. In particular, a transformer has the advantage of encoding the input data as powerful features via the attention mechanism, which enables a good modelling of the global context. In contrast, CNNs are are particularly powerful in capturing samples' details [15].

Hence, to benefit from both and being able to learn sample details as well as global context, we propose to parallelised VGG net and transformers for the detection of highlighted speech. To the best our knowledge, this is the first time to parallelise VGG net and a transformer as a framework on a highlights recognition task.

The framework of the proposed model (cf. Figure 2) comprises four components:

(i) extraction of MFCCs features in grayscale as input layer;

(ii) VGG-13 net component;

(iii) 4 layers of transformer embeddings component;

(iv) combination of all embeddings into a softmax classifier aimed to make the final prediction.

A. Features

MFCCs are one of the most commonly used filterbank-based parameterisation methods for speech processing applications, such as Automatic Speech Recognition (ASR) [16], speaker verification/identification [17], and language identification [18]. MFCCs are derived from logMEL frequencies by computing the cepstrum of the melodic frequencies [19]. We extract MFCCs with the openSMILE toolkit [19] as features in our study. We gain the advantage of its low dimensionality and independence of the corruption across feature dimensions.

B. Data Partitioning

To carry out the experiments, we perform a Leave Eight Speakers Out (LESO) cross validation strategy, similarly with the work outlined in previous work [20], [21] to satisfy the speaker independence evaluation constraint. In this context, all the 64 speakers were divided into eight sets. In each set, the amount of female speakers and males speakers are four. One set was employed as fixed test set; the other seven sets were considered as training (Six subsets) and LESO validation set (One subset). During the training process, one subset was chosen as validation set while the other six sets were utilised as training set. Then, this process was repeated seven times until seven subsets were utilised as the LESO validation set.

C. Framework Components

For the VGG net component, we apply a classical VGG net and construct a multiple-layers CNN by using small 3x3 convolutional kernels with rectified linear unit (ReLU) and non-linear functions without pooling between these layers.

For the transformer component, we utilise a multihead self-attention mechanism, which is able to model relationships between all time steps in a sequence. To this end, we firstly maxpool the input MFCC map to the transformer block to reduce the number of parameters that the network needs to learn. Then, we apply four multi-head self-attention layers

of the transformer, by this enabling the network to look at multiple previous time steps when predicting the next.

Finally, we defined a feed-forward layer to combine the previously introduced components, , the VGG embeddings and the transformer embeddings. Note that the VGG embeddings were flattened into a 1D array of the transformer layers, we maxpool the input feature map and then squeeze it to remove the first channel dimension. Afterwards, the reduced input feature map was passed into transformer encoder layers. Finally, the VGG net embeddings and transformer embeddings are concatenated and sent to a linear softmax classifier to get the predictions.

D. Experimental Details

To perform the experimental baselines used for comparisons, besides the classical VGG net, we also consider ResNet [22], [23], , a 'classic' CNN architecture increasingly popular in recent days. ResNet was chosen due to its capability of capture morphological details effectively (as CNNs), but at the same time avoid the degradation problem (typical of deep neural networks) through its shortcut links between its layers.

For the VGG net, we apply VGG-11, VGG-13, VGG-16, and VGG-19 in our experiments. For ResNet, we apply ResNet-18, ResNet-34, and ResNet-50. Besides as baselines, all of these architectures are also combined with the same transformer component. For training details, we apply the following parameters: as learning rate 0.01, as batch size 16, 100 epochs, as optimizer SGD with momentum 0.8, and as loss function CrossEntropyLoss. All the models were developed on Pytorch 1.8.1 and trained on a single Nvidia RTX 3090 GPU.

IV. Results

As evaluation metrics to discuss the experimental results, we report Accuracy and Unweighted Average Recall (UAR). UAR was chosen due to the imbalanced distribution of samples across the two classes. In Table 2, the baseline results from both VGG net and ResNet, as well as the proposed models, all CNN architectures combined with transformer embeddings, are shown.

From the results, it can be seen that the methods with transformer embedding outperform all the baseline models without transformer – cf. every model on the upper part the one in the lower part (the same + transformer), in Table 2.

Our proposed VGG-13 with Transformer model achieves the best result: Accuracy $= 78.2\%$, UAR $= 78.8\%$ (cf. VGG-13 + Transformer); which as already indicated, outperforms the baseline model: Accuracy $= 77.9\%$, UAR $= 78.1\%$ (cf. VGG-13). This results makes the efficiency of parallelising VGG and transformers in a unique architecture evident.

