FUSING ASR OUTPUTS IN JOINT TRAINING FOR SPEECH EMOTION RECOGNITION

Yuanchao Li, Peter Bell, Catherine Lai

Centre for Speech Technology Research
University of Edinburgh, Scotland, UK
y.li-385@sms.ed.ac.uk, {peter.bell, c.lai}@ed.ac.uk

ABSTRACT

Alongside acoustic information, linguistic features based on speech transcripts have been proven useful in Speech Emotion Recognition (SER). However, due to the scarcity of emotion labelled data and the difficulty of recognizing emotional speech, it is hard to obtain reliable linguistic features and models in this research area. In this paper, we propose to fuse Automatic Speech Recognition (ASR) outputs into the pipeline for joint training SER. The relationship between ASR and SER is understudied, and it is unclear what and how ASR features benefit SER. By examining various ASR outputs and fusion methods, our experiments show that in joint ASR-SER training, incorporating both ASR hidden and text output using a hierarchical co-attention fusion approach improves the SER performance the most. On the IEMOCAP corpus, our approach achieves 63.4% weighted accuracy, which is close to the baseline results achieved by combining ground-truth transcripts. In addition, we also present novel word error rate analysis on IEMOCAP and layer-difference analysis of the Wav2vec 2.0 model to better understand the relationship between ASR and SER.

Index Terms— Speech emotion recognition, Automatic speech recognition, Multi-task learning, Wav2vec 2.0

1. INTRODUCTION

Despite the rapid development of SER, its use in the wild remains difficult, due in part to the variability and vulnerability of human speech. To address this issue, researchers have proposed combining linguistic (i.e. transcript-derived) and acoustic information to improve the stability of SER. Traditional feature-level fusion and decision-level fusion, as well as the current tensor-level fusion trend, are widely used [1]. Furthermore, many researchers have also compared the performance of various fusion methods [2]. All of these works demonstrated the utility of combining linguistic information in SER.

Nonetheless, there is limited reliable text information in this area, particularly in real-life SER. First, when compared to ASR, the corpora sizes for SER are relatively small. For example, the popular IEMOCAP [3] only has about 12 hours of speech, whereas LibriSpeech [4] has about 1,000 hours. SER models built on such limited sized corpora don't generalize well to out-of-domain speech. Second, while previous studies proposed using ASR to generate transcripts for SER [5], ASR on emotional speech can often result in relatively high error rates. Previous research has shown that emotion in speech degrades ASR performance, with emotional speech assumed to be a distortion of neutral speech [6]. However, with the advancement of deep learning technologies, transfer learning for SER from ASR and joint training of ASR and SER have recently emerged [7, 8]. Nevertheless, the relationship between ASR and SER is still poorly studied, particularly what and how ASR features can benefit SER.

In this paper, we investigate various ASR outputs and fusion methods for a joint ASR-SER training model. We analyze the Word Error Rate (WER) on IEMOCAP using Wav2vec 2.0 (W2V2), elucidating the reasons why ASR fails on emotional speech, which has not previously been reported. We also compare four ASR outputs from W2V2 and three fusion methods, and find that our proposed hierarchical co-attention fusion achieves 63.4% Weighted Accuracy (WA), comparable performance to the baseline result using ground-truth transcripts.

2. RELATED WORK

The relationship between ASR and SER is an important but understudied topic. Although both tasks use speech signals as input, ASR works more at the frame level, whereas SER recognizes emotion on larger timescales. Previous work [9] has demonstrated that features from the initial layers of both ASR and SER tasks are transferable and that the relevance of features gradually fades through deep layers. However, ASR features closer to the audio modality can outperform those closer to the text modality for arousal prediction and vice versa for valence prediction [10].

These findings suggest it is possible to use transfer learning for SER from ASR. For example, [11] used hidden layer output from a pre-trained ASR model as linguistic features for SER and obtained promising results. Similarly, [8] used a single W2V2 model for both ASR and SER tasks and achieved State-Of-The-Art (SOTA) performance on 10-fold

ICASSP 2022

Cross-Validation (CV). However, gaps remain on foundational issues such as whether WER varies with emotion type, and where ASR features should enter SER models. In this paper, we build on this line of work, conducting new analyses on WER, ASR outputs, and fusion methods, with the goal of understanding how to best bring ASR into SER research.

3. EXPERIMENT PREPARATION AND PRELIMINARY ANALYSIS

3.1. Corpus and ASR Model

In these experiments, we use the IEMOCAP corpus [3] as it is widely used in SER research. It consists of five dyadic sessions with ten actors, each with a scripted and improvised multimodal interaction. The corpus contains approximately 12 hours of speech that has been annotated by three annotators using ten emotion classes. We combine happy and excited, and use four categories: angry, happy, neutral, and sad, based on prior research [12] . We removed utterances that did not have transcripts, bringing the total number of utterances used in this study to 5500.

We use the well-known W2V2 model [13] for the ASR model. W2V2 is a self-supervised learning framework composed of three major components: a CNN-based local encoder that extracts a sequence of embeddings from raw audio as latent representation, a Transformer network for context representation, and a quantization module for discretization. In this paper, we use the "Wav2vec2-base-960h" model, which has been fine-tuned using 960 hours of Librispeech.

3.2. Preliminary Analysis of WER

We calculate the WER on IEMOCAP to investigate the types of errors that occur in emotional speech (Table 1). Surprisingly, we see that the neutral WER is the second highest– we would expect WER for neutral speech to be the lowest as the least emotionally 'distorted'. Furthermore, while Angry and Happy can have similarly variable prosodic characteristics (high f0 and intensity) [14] , Angry has the lowest WER while Happy has the highest. To determine the reasons for this, we conducted a word count analysis. The results in Table 1 show that emotion classes with low WER typically have a low ratio of short utterances (word count less than 10). The fact that neutral speech is less prosodically variable may be the reason it has a lower WER than happy, even though its short utterances rate is the highest.

In fact, we calculated the WER based on the number of words, and found that the shorter the utterance, the higher the WER (see Table 2). This finding contradicts a previous finding that longer utterances are more likely to have higher error rates [5]. This may be due to the fact that, when using W2V2, longer utterances contain more contextual information that the ASR can use to compensate for the negative effects of emotion. However, short utterances, and similarly

Table 1. WER and word count analysis results. *Utter* is the number of utterances of each emotion class, and *Short%* is the ratio of short utterances.

	Ang	Hap	Neu	Sad	Overall
WER	22.8%	38.9%	36.3%	29.5%	32.7%
Utter	1103	1615	1704	1078	5500
Short%	44.8%	55.7%	60.5%	52.7%	

Table 2. WER according to different number of words N.

N	Ang	Hap	Neu	Sad	Overall
≤ 10	26.5%	54.8%	48.4%	47.2%	46.4%
11-**20**	21.8%	37.5%	36.1%	30.4%	32.2%
21-**30**	21.4%	31.0%	27.2%	22.6%	26.3%
≥ 30	20.9%	31.0%	26.1%	20.1%	25.0%

disfluencies, are clearly very important in understanding affect in spoken dialogues [15]. We hope that this preliminary analysis and novel findings will aid future ASR-SER research, by motivating the joint training to potentially learn these sorts of complex interactions.

4. METHOD

4.1. Feature Extraction and Encodings

In our model, the raw audio time-series is encoded by the W2V2 model (for ASR), and separately as MFCCs (for SER). On the ASR path, the W2V2 representations are then decoded to text by a word-level Connectionist Temporal Classification (CTC) decoder. Both the hidden state output and the text are extracted. A BERT model [16] is then used to extract linguistic features from text output. On the SER path, a max pooling with kernel size 2 is conducted on the MFCC features to obtain an acoustic representation.

The ASR hidden output (short for hidden layer output), text output, and pooled MFCCs, are then encoded using 2-layer Bidirectional Long Short-Term Memory (Bi-LSTM) networks. Each layer contains 32 hidden units followed by a dropout with probability of 0.5. A self-attention layer with 16 heads and 64 nodes is applied on the output of the Bi-LSTM, which generates a fixed-length vector as the feature encoding.

4.2. Feature Fusion and Emotion Output

We fuse the three encodings from the self-attention layer to produce a final vector for emotion classification. We compared three fusion methods (Fig. 1): (a) concatenation, (b) concatenation with co-attention fusion, and (c) our proposed hierarchical co-attention fusion. The co-attention mechanism
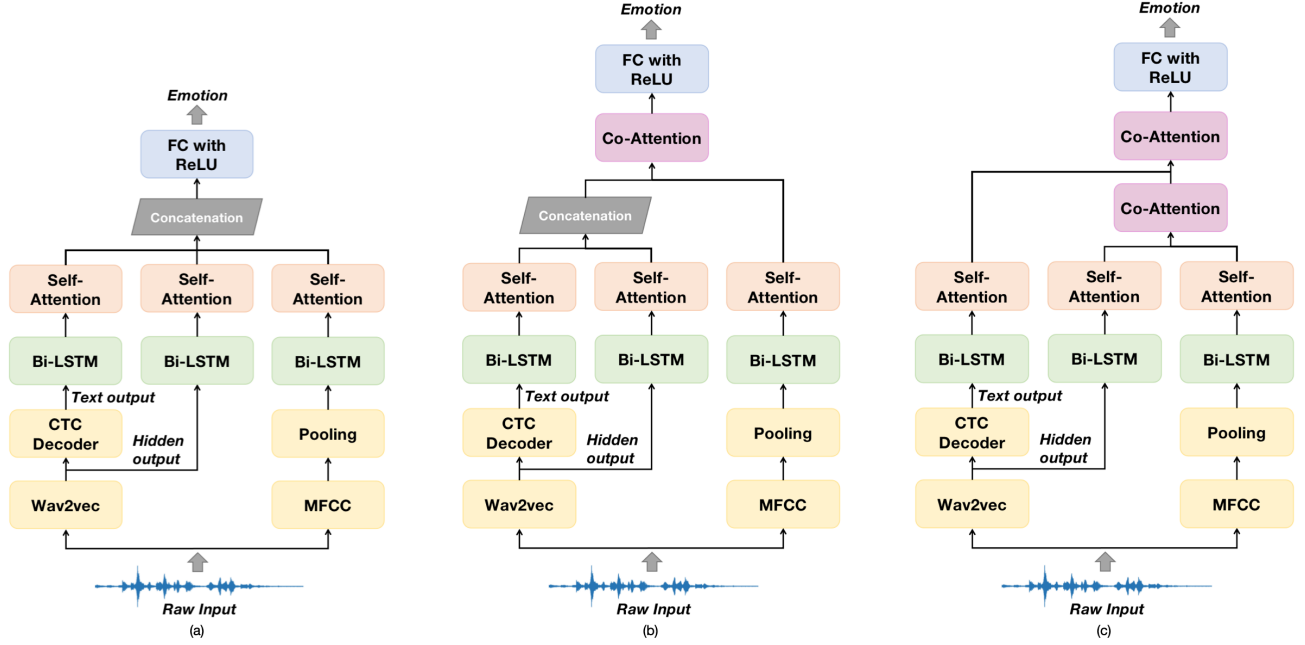
Fig. 1. Proposed model with different fusion methods.

[17] concatenates two hidden-state vectors which exchange key-value pairs in self-attention, allowing features from one input channel to be incorporated into the other:

$$H_C = Concat(H_{C_1}, H_{C_2}) \tag{1}$$

$$H_{C_i} = MultiHead(Q_A, K_B, V_B)W^O \tag{2}$$

$$= Concat(head_1, ..., head_n) \tag{3}$$

$$head_i = Attention(Q_A W_i^Q, K_B W_i^K, V_B W_i^V) \tag{4}$$

where W^O, W_i^Q, W_i^K, and W_i^V are trainable parameters. Q_A represents the query from one input channel, while K_B and V_B represent the key and value from the other. The value of n is 16, and H_C denotes the final concatenated hidden states of co-attention.

Hierarchical approaches for emotion recognition have proved useful in fusing different-level or different-type of features in previous works [18]. Inspired by these works and the hierarchical characteristics of speech [19], we propose to fuse the input features from low to high level in a hierarchical manner using co-attention. Because MFCC features are extracted within frames, W2V2 features contain within-utterance context, and BERT features can learn cross-sentence context, we refer to them as frame-level, utterance-level, and dialog-level respectively, and hierarchically fuse them in this order to generate a fixed-length vector. This is passed to a fully connected output layer with *Softmax* activation function to generate the probability distributions over emotion classes.

