論文 / 著書情報
Article / Book Information

Title	MULTIMODAL EMOTION RECOGNITION WITH HIGH-LEVEL SPEECH AND TEXT FEATURES
Author	Mariana Rodrigues Makiuchi, Kuniaki Uto, Koichi Shinoda
Journal/Book name	2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) Proceedings, , , pp. 350-357
Pub. date	2021, 12
DOI	https://doi.org/10.1109/ASRU51503.2021.9688036
Copyright	(c)2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
Note	This file is author (final) version.

MULTIMODAL EMOTION RECOGNITION WITH HIGH-LEVEL SPEECH AND TEXT FEATURES

Mariana Rodrigues Makiuchi, Kuniaki Uto, Koichi Shinoda

Tokyo Institute of Technology

ABSTRACT

Automatic emotion recognition is one of the central concerns of the Human-Computer Interaction field as it can bridge the gap between humans and machines. Current works train deep learning models on low-level data representations to solve the emotion recognition task. Since emotion datasets often have a limited amount of data, these approaches may suffer from overfitting, and they may learn based on superficial cues. To address these issues, we propose a novel cross-representation speech model, inspired by disentanglement representation learning, to perform emotion recognition on wav2vec 2.0 speech features. We also train a CNN-based model to recognize emotions from text features extracted with Transformer-based models. We further combine the speech-based and text-based results with a score fusion approach. Our method is evaluated on the IEMOCAP dataset in a 4-class classification problem, and it surