INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Discrimination of the Different Intents Carried by the Same Text Through Integrating Multimodal Information

Zhongjie Li[1,2], *Gaoyan Zhang*[1,*], *Longbiao Wang*[1], *Jianwu Dang*[1,2,*]

[1]CCA Lab, College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Information Science, Japan Advanced Institute of Science and Technology
{zhongjie_li2019, zhanggaoyan, longbiao_wang, dangjianwu}@tju.edu.cn

Abstract

Many intent understanding studies neglect the impact of paralinguistic information, resulting in misunderstandings during speech interactions, particularly when different intentions are conveyed by the same text with varying paralinguistic information. To address this issue, this study developed a Chinese multimodal spoken language intention understanding dataset that features different spoken intentions for identical texts. Our proposed attention-based BiLSTM model integrates textual and acoustic features and introduces an acoustic information gate mechanism to supplement or correct linguistic intention with paralinguistic intention. Experimental results demonstrate that our multimodal integration model improves intent discrimination accuracy by 11.0% compared to models that incorporate only linguistic information. The result highlights the effectiveness of our proposed model for intent discrimination, particularly in cases with identical text but varying intentions.

Index Terms: spoken language intent understanding, human-computer interaction, multimodal information integration

1. Introduction

In the last decade, with the development of artificial intelligence and the popularization of smart devices, human-computer intelligent dialogue technology has received extensive attention. Spoken intent understanding is the core module of the whole dialogue system, so it is a key issue to accurately obtain the comprehensive intent information transmitted by the speaker, including linguistic intents carried by textual information and paralinguistic intents carried by acoustic information.

Previous studies have shown that paralinguistic information, such as the speaker's emotion, attitude, and confidence, plays an important role in intent understanding [1, 2]. Recently, some researchers directly use the acoustic modality for intention understanding [3, 4], while a majority of studies use the text modality to discriminate linguistic intention. For instance, in task-domain dialogue research, relying solely on textual information has been shown to be effective in decoding the speaker's intent [5, 6, 7, 8, 9]. Furthermore, pre-trained models have gained significant attention and demonstrated remarkable performance in the field of intent recognition [10, 11, 12].

However, in some ambiguous cases, it is difficult to get the true intent of the speaker through only textual information, even with the most advanced pre-training language models. For example, as shown in Table 1, for the identical Chinese text "我一点也不生气", the intent is completely different when using different prosodies to pronounce it. As demonstrated in case 2 of Figure 1, the linguistic intent conveyed by the textual information conflicts with the paralinguistic intent conveyed by the acoustic information. Likewise, this phenomenon is prevalent in English, where the absence of paralinguistic information during human-computer interaction can impede the computer's ability to apprehend genuine human intention. Therefore, the challenge of this study is to get the true intention when the modalities have conflicting information.

To address this issue, we intend to integrate text and acoustic information together to decode the true intention. Considering the lack of datasets that include different intentions embedded in the same text, in this work, we first built a Chinese multimodal dataset including audio, text, and electroencephalogram (EEG) signals [13] (the EEG signal is not used in this study) when subjects transmitted different intents using identical text. Afterward, we proposed a multimodal information fu-

Table 1: *Examples of the same Chinese text with different user intents in the CMSLIU dataset.*

Input message	Intent
E.g. 1: 这首歌真好听	
Case 1: This song sounds wonderful!	Praise
Case 2: This song sounds terrible!	Irony
E.g. 2: 我一点也不生气	
Case 1: I am not angry at all.	Literal
Case 2: I am very angry.	Irony
E.g. 3: 你会不会开车	
Case 1: Can you drive?	Query
Case 2: Your car skills are too bad.	Antipathy

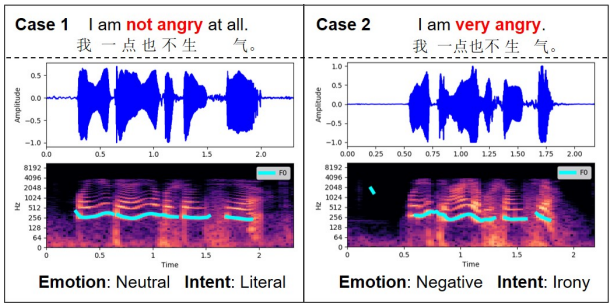

Figure 1: *Examples about the influence of paralinguistic information on intent understanding. The user expresses the same Chinese text with different prosodies (Case 1 is literal and case 2 is irony). Thus the responses of the speech device should be different if it understands the user's true intention.*

*Corresponding author. This work is supported in part by the National Natural Science Foundation of China (NSFC) No.62276185, No.61876126, and in part by JSPS KAKENHT Grant (No. 20K11883).

10.21437/Interspeech.2023-2444

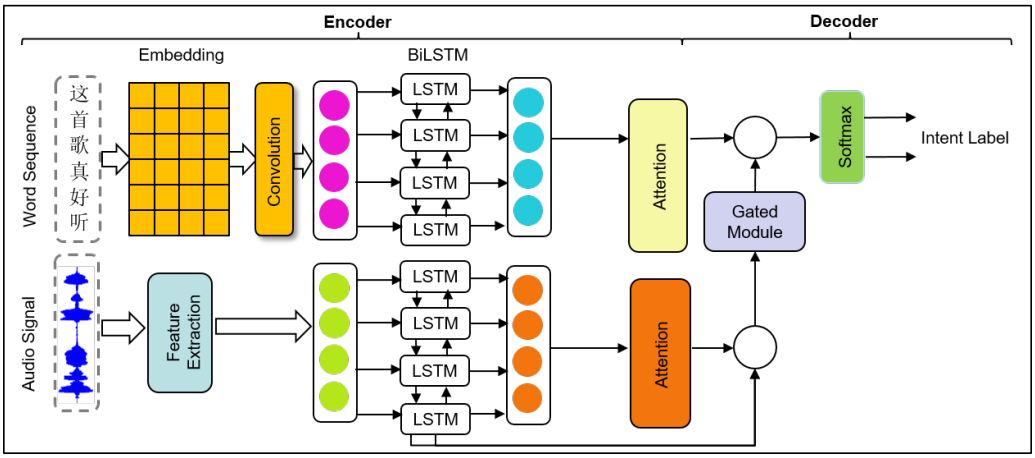

Figure 2: *The overview of the proposed intention discrimination model based on multimodal information integration.*

sion model using an attention-based Bidirectional Long Short-Term Memory (BiLSTM) network to integrate features related to linguistic intent from text sequence, and features from audio signal related to paralinguistic intent to get the speaker's comprehensive intent.

The main contributions of this work are as follows:

1) We designed and collected the Chinese spoken language intent understanding dataset to simulate the situation in a real conversation in which speakers express different intentions using identical text.

2) We develop a multimodal information fusion model for intent discrimination, and introduce a gate mechanism to learn the relationship between linguistic intents and paralinguistic intents, so as to achieve the supplement or correction of paralinguistic intents on linguistic intents.

2. Dataset construction

2.1. Dataset introduction

In this work, we collected the first Chinese multimodal dataset **CMSLIU**[1] that considered situations of the same text embedded by different intentions. The CMSLIU dataset has 5520 audio-text utterances in total, including task-oriented utterances, such as weather queries and booking tickets, and the open domain Chit-Chat. It is annotated with six intent labels including *Query*, *Directive*, *Irony*, *Praise*, *Literal* and *Antipathy*. It is also marked with three emotion labels, namely, positive, neutral, and negative. Examples of the different intentions embedded in the same Chinese text are shown in Table 1.

2.2. Data collection

Thirty native Mandarin speakers (20-28 years old) participated in data collection. The collection was performed in a sound-proof, electromagnetically shielded room. During the task, the text presentation and audio collection were conducted with Psychtoolbox-3 (www.psychtoolbox.org) running in MATLAB R2019a. The experiment included 15 sessions and 184 sentences. Each session started with a 10-s silent period, followed by a video clip lasting about 3 to 5 minutes to induce positive, neutral, or negative emotions of the subjects. Then, a set of sentence texts with the same emotional tags were sequentially presented on a screen for the participants to read out with predefined intentions. The audio recording started with a button press of 'S' and ended with a button press of 'E'. The speech signals were recorded using an electric condenser microphone (iCON Ultra 4) at a sample rate of 44100 Hz. For more details about the audio data corresponding to the Chinese texts given in Table 1 and more details about data collection, please visit the CMSLIU dataset URL[1].

Table 2: *Statistics of CMSLIU, ATIS, and Snips datasets.*

	CMSLIU	**ATIS**	**Snips**
# Intents	6	21	7
# Slots	17	120	72
# Emotion	3	/	/
Training set size	4,380	4,478	13,084
Testing set size	1,140	893	700

In addition, to verify the generalization of the proposed method, we used another two widely used English datasets in intent understanding research, that is, the **ATIS** (Airline Travel Information Systems) [14] and **Snips**[2]. They mainly focus on task domain dialogues. The statistic descriptions of the three datasets are shown in Table 2.

3. Proposed Method

An overview of the proposed multimodal information fusion model is illustrated in Figure 2. It includes a feature extraction module, a multimodal information encoder, and a decoder for intent discrimination.

3.1. Feature extraction

We extracted both textual features and acoustic features for intention discrimination. Previous research has demonstrated that RoBERTa [15], which is optimized from BERT [16] with larger training data and longer training time, generates more accurate Chinese text embeddings in single Chinese text intent recognition tasks. Therefore, for text modality, we apply the RoBERTa model with Whole Word Masking to encode each Chinese char-

[1]https://drive.google.com/drive/folders/1w76HxNj4zWK3snpdjlr9-aDNRddOIrlD?usp=sharing

[2]https://github.com/snipsco/nlu-benchmark/tree/master/2017-06-custom-intent-engines

acter into a high-dimensional vector to represent the textual content.

For the acoustic signal, we have also tried the speech-based pre-training models such as wav2vec2.0 [17], HuBERT [18] and Conformer [19], but those pre-training models did not achieve good performance. The reason may be that these pre-training models were trained with data that have consistent information across modalities, while our study deal with the data that have conflicting information across modalities. Therefore, instead of using the pre-training model, we introduce prior knowledge to construct distributed acoustic features. For example, the longer duration of the *"真 (really)"* in *"这首歌真好听 (This song sounds **rea∼lly** good)"*, is more likely to express irony intent, while the shorter the duration, is more likely to express praise intent. Moreover, F0 tends to decline when people express irony intention. In sum, all these advantages are not available in the pre-training model. As a result, we use a handcrafted feature extraction method to get effective paralinguistic information from speech signals [20]. The details about acoustic feature extraction are described as follows:

a) *F0 and Energy*: The paralinguistic information in speech, i.e. the speaker's emotion and confidence, is commonly described in terms of prosody features such as F0 and energy. In this work, the standard Root Mean Square Energy (RMSE) is employed to calculate speech energy using:

$$E = \sqrt{\frac{1}{n}\sum_{i=1}^{n} y[i]^2} \tag{1}$$

The RMSE is computed frame by frame, and we take the mean and standard deviations as features.

b) *Pause*: We adopt this feature to represent the "silent" part of the audio signal. This value is closely related to our confidence as well as emotion when we talk. The pause is given by:

$$Pause = Pr(y[n]<s) \tag{2}$$

where *s* represents the suitable threshold, which is approximately equal to 0.4 * RMSE.

c) *Harmonics*: The median-based filter described in [21] is used to calculate harmonics. For a given input vector *x(n)*, we can create a median filter for a given window size *l*:

$$y[n] = med(x[n-k:n+k], k = (l-1)/2) \tag{3}$$

where *l* is odd and *y*[*n*] is the output of median filter. In addition, the median is obtained as the average of the two values in the middle of the sorted list when *l* is an even number. Then median filtering is performed on the *h*th frequency slice S_h of a given spectrogram **S**, so as to obtain harmonic-enhanced spectrogram frequency slice H_h :

$$H_h = M(S_h, l_{harm}) \tag{4}$$

where *M* is the median filter, and l_{harm} is the length of the harmonic filter.

d) *Central moments*: At last, we "summarize" the input information using the average and standard deviation of the audio signal amplitude.

3.2. Encoder for textual and acoustic information

For textual information, word embedding is carried out by the Chinese pre-training RoBERTa model first. Next, a convolution layer performs a discrete convolution on the input matrix to extract sentence features [22, 23], where each row of the matrix is the word embedding of the corresponding word. Then, the newly generated matrix is input into the BiLSTM [24] model. And the final hidden state h_j at time step *j* is a concatenation of forward hidden state $\overrightarrow{h_j}$ and backward hidden state $\overleftarrow{h_j}$.