5. EXPERIMENTS AND RESULTS

The model is optimized by the multi-task loss function:

$$\mathcal{L} = \lambda\mathcal{L}_{ASR} + (1 - \lambda)\mathcal{L}_{SER}, \tag{5}$$

where $\mathcal{L}_{ASR}$ and $\mathcal{L}_{SER}$ are the losses for ASR and SER, respectively. We set λ to 0.2, and use cross-entropy as the loss function. We use the Adam optimizer, with a learning rate of 10^{-4} and decay rate of 10^{-5}. The gradients are clipped with a threshold of 5.0. The batch size is set to 20, and the number of epochs is limited to 100. We perform 5-fold CV and use WA to assess the performance.

The experiment results are shown in Table 3. Here, we present (1) the ground-truth model incorporating human transcripts in the SER model (i.e. no ASR); results of including (2) ASR hidden output or (3) text output; (4) the results of our proposed full model. We used the middle layer ASR output from W2V2 model because of its highest WA (see Table 4).

As expected, the ground-truth model achieves the best performance as the ASR outputs contain recognition errors. Both ASR hidden output and text output help improve the SER performance over the acoustic features, although the difference is small. In general, the performance of co-attention fusion outperforms that of concatenation. It appears that the relatedness of two input channels is learned by attention. We also see that when using ground-truth transcripts, co-attention improves performance more than when using ASR outputs. It is plausible because the incorrectly recognized ASR features may have little or no relatedness to acoustic features, reducing the effectiveness of attention. Finally, our proposed ap-

Table 3. Performance comparison of using different models, features and fusion methods.

Model	Feature	Fusion approach	WA
SER (*baseline*)	Acoustic + Ground-truth transcripts	Concatenation	63.5%
		Co-attention	**64.9%**
ASR-SER	Acoustic + Hidden output (middle layer)	Concatenation	58.7%
		Co-attention	59.1%
ASR-SER	Acoustic + Text output	Concatenation	59.4%
		Co-attention	60.3%
ASR-SER	Acoustic + Hidden output (middle layer) + Text output	Concatenation	62.2%
		Co-attention	62.9%
		Hierarchical co-attention	**63.4%**

Table 4. Layer-difference analysis results.

Model	Feature	WA
SER (*baseline*)	Acoustic	51.4%
	Ground-truth transcripts	60.0%
ASR-SER	Hidden output (first layer)	45.6%
	Hidden output (middle layer)	46.1%
	Hidden output (final layer)	42.0%
	Text output	47.7%

proach: the joint ASR-SER model incorporating both ASR hidden output and text output using hierarchical co-attention achieves 63.4% WA, which is nearly the same as the ground-truth transcript concatenation result and only 1.5% less than the ground-truth co-attention result. This indicates that hierarchical fusion can help ameliorate ASR errors.

We conducted a layer-difference analysis study to determine which W2V2 layer contributes the most to the SER task. [20] examined the layer-specific information in W2V2 intermediate speech representations and found that various acoustic and linguistic properties are encoded in different layers. As such, we compared the hidden-state outputs from the first layer (initial embeddings), middle layer, and final layer and found a surprising difference in SER performance. Table 4 shows that the WA of using the hidden outputs from first and middle layers are close, but decreases significantly when using the final layer. Furthermore, we discovered that the text output outperforms the hidden outputs, which contradicts prior assertions that "*ASR features can be more robust than the text output of ASR*" [11]. This difference may be specific to W2V2 and requires further investigation, which we leave for future work.

We present previous works that also use ASR outputs in SER, but note that due to variation in experimental setup, results are not directly comparable. [21] used the Google Cloud Speech API to generate transcripts that were then fused with MFCCs for SER. They achieved 69.1% WA, likely due to their low WER (5.53%). [5] concatenated GloVe features obtained from Wit.ai API text transcription with acoustic features, obtaining a 62.9% Unweighted Accuracy (UA) using an LSTM based model. [11] jointly trained ASR-SER using log mel-scale filter bank outputs and decoder outputs from a pre-trained ASR model, and achieved 68.6% WA. [8] used a single W2V2 as the training model for both ASR and SER rather than separating two tasks, and used 10-fold CV, which usually yields better results than 5-fold; [7] carefully fine-tuned the pre-trained ASR model on IEMOCAP before transferring the ASR features for SER. Note that since our goal is to analyze potential role of ASR in ASR-SER rather than achieve the SOTA performance, we did not focus on fine-tuning the model and parameters, but we expect that the performance of our model can be further improved by parameter tuning.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a joint ASR-SER training model that incorporates both ASR hidden output and text output using a hierarchical co-attention fusion approach. On IEMOCAP, it achieves 63.4% WA, which is comparable to the ground-truth transcript model. By conducting WER analysis, we demonstrate situations where ASR can fail on emotional speech. We hope that the findings, which have not previously been reported can contribute to understanding the ASR-SER relationship. In our future work, we will conduct detailed layer-wise analysis to understand what meanings the intermediate representations represent as the W2V2 layer becomes deeper. We will also try character-level or hybrid CTC decoders to solve the language model mismatch problem.

7. ACKNOWLEDGEMENTS

8. REFERENCES

[1] Björn Schuller, Gerhard Rigoll, and Manfred Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *Proceedings of ICASSP 2004*. IEEE, 2004.

[2] Jilt Sebastian, Piero Pierucci, et al., "Fusion Techniques for Utterance-Level Emotion Recognition Combining Speech and Transcripts," in *Proceedings of Interspeech 2019*, 2019, pp. 51–55.

[3] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[4] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proceedings of ICASSP 2015*, 2015, pp. 5206–5210.

[5] Saurabh Sahu, Vikramjit Mitra, Nadee Seneviratne, and Carol Espy-Wilson, "Multi-Modal Learning for Speech Emotion Recognition: An Analysis and Comparison of ASR Outputs with Ground Truth Transcription," in *Proceedings of Interspeech 2019*, 2019, pp. 3302–3306.

[6] Raul Fernandez, *A computational model for the automatic recognition of affect in speech*, Ph.D. thesis, Massachusetts Institute of Technology, 2004.

[7] Sitong Zhou and Homayoon Beigi, "A transfer learning method for speech emotion recognition from automatic speech recognition," *arXiv preprint arXiv:2008.02863*, 2020.

[8] Xingyu Cai, Jiahong Yuan, Renjie Zheng, Liang Huang, and Kenneth Church, "Speech emotion recognition with multi-task learning," in *Proceedings of Interspeech 2021*, 2021.

[9] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon, "On the Correlation and Transferability of Features Between Automatic Speech Recognition and Speech Emotion Recognition," in *Proceedings of Interspeech 2016*, 2016, pp. 3618–3622.

[10] Noé Tits, Kevin El Haddad, and Thierry Dutoit, "ASR-based features for emotion recognition: A transfer learning approach," in *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*. 2018, pp. 48–52, Association for Computational Linguistics.

[11] Han Feng, Sei Ueno, and Tatsuya Kawahara, "End-to-End Speech Emotion Recognition Combined with Acoustic-to-Word ASR Model," in *INTERSPEECH*, 2020, pp. 501–505.

[12] Yuanchao Li, Tianyu Zhao, and Tatsuya Kawahara, "Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning," in *Proceedings of Interspeech 2019*, 2019, pp. 2803–2807.

[13] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, 2020, pp. 12449–12460.

[14] Astrid Paeschke and Walter F Sendlmeier, "Prosodic characteristics of emotional speech: Measurements of fundamental frequency movements," in *ISCA Tutorial and Research Workshop on Speech and Emotion*, 2000.

[15] Leimin Tian, Johanna D Moore, and Catherine Lai, "Emotion recognition in spontaneous and acted dialogues," in *Proceedings of ACII 2015*. IEEE, 2015, pp. 698–704.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of NAACL-HLT 2019*, pp. 4171–4186, 2019.

[17] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Advances in neural information processing systems*, 2019.

[18] Leimin Tian, Johanna Moore, and Catherine Lai, "Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features," in *Proceedings of SLT 2016*. IEEE, 2016, pp. 565–572.

[19] Santiago Pascual, Mirco Ravanelli, Joan Serra, Antonio Bonafonte, and Yoshua Bengio, "Learning problem-agnostic speech representations from multiple self-supervised tasks," in *Proceedings of Interspeech 2019*, 2019.

[20] Ankita Pasad, Ju-Chieh Chou, and Karen Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *Proceedings of ASRU 2021*, 2021.

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