Multimodal automatic speech fluency evaluation method for Putonghua Proficiency Test propositional speaking section

Jiajun Liu[1]*, Huazhen Meng*[2]*, Yunfei Shen*[1]*, Linna Zheng*[2]*, Aishan Wumaier*[2,*]

[1]School of Software, Xinjiang University, Urumqi, China
[2]School of Science and Engineering, Xinjiang University, Urumqi, China
{*liujiajun, menghuazhen, shenyunfei, zln*} @*stu.xju.edu.cn*, *Hasan*1479@*xju.edu.cn*

Abstract

The Putonghua Proficiency Test (Putonghua Shuiping Ceshi, PSC) is a valid form of speaking test in China. The propositional speaking section in PSC focuses on the speakers' ability to express ideas fluently and accurately without textual reference. However, unlike the other sections of the PSC, the propositional speaking section is still scored manually. Aiming at the problem of inefficiency, high cost, and subjectivity of manual scoring in the propositional speaking section, a multimodal method is proposed based on textual and acoustic modalities for automatic speech fluency evaluation. First, different neural networks are used to extract unimodal features. Then, cross-modal attention is applied to achieve multimodal fusion. Finally, fluency evaluation results are obtained by applying self-attention to reinforce the information with high contribution. The accuracy of the proposed method for automatic speech fluency evaluation is 81.67% on the self-built dataset. It shows that the textual and acoustic features used in this paper provide complementary information to improve the accuracy of fluency evaluation. And the fused features can be effectively applied to automatic speech fluency evaluation tasks.

Index Terms: Putonghua Shuiping Ceshi, automatic speech evaluation, multimodal, propositional speaking

1. Introduction

Speech fluency is a vital evaluation criterion of the Putonghua Proficiency Test (Putonghua Shuiping Ceshi, PSC), which reflects the speakers' oral expression ability. As an open-ended question, the propositional speaking section requires speakers to complete an impromptu presentation within a set time limit [1]. The propositional speaking section focuses on the speakers' ability to express ideas fluently and accurately. There are several criteria for evaluating propositional speaking. Accuracy in both pronunciation and grammar is regarded as the most common criterion in PSC. In contrast, the evaluation of speech fluency is more subjective. Factors affecting fluency in propositional speaking include speaking rate, articulation rate, average length of pause, phonation time ratio, repetition, correction, etc [2]. Due to the diversity of the evaluation criteria, manual scoring is still used. However, differences between evaluators lead to significant gaps in manual evaluations. And the manual scoring of propositional speaking is limited by cost, time, and space.

In this paper, we provide a new perspective for automatic speech fluency evaluation. Specifically, a 29.83-hour dataset is firstly constructed to address the problem of scarce resources for PSC propositional speaking section. Then, textual and acoustic features that influence the evaluation of speech fluency are explored. By analyzing the different expressions of fluency in textual and acoustic features, we propose a multimodal method that combines acoustic features with textual features through cross-modal attention to achieve audio-text fusion. Finally, we conduct experiments on the PSC propositional speaking dataset and compare the results with other models to prove the effectiveness of the proposed method. Experiments show that the accuracy of the proposed method is 81.67% on the self-built dataset.

2. Related work

Several scholars have researched automatic speech fluency evaluation. Early approaches to automatic speech fluency evaluation were based on acoustic features extraction and fluency scorer construction. Many studies used features such as speaking rate and pause frequency as input, then used simple machine learning models to build automatic evaluation models and trained linear regression or feedforward neural network models to fit manually annotated fluency scores. A statistical approach was presented based on the n-gram model to evaluate sentence fluency on a Chinese corpus in 2003 [3]. A logistic regression model was used by Bhat to score fluency on eight signal features and obtained a manual correlation coefficient of 0.6 [4]. Authors in [5] used seven regression models, such as Random Forest (RF) and Multi-Layer Perceptron (MLP), for automatic scoring experiments. Multiple linear regression [6] and ordinal regression neural networks [7] were used to predict the experts' mean fluency ratings in second language acquisition. Molholt [8] proposed a hybrid approach to evaluate spoken fluency by combining three metrics with the Support Vector Machine (SVM). Deng conducted a series of SVM classification experiments on a Japanese spontaneous speech corpus [9]. In addition, Deng evaluated speech fluency through Long-Short Term Memory (LSTM) on the same dataset [10]. Zhang proposed to use Bi-directional Long Short Term Memory (BLSTM) to capture better dynamic changes in phone-level fluency features [11]. In recent years, prosodic features have been used in fluency evaluation tasks. Authors in [12] combined novel prosodic and lexical features to compute the fluency score. Sammit presented an automated prosody classification for oral reading fluency with quadratic kappa loss and attentive X-Vectors [13]. Yang proposed a method for predicting the prosodic word and prosodic phrase boundaries to improve the Mandarin spoken fluency of international students [14].

In summary, most of the above methods treated the fluency evaluation as a classification or regression task and used deep learning methods to achieve satisfactory results. Few studies combined fluency with semantic, grammatical, and other text-related content for a comprehensive evaluation. In this work, we get inspiration from speech recognition and speech emotion classification models to learn useful features from the raw data. Inspired by Yoon [15], we use text data and audio signals si-

979-8-3503-9796-3/22/$31.00

multaneously to obtain more useful information for automatic speech fluency evaluation.

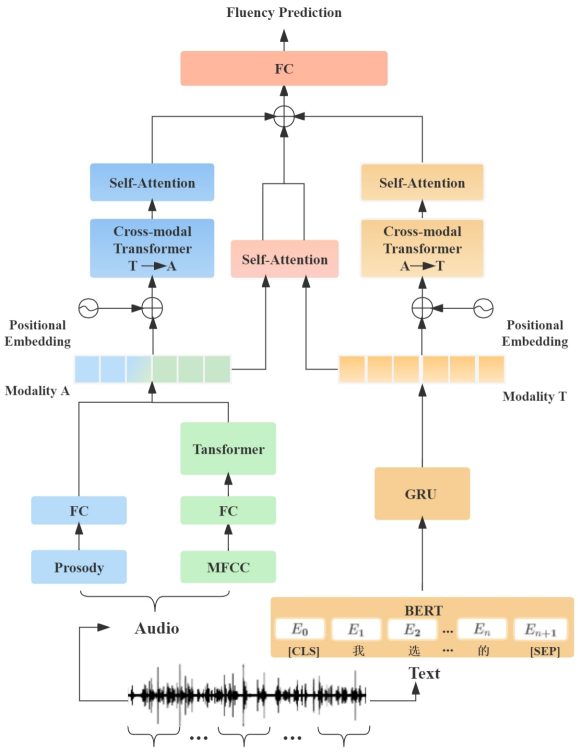

Figure 1: *The architecture of the model*

3. Method

The architecture of the model is shown in Figure 1. The whole model consists of an acoustic encoder, a textual encoder, and a multimodal fusion network. The measured data consists of textual and acoustic modalities, which are 2D tensors denoted by $X_t \in R^{T_t \times d_t}$ and $X_a \in R^{T_a \times d_a}$, where T_m and d_m represent sequence length and feature vector size of modality. For the rest of the paper, t and a represent textual and acoustic modalities, respectively. We encode all sequences from different modalities to guarantee a better fusion outcome.

3.1. Textual encoder

First, pre-trained Chinese BERT-base word embeddings are used to obtain word vectors from the text. Specifically, the head embedding of the last layer encoded by Bert [16] is used. We denote the text sequences as $X_t = \{x_0, x_1, ..., x_{n+1}\}$. Then, Gated Recurrent Unit (GRU) is applied to capture the contextual timing information.

$$\hat{X}_t = GRU(X_t) \in R^{T_t \times d_t} \tag{1}$$

Where $\hat{X}_t$ represents the text sequences obtained from the end-state of the GRU.

3.2. Acoustic encoder

The acoustic encoder is used to obtain sequential and statistical information from the acoustic modality by integrating data from the Mel Frequency Cepstral Coefficient (MFCC) and prosodic features.

A fully connected network is used to process prosodic sequences. For MFCC sequences, we first use a layer of the fully connected network to process MFCC features, after which we input the processed features to the transformer coding layer to obtain more useful MFCC sequences:

$$\begin{aligned} \hat{X}_p &= W_p X_p + b_p \in R^{T_p \times d_p} \\ \hat{X}_m &= Transformer(W_m X_m + b_m) \in R^{T_m \times d_m} \end{aligned} \tag{2}$$

Where W is the weight vectors, and b is the bias. $\hat{X}_m$ refers to the final stages of the MFCC output. Then the MFCC feature vector $\hat{X}_m$ is concatenated with another prosodic feature vector, $\hat{X}_p$, to generate a more informative vector representation of the signal.

$$\hat{X}_a = Concat(\hat{X}_p, \hat{X}_m) \tag{3}$$

Where $\hat{X}_a$ represents the fused acoustic feature vector.

3.3. Multimodal fusion network

In this paper, we construct a Transformer [17] based sequence encoding and use the ideas in [18, 19, 20] to encode multiple cross-modal attention blocks. Then self-attention is applied to reinforce the information from the raw and processed features.

First, we add the positional embedding (PE) with temporal information to $\hat{X}_a$ and $\hat{X}_t$.

$$\begin{aligned} Z_t &= \hat{X}_t + PE(T_{\hat{X}_t}, d) \\ Z_a &= \hat{X}_a + PE(T_{\hat{X}_a}, d) \end{aligned} \tag{4}$$

Then, textual and acoustic modalities are used as examples to introduce cross-modal attention in detail. Each attention module requires three inputs, Query (Q), Key (K), and Value (V). We define the three inputs as follows:

$$\begin{aligned} Q_t &= X_t W_{Q_t} \\ K_a &= X_a W_{K_a} \\ V_a &= X_a W_{V_a} \end{aligned} \tag{5}$$

Where $W_{\{Q_t, K_a, V_a\}} \in R^{T_{\{t,a,a\}} \times d_{\{k,k,k\}}}$ is weight. In the following, we use the example for passing acoustic (a) modality to textual (t) modality, which is denoted by "a $\rightarrow$ t". And the fused attention output vector Y from a to t can be represented as follows:

$$\begin{aligned} Y_t &= CM_{a \rightarrow t}(X_t, X_a) \\ &= softmax(\frac{Q_t K_a^T}{\sqrt{d_k}}) V_a \\ &= softmax(\frac{X_t W_{Q_t} W_{K_a}^T X_a^T}{\sqrt{d_k}}) X_a W_{V_a} \end{aligned} \tag{6}$$

The audio-to-text and text-to-audio features obtained through cross-modal attention to fuse different modalities are as follows:

$$\begin{aligned} Z_a^t &= CM_{a \rightarrow t}(Z_a, Z_t) \\ Z_t^a &= CM_{t \rightarrow a}(Z_t, Z_a) \end{aligned} \tag{7}$$

The self-attention is a variation of the attention mechanism, which relies less on external information and is better at capturing the internal relevance of data [21]. In this paper, we put the textual feature vector $\hat{X}_t$, the acoustic feature vector $\hat{X}_a$, the enhanced textual feature vector Z_a^t and the enhanced acoustic feature vector Z_t^a through the self-attention to reinforce the information, in order to obtain:

$$Z = \hat{X}_a + \hat{X}_t + Z_a^t + Z_t^a \tag{8}$$

Finally, the fluency level is obtained by a layer of the fully connected network. Where W and b denote the weight vector and bias, respectively. And d_{out} is the output dimensions of fluency levels.

$$Fluency = WZ + b \in R^{d_{out}} \tag{9}$$

4. Experiments

4.1. Dataset

We evaluate our model on the self-built PSC propositional speaking dataset. The dataset consists of 600 acoustic files from 205 speakers, including actual and simulated exam data. For all files, speakers must choose one of two topics and speak freely for three minutes.

We make the textual annotations and manual scoring annotations for each audio. As there is no reference text for the PSC propositional speaking section, Praat is used to proofread the audio and text files. The manual scoring annotations of the dataset are evaluated in seven dimensions according to the PSC propositional speaking scoring criteria. The main focus of this paper is fluency, which uses a five-point system with three levels, as shown in Table 1.

Table 1: *Definition of fluency levels*

Fluency levels	**Expression**
level1	Natural and coherent speech.
level2	Coherent speech in general, poor oral expression (a performance of recitation).
level3	The speech is incoherent, with many pauses, repetition, and stammering.

Three qualified and experienced evaluators are invited to score our data. The correlation coefficients for the three evaluators' overall scoring of the propositional speaking data are 88.20%, 88.60%, and 83.59%. The average correlation coefficient on fluency is 78.87%. It shows that the evaluators' scoring results can be used as labels for the training model. Table 2 shows the statistical results for the dataset.

4.2. Feature Extraction

First, each audio is split into several segments. Then MFCC and prosodic features are extracted as fluency features from these segments. The 39-dimensional MFCC feature vector extracted in this paper is set to have a frame size of 25 ms at a rate of 10 ms with the Hamming function. The prosodic features are composed of 120 features, which include the F0 frequency (F0), the Short Term Energy (STE), the Zero-Crossing Rate (ZCR), Sound Pressure Level (SPL), etc. The MFCC and prosodic features mentioned above are obtained using the OpenSMILE toolkit [22] and librosa [23].

Table 2: *Dataset statistics table*

Data information	**Results**
Total duration of speech (hours)	29.83
Total number of sentences	8682
Average duration of words each sentence (seconds)	11.53
Total number of words	303274
Total unique words	3143
Maximum words in a sentence	150
Minimum words in a sentence	1
Average words in a sentence	36
Ratio of actual exam data to simulated exam data	1:5
Degree of fluency ratio	1:1:1

4.3. Setup

Our model is implemented using PyTorch [24]. During training, we train each model with a batch size of 8 for 20 epochs and monitor its performance on the validation set. The mean square error (MSE) loss is applied to train the network using the Adam optimizer, with a learning rate of 1e-3. The input dimensions of the text, MFCC, and prosodic features are 768, 39, and 120, respectively. The two-layers GRUs with hidden layer dimensions are set to 128, 32. The fully connected network hidden layer dimension is set to 32. The dimension and number of heads in cross-modal attention are set to 16 and 8. All our experiments are conducted on NVIDIA TESLA V100 GPU.

4.4. Results and analysis

In this section, we first conduct comparative experiments on the unimodal models. Then, we explore features from different modalities that influence the fluency evaluation. Finally, we compare our model with other multimodal models and design an ablation study to evaluate the impact of different inputs and the use of modules. The mean absolute error (MAE), accuracy (Acc), F1 score (F1), and Pearson correlation (Corr) are used in experiments to evaluate model performance.

4.4.1. Unimodal Comparison Results

The performances of textual and acoustic features in speech fluency evaluation are verified under unimodal conditions. As shown in Table 3, Multinomial Naive Bayes (MNB), Logistic Regression (LR), SVM, and MLP are used to obtain unimodal comparison results. The experimental results show that the prosodic features are particularly well represented in fluency evaluation. And we note that textual features are informative in this paper. Then, we extract several prosodic features and analyze their performances at different levels of fluency. The results are shown in Figures 2, 3. Additionally, XGBoost [25] is used to rank all prosodic features. In Figure 4, we see that F0 and STE make significant contributions to the study of fluency.

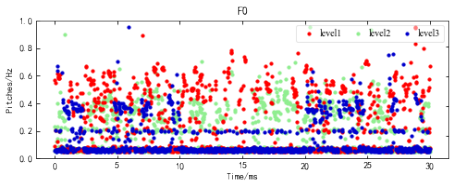

Figure 2: *F0 at different levels of fluency*

Table 3: *Comparison of unimodal with different models*

Feature	Models	Acc(%)	F1(%)	Corr(%)
Text	MNB	62.50	59.30	64.38
	LR	**65.00**	63.84	72.10
	SVM	64.17	**64.36**	**74.60**
	MLP	62.50	63.53	67.85
MFCC	MNB	**52.50**	**52.27**	32.20
	LR	51.67	51.71	**38.56**
	SVM	42.50	42.45	36.24
	MLP	47.50	40.77	34.18
Prosody	MNB	51.67	50.10	26.45
	LR	74.17	73.87	72.69
	SVM	70.83	70.77	77.06
	MLP	**75.83**	**75.61**	**80.77**

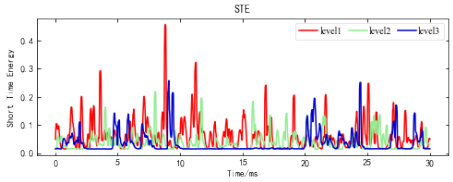

Figure 3: *STE at different levels of fluency*

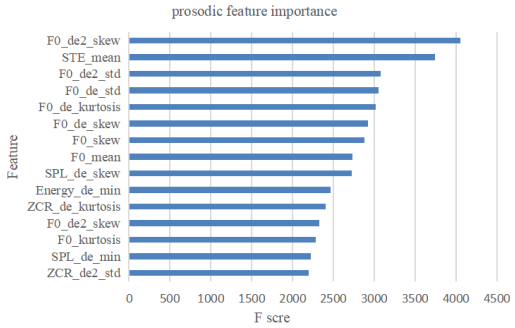

Figure 4: *Top 15 prosodic features that affect fluency*

4.4.2. *Multimodal Comparison Results*

The classical multimodal encoders [15] mentioned in Session 2 are used as the comparison models for this experiment, which include the Audio Recurrent Encoder (ARE), the Text Recurrent Encoder (TRE), the Multimodal Dual Recurrent Encoder (MDRE), and the Multimodal Dual Recurrent Encoder with Attention (MDREA). In addition, our modal is compared with the existing state-of-the-art baselines for multimodal analysis. The Later Fusion DNN (LF-DNN) first learns unimodal features and then concatenates these features before classification. The Tensor Fusion Network (TFN) [26] calculates a multi-dimensional tensor to capture unimodal and bimodal interactions. The Low-rank Multimodal Fusion (LMF) [27] performs effective multimodal fusion by using modality-specific low-rank factors. The Multimodal Transformer (MULT) [18] uses the cross-modal transformer to fuse different modalities. The comparative results are shown in Table 4. From these results, we find that our proposed model surpasses other models in most of the evaluations.

Table 4: *The results of different modals on PSC propositional speaking dataset*

Models	Acc(%)	F1(%)	Corr(%)	MAE
ARE	63.33	62.60	73.00	0.48
TRE	64.17	63.22	73.86	0.45
MDRE	68.33	67.48	80.03	0.43
MDREA	65.00	65.01	75.86	0.44
LF-DNN	72.50	72.05	79.21	0.42
TFN	70.83	70.35	79.92	0.42
LMF	73.33	72.76	81.41	0.39
MULT	80.83	81.11	84.05	**0.30**
Ours	**81.67**	**81.66**	**85.51**	0.34

4.4.3. *Ablation study*

The ablation study of model performance with varying modalities combination is shown in Table 5. The first three rows are the results with only one feature. The next three rows summarize the results from the combination of different features. The following two rows evaluate the influence of the feature extraction and self-attention mechanism. Through ablation study, we observe that combining textual and acoustic features significantly can improve the model's performance.

Table 5: *Ablation study on the proposed modal*

Description	Acc(%)	F1(%)	Corr(%)
Text only	63.33	62.54	70.61
MFCC only	75.00	74.71	82.44
Prosody only	38.33	43.69	43.35
MFCC+Text	76.67	76.03	**85.91**
Prosody+Text	65.83	64.88	74.53
MFCC+prosody	73.33	72.92	81.70
w/o Transformer&GRU	76.67	76.09	81.98
w/o Self-Attention	75.00	75.08	84.02
Ours	**81.67**	**81.66**	85.51

5. Conclusions

Speech fluency is one of the critical evaluation criteria for PSC. However, little work has been done to achieve automatic speech fluency evaluation. This paper introduces a multimodal method based on textual and acoustic modalities for automatic speech fluency evaluation. Our research contributions are as follows: First, a 29.83-hour dataset of PSC propositional speaking is constructed. Then, some features that influence the evaluation of speech fluency are explored. Finally, fluency evaluation results are obtained using cross-modal attention and self-attention to achieve multimodal fusion and information reinforcement. Experiments demonstrate that the fused features used in this paper provide complementary information to improve the accuracy of fluency evaluation. Future work will focus on improving the accuracy of automatic speech fluency evaluation.

6. Acknowledgements

This paper is supported by Open Research Fund Program of Beijing National Research Center for Information Science and Technology (BNR2021KF02005).

7. References

[1] The Putonghua training and testing center of the national language committee, *Implementation Outline for Putonghua Proficiency Test*, The Commercial Press, Beijing China, pp.4-5, 2004 (in Chinese).

[2] X. Guo, "The quantitative measure and scoring of fluency of spoken chinese as a second language," *Journal of Xiangtan Normal University (Social Science Edition)*, vol. 29, no. 04, pp. 91–94, 2007 (in Chinese).

[3] D. Liu, Y. Zhou, C. Zong, and F. Ren, "Automatic evaluation of sentence fluency," in *SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483)*, vol. 2. IEEE, 2003, pp. 1687–1692.

[4] S. Bhat, M. Hasegawa-Johnson, and R. Sproat, "Automatic fluency assessment by signal-level measurement of spontaneous speech," in *Second Language Studies: Acquisition, Learning, Education and Technology*, 2010.

[5] A. Loukina, K. Zechner, J. Bruno, and B. B. Klebanov, "Using exemplar responses for training and evaluating automated speech scoring systems," in *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, 2018, pp. 1–12.

[6] L. Fontan, M. L. Coz, and S. Detey, "Automatically measuring l2 speech fluency without the need of asr: A proof-of-concept study with japanese learners of french," in *Interspeech 2018*, 2018.

[7] S. Mao, Z. Wu, J. Jiang, P. Liu, and F. K. Soong, "Nn-based ordinal regression for assessing fluency of esl speech," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[8] G. Molholt and L. Liao, "A hybrid approach to assessing spoken fluency combining three metrics with support vector machines," in *2011 4th International Conference on Biomedical Engineering and Informatics (BMEI)*, vol. 4. IEEE, 2011, pp. 2228–2231.

[9] H. Deng, Y. Lin, T. Utsuro, A. Kobayashi, H. Nishizaki, and J. Hoshino, "Automatic fluency evaluation of spontaneous speech using disfluency-based features," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 9239–9243.

[10] H. Deng, T. Utsuro, A. Kobayashi, and H. Nishizaki, "Comparison of static and time-sequential features in automatic fluency detection of spontaneous speech," in *2021 24th Conference of the Oriental COCOSDA International Committee for the Coordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, 2021, pp. 158–163.

[11] H. Zhang, K. Shi, and N. F. Chen, "Multilingual speech evaluation: Case studies on english, malay and tamil," in *Interspeech 2021*, 2021.

[12] O. D. Deshmukh, K. Kandhway, A. Verma, and K. Audhkhasi, "Automatic evaluation of spoken english fluency," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4829–4832.

[13] G. Sammit, Z. Wu, Y. Wang, Z. Wu, A. Kamata, J. Nese, and E. C. Larson, "Automated prosody classification for oral reading fluency with quadratic kappa loss and attentive x-vectors," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3613–3617.

[14] H. Yang, D. Li, and Y. Yan, "Improving fluency of spoken mandarin for nonnative speakers by prosodic boundary prediction based on deep learning," *Wireless Communications and Mobile Computing*, vol. 2022, 2022.

[15] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 112–118.

[16] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019. NIH Public Access, 2019, p. 6558.

[19] B. Xie, M. Sidulova, and C. H. Park, "Robust multimodal emotion recognition from conversation with transformer-based crossmodality fusion," *Sensors*, vol. 21, no. 14, p. 4913, 2021.

[20] V. Rajan, A. Brutti, and A. Cavallaro, "Is cross-attention preferable to self-attention for multi-modal emotion recognition?" in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4693–4697.

[21] J. Cheng, R. Liang, and L. Zhao, "Dnn-based speech enhancement with self-attention on feature dimension," *Multimedia Tools and Applications*, vol. 79, no. 43, pp. 32 449–32 470, 2020.

[22] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.

[23] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python." Citeseer, 2015.

[24] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[25] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[26] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint arXiv:1707.07250*, 2017.

[27] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," *arXiv preprint arXiv:1806.00064*, 2018.