For acoustic information, we first perform feature extraction on the input audio signal, and then the extracted feature sets are fed into an attention-based BiLSTM model. For each hidden state h_j, we compute the intent context vector C^{AI} based on audio as the weighted sum of LSTM's hidden states $h_1, \cdots, h_L$, by the learned attention weights α_j^{AI}:

$$C^{AI} = \sum_{j=1}^{L} \alpha_j^{AI} h_j \tag{5}$$

where the acoustic attention weights are computed as below:

$$\alpha_j^{AI} = \frac{exp(e_j)}{\sum_{m=1}^{L} exp(e_m)} \tag{6}$$

$$e_m = \sigma(W_{he}^{AI} h_m) \tag{7}$$

Where σ is the activation function, and W_{he}^{AI} is the weight matrix of a feed-forward neural network. Similarly, the intent context vector C^{TI} based on text can also be computed as C^{AI}.

In this study, we additionally introduced a gating mechanism for achieving the supplement or correction effect of paralinguistic intents carried by acoustic information for linguistic intents carried by textual information.

$$g = v \cdot tanh(C^{TI} + W \cdot C^{AI}) \tag{8}$$

where C^{TI} and C^{AI} represent the context vectors of linguistic intent and paralinguistic intent respectively. Here, *v* and *W* are trainable vector and matrix respectively. Significantly, g can be seen as a weighted feature of the context vector C^{TI} and C^{AI}.

3.3. Decoder and training for intent discrimination

We use a decoder to perform intent discrimination, which can be denoted as below:

$$y^I = softmax(W_{hy}^I(h_L^T + h_L^A \cdot g)) \tag{9}$$

where W_{hy}^I is weight matrix.h_L^T and h_L^A represents the last hidden state of the text and audio BiLSTM model, respectively.

Finally, the intent discrimination objection is formulated as:

$$L = -\sum_{j=1}^{T} \hat{y}_j^I log(y_j^I) \tag{10}$$

where $\hat{y}_j^I$ is the gold intent label.

4. Experiments

4.1. Experiment setting

In this study, the CMSLIU dataset is split into the training set and testing set, which contains 4,380 and 1,140 text-audio samples, separately. For text modality, to justify the generalization of the proposed model, we also implemented experiments on ATIS and Snips benchmark datasets. For audio modality, we down-sample the audio samples to 8 kHz first, and then the feature sets are extracted by librosa[3] toolkit.

[3]https://github.com/librosa/librosa

Table 3: *Comparisons of intent discrimination accuracy (%) on three datasets among different models (where T is for Text and A is for Acoustic).*

Model	Year	ATIS	Snips	CMSLIU
LIDSNet [25]	2021	96.0	98.0	50.5
Stack-Prop [26]	2019	96.9	98.0	52.3
DCA-Net [5]	2021	97.7	98.8	52.4
Bilinear [27]	2022	**98.2**	98.9	52.9
Our (T)		98.0	**98.9**	53.1
Our (A)		/	/	59.2
Our (T+A)		/	/	60.6
Our (T+A+Gate)		/	/	**64.1**

In all experiments, the convolution layer used one 3×3 filter with the SELU activation function. The size of hidden units and maximum training epochs are set as 128 and 50, respectively. The Adam optimizer is applied to optimize the parameters in our model. The initial learning rate is 10^{-3}, and the learning rate is set to 5×10^{-4} when the training accuracy is greater than 60% but less than 80%. The learning rate is set to 10^{-4} while training accuracy exceeds 80%. Following Qin et al. [5], we use accuracy to evaluate the intent recognition performance.

4.2. Results and analysis

Table 3 shows that the six intents discrimination accuracy of our proposed multimodal information integration model on the CMSLIU dataset reached 64.1%. To verify the supplement or correction of paralinguistic intent carried by acoustic information on linguistic intent transmitted by textual information, we implemented ablation experiments on the CMSLIU dataset. The confusion matrix of the ablation experiment is given in Figure 3. And the accuracy of intent discrimination decreases to 53.1% when we input only the textual information. We also conducted experiments on ATIS and Snips datasets to exclude the influence of models and datasets as shown in Table 3, where the compared baselines for intent recognition include the state-of-the-art (SOTA) models LIDSNet [25], Stack-Propagation [26], DCA-Net [5], and Bilinear attention [27].