We can also observe that an increment in models' depth goes along with improvements in the results. for VGG-11 vs VGG13: Accuracy $= 76.1\%$, UAR $= 76.2\%$ vs Accuracy $= 77.9\%$, UAR $= 78.1\%$; for VGG-11 + Transformer vs VGG13 + Transformer: Accuracy $= 77.3\%$, UAR $= 77.0\%$ vs Accuracy $= 78.2\%$, UAR $= 78.8\%$. However, for layers deeper than 16 and 19, this trend is inverted, which might be due to the fact that deeper models tend to overfitting. Differently, for ResNet architectures, the deeper models perform better than the others: Accuracy $= 76.5\%$, UAR $= 76.3\%$ (for ResNet-50); Accuracy $= 77.3\%$, UAR $= 77.4\%$ (for ResNet-50 + Transformer). This might indicate that ResNet architectures are more efficient to solve the vanishing gradient problem when networks are deeper, as suggested by the outcomes from previous research [24].

TABLE III
EVALUATION ACCURACY AND UNWEIGHTED AVERAGE RECALL (UAR) [%] FOR THE BASELINE AND PROPOSED MODELS, , COMBINATION OF VGG TYPES AND A TRANSFORMER. RESULTS ARE PRESENTED FOR THE LESO VALIDATION AND TEST SET; THE BEST RESULTS ON TEST SET ARE HIGHLIGHTED IN BOLD. NOTE, THE CHANCE LEVEL IN OUR EXPERIMENTS IS 50%.

Methods	Accuracy		UAR	
	LESO	Test	LESO	Test
Baselines				
VGG-11	76.4	76.1	76.5	76.2
VGG-13	78.0	77.9	78.2	78.1
VGG-16	76.8	76.3	76.7	76.8
VGG-19	75.8	75.4	75.8	75.7
ResNet-18	74.0	74.1	74.0	73.8
ResNet-34	73.0	73.3	72.7	72.9
ResNet-50	76.2	76.5	76.1	76.3
Proposed models				
VGG-11 + Transformer	77.2	77.3	77.0	77.0
VGG-13 + Transformer	78.0	**78.2**	78.3	**78.8**
VGG-16 + Transformer	77.6	77.4	77.2	77.3
VGG-19 + Transformer	76.2	76.2	76.3	76.7
ResNet-18 + Transformer	74.4	74.6	74.5	74.1
ResNet-34 + Transformer	74.0	74.1	73.7	73.8
ResNet-50 + Transformer	77.4	77.3	77.5	77.4

V. Conclusion

Along with the corpus, we presented a robust and efficient deep learning framework, based on parallelising VGG net and a transformer, for the automatic recognition of highlighting from speech. Our experimental results demonstrate that by combining VGG and transformer frameworks, a good accuracy can be achieved in the task of speech highlighting detection.

The presented framework could be successfully implemented in a variety of educational scenarios. For instance, since highlights can reactivate students' engagement when their attention may decay (something particularly easily happening during distance-learning sessions) a system that automatically retrieves highlights from pre-recorded lectures would enable students to easily retrieve and re-watch key contents.

For our next steps, one of our main priorities is to explore the facial features, pose, and gestures in our dataset. By this, we plan to further develop the presented work (focused on speech) towards a multi-modal framework.

VI. Acknowledgement

This research was partly supported by the Affective Computing HCI Innovation Research Lab between Huawei Technologies and University of Augsburg.

REFERENCES

[1] S. Amiriparian, A. Sokolov, I. Aslan, L. Christ, M. Gerczuk, T. Hübner, D. Lamanov, M. Milling, S. Ottl, I. Poduremennykh *et al.*, "On the impact of word error rate on acoustic-linguistic speech emotion recognition: An update for the deep learning era," *arXiv preprint arXiv:2104.10121*, 2021.

[2] J. Han, K. Qian, M. Song, Z. Yang, Z. Ren, S. Liu, J. Liu, H. Zheng, W. Ji, T. Koike *et al.*, "An early study on intelligent analysis of speech under covid-19: Severity, sleep quality, fatigue, and anxiety," *arXiv preprint arXiv:2005.00096*, 2020.

[3] A. Baird, M. Song, and B. Schuller, "Interaction with the soundscape: Exploring emotional audio generation for improved individual well-being," in *International Conference on Human-Computer Interaction*. Springer, 2020, pp. 229–242.

[4] D. Kučak, V. Juričić, and G. ambić, "Machine learning in education-a survey of current research trends." *Annals of DAAAM & Proceedings*, vol. 29, 2018.

[5] G. Deshpande and B. Schuller, "An overview on audio, signal, speech, & language processing for covid-19," *arXiv preprint arXiv:2005.08579*, 2020.

[6] I. Aslan, M. Schwalm, J. Baus, A. Krüger, and T. Schwartz, "Acquisition of spatial knowledge in location aware mobile pedestrian navigation systems," in *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services*, 2006, pp. 105–108.

[7] E. H. Chi, L. Hong, M. Gumbrecht, and S. K. Card, "Scenthighlights: highlighting conceptually-related sentences during reading," in *Proc. International Conference on Intelligent User Interfaces(IUI)*, San Diego,USA, 2005, pp. 272–274.

[8] "Teacher talk and meaning making in science classrooms: A vygotskian analysis and review," *Studies in Science Education*.

[9] X. Che, H. Yang, and C. Meinel, "Automatic online lecture highlighting based on multimedia analysis," *IEEE Transactions on Learning Technologies*, vol. 11, no. 1, pp. 27–40, 2017.

[10] F. R. Chen and M. Withgott, "The use of emphasis to automatically summarize a spoken discourse," in *Proc. International Conference on Acoustics, Speech, and Signal Processing(ICASSP)*, vol. 1, San Francisco, USA,, 1992, pp. 229–232.

[11] T. Baur, S. Clausen, A. Heimerl, F. Lingenfelser, W. Lutz, and E. André, "Nova: a tool for explanatory multimodal behavior analysis and its application to psychotherapy," in *Proc. International Conference on Multimedia Modeling(MMM)*, Daejeon, Korea, 2020, pp. 577–588.

[12] S. M. Vieira, U. Kaymak, and J. M. Sousa, "Cohen's kappa coefficient as a performance measure for feature selection," in *International Conference on Fuzzy Systems*. IEEE, 2010, pp. 1–8.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[14] W. Hartmann, R. Hsiao, T. Ng, J. Z. Ma, F. Keith, and M.-H. Siu, "Improved single system conversational telephone speech recognition with vgg bottleneck features." in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 112–116.

[15] P. Karpov, G. Godin, and I. V. Tetko, "Transformer-cnn: Swiss knife for qsar modeling and interpretation," *Journal of Cheminformatics*, vol. 12, no. 1, pp. 1–12, 2020.

[16] B. T. Meyer, S. V. Ravuri, M. R. Schädler, and N. Morgan, "Comparing different flavors of spectro-temporal features for asr," in *Proc. International Speech Communication Association*, Florence, Italy, 2011.

[17] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining mfcc and phase information," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 4, pp. 1085–1095, 2011.

[18] H. Mukherjee, S. M. Obaidullah, K. Santosh, S. Phadikar, and K. Roy, "A lazy learning-based language identification from speech using mfcc-2 features," *International Journal of Machine Learning and Cybernetics*, vol. 11, no. 1, pp. 1–14, 2020.

[19] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proc. Multimedia*, Florence, Italy, 2010, pp. 1459–1462.

[20] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang, and G. Rigoll, "Speaker independent speech emotion recognition by ensemble classification," in *Proc. International Conference on Multimedia and Expo(ICME)*, Amsterdam, Netherlands, 2005, pp. 864–867.

[21] M. Song, Z. Yang, A. Baird, E. Parada-Cabaleiro, Z. Zhang, Z. Zhao, and B. Schuller, "Audiovisual analysis for recognising frustration during game-play: introducing the multimodal game frustration database," in *Proc. IEEE International Conference on Affective Computing and Intelligent Interaction (ACII)*, Cambridge, United Kingdom, 2019, pp. 517–523.

[22] Z. Yang, S. Liu, M. Song, E. Parada-Cabaleiro, and B. W. Schuller, "Adventitious respiratory classification using attentive residual neural networks." in *Proc. INTERSPEECH*, 2020, pp. 2912–2916.

[23] M. Song, K. Qian, B. Chen, K. Okabayashi, E. Parada-Cabaleiro, Z. Yang, S. Liu, K. Togami, I. Hidaka, Y. Wang *et al.*, "Predicting group work performance from physical handwriting features in a smart english classroom," in *2021 5th International Conference on Digital Signal Processing*, 2021, pp. 140–145.

[24] F. Huang, J. Ash, J. Langford, and R. Schapire, "Learning deep resnet blocks sequentially using boosting theory," in *Proc. International Conference on Machine Learning*. PMLR, 2018, pp. 2058–2067.