Based on the experimental analysis, we have the following findings:

1) The proposed model achieves satisfactory performance on all three datasets when only textual information is used as input. It should be noted that the goal of this study is not to achieve SOTA performance on intent recognition in the ATIS and Snips datasets, but rather to validate that our proposed model performs competitively on text-based datasets compared to other mainstream models. Although the previous SOTA models have achieved good intent recognition results on ATIS and Snips datasets, the intent discrimination accuracy of each model has decreased significantly on the CMSLIU dataset. In fact, this phenomenon is quite normal since the samples in the CMSLIU dataset with the same text but different intent labels, which cannot be correctly discriminated by text-based linguistic information alone. That is, the textual information input into the model is exactly identical, but corresponds to different intent labels. In such cases, even the most ideal text-based model cannot work in the classification task of the CMSLIU dataset.

2) Comparing Figure 3 (A) with (B), it is found that the textual-based method can achieve high accuracy in identifying intentions that are insensitive to paralinguistic information, such as *Query* and *Directive*. However, for the other four intentions, the accuracy is significantly lower than that audio-based method due to the lack of paralinguistic information.

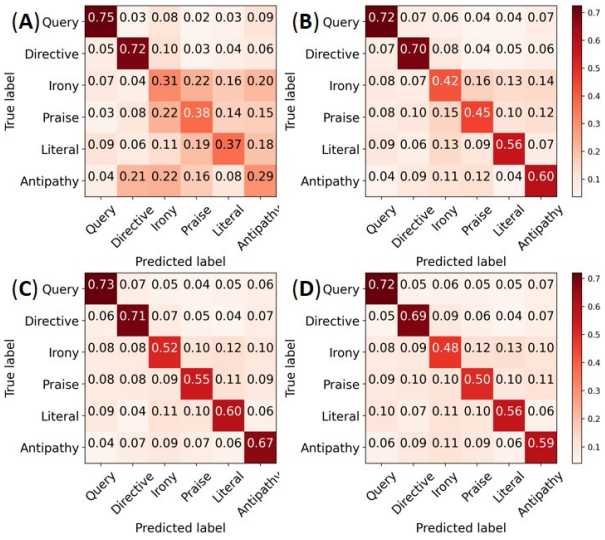

Figure 3: *The confusion matrix of (A) Text-only (B) Acoustic-only (C) Text + Acoustic + Gate (D) Text + Acoustic.*

3) As shown in Figure 3 (A) and (C), after fusing acoustic and textual information, the overall intent recognition accuracy of our framework improves by 11.0%, especially for *Irony*, *Praise* and *Antipathy*. The reason may be that the paralinguistic information contained in the audio modulates the textual information through the gate mechanism to distinguish which intention the speaker expresses. In contrast, the intention discrimination accuracy is significantly declined when the gating module is removed and the features of the two modalities are directly concatenated, as shown in Figure 3 (D), which proves the effectiveness of the gating mechanism.

5. Conclusions

In this work, we proposed a novel attention-based BiLSTM multimodal information fusion model for decoding comprehensive intent information containing paralinguistic intents carried by acoustic information and linguistic intents carried by textual information.

First, we constructed a Chinese spoken language intention understanding dataset and then proposed a method to fuse the textual and acoustic information together for intent discrimination, in which the acoustic information gate mechanism was introduced to supplement or correct linguistic intention with the paralinguistic intention. The result highlights that our proposed model can effectively utilize paralinguistic information in intent discrimination, particularly in cases where the identical text could not provide any useful information to distinguish varying intentions.

In future work, we will additionally fuse EEG features [28, 29] to obtain the intent representations in the brain so as to further improve the intent discrimination performance.

6. Acknowledgements

We are deeply grateful to Professor Shogo Okada from JAIST for his invaluable guidance and feedback on this research.

7. References

[1] S. Matsubara, S. Kimura, N. Kawaguchi, Y. Yamaguchi, and Y. Inagaki, "Example-based speech intention understanding and its application to in-car spoken dialogue system," in *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002.

[2] Y. Gu, X. Li, S. Chen, J. Zhang, and I. Marsic, "Speech intention classification with multimodal deep learning," in *Canadian conference on artificial intelligence*. Springer, 2017, pp. 260–271.

[3] Y. Ning, J. Jia, Z. Wu, R. Li, Y. An, Y. Wang, and H. Meng, "Multi-task deep learning for user intention understanding in speech interaction systems," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[4] J. Lehmann, N. Christen, Y. Barilan, and I. Gannot, "Age-related hearing loss, speech understanding and cognitive technologies," *International Journal of Speech Technology*, vol. 24, no. 2, pp. 509–516, 2021.

[5] L. Qin, Z. Li, W. Che, M. Ni, and T. Liu, "Co-gat: A co-interactive graph attention network for joint dialog act recognition and sentiment classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 15, 2021, pp. 13 709–13 717.

[6] H. Tang, D. Ji, and Q. Zhou, "End-to-end masked graph-based crf for joint slot filling and intent detection," *Neurocomputing*, vol. 413, pp. 348–359, 2020.

[7] O. Hamed and H. J. Steinhauer, "Pedestrian's intention recognition, fusion of handcrafted features in a deep learning approach," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 18, 2021, pp. 15 795–15 796.

[8] F. Cai, W. Zhou, F. Mi, and B. Faltings, "Slim: Explicit slot-intent mapping with bert for joint multi-intent detection and slot filling," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7607–7611.

[9] P. Wei, B. Zeng, and W. Liao, "Joint intent detection and slot filling with wheel-graph attention networks," *Journal of Intelligent & Fuzzy Systems*, no. Preprint, pp. 1–12, 2022.

[10] Z. Ma, B. Sun, and S. Li, "A two-stage selective fusion framework for joint intent detection and slot filling," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[11] F. Yang, X. Zhou, Y. Wang, A. Atawulla, and R. Bi, "Diversity features enhanced prototypical network for few-shot intent detection," in *Proc. International Joint Conference on Artificial Intelligence*, vol. 7, 2022, pp. 4447–4453.

[12] A. Kumar, V. Malik, and J. Vepa, "Does utterance entails intent?: Evaluating natural language inference based setup for few-shot intent detection," *Proc. Interspeech 2022*, pp. 4501–4505, 2022.

[13] Z. Li, B. Zhao, G. Zhang, and J. Dang, "Brain network features differentiate intentions from different emotional expressions of the same text," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[14] G. Tur, D. Hakkani-Tür, and L. Heck, "What is left to be understood in atis?" in *2010 IEEE Spoken Language Technology Workshop*. IEEE, 2010, pp. 19–24.

[15] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-training with whole word masking for chinese bert," 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9599397

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[17] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[18] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[19] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[20] S. H. Dumpala, S. Rempel, K. Dikaios, M. Sajjadian, R. Uher, and S. Oore, "Estimating severity of depression from acoustic features and embeddings of natural speech," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7278–7282.

[21] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, vol. 13, 2010.

[22] Y. Liu, H. Liu, L.-P. Wong, L.-K. Lee, H. Zhang, and T. Hao, "A hybrid neural network rbert-c based on pre-trained roberta and cnn for user intent classification," in *Neural Computing for Advanced Applications: First International Conference, NCAA 2020, Shenzhen, China, July 3–5, 2020, Proceedings 1*. Springer, 2020, pp. 306–319.

[23] S. Rathor and S. Agrawal, "Sense understanding of text conversation using temporal convolution neural network," *Multimedia Tools and Applications*, vol. 81, no. 7, pp. 9897–9914, 2022.

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

[25] V. Agarwal, S. D. Shivnikar, S. Ghosh, H. Arora, and Y. Saini, "Lidsnet: A lightweight on-device intent detection model using deep siamese network," in *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2021, pp. 1112–1117.

[26] L. Qin, W. Che, Y. Li, H. Wen, and T. Liu, "A stack-propagation framework with token-level intent detection for spoken language understanding," *EMNLP/IJCNLP, Association for Computational Linguistics, pp. 2078-2087*, 2019.

[27] D. Chen, Z. Huang, and Y. Zou, "Leveraging bilinear attention to improve spoken language understanding," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7142–7146.

[28] Z. Li, G. Zhang, J. Dang, L. Wang, and J. Wei, "Multi-modal emotion recognition based on deep learning of eeg and audio signals," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–6.

[29] Z. Li, G. Zhang, L. Wang, J. Wei, and J. Dang, "Emotion recognition using spatial-temporal eeg features through convolutional graph attention network," *Journal of Neural Engineering*, vol. 20, no. 1, p. 016046, 2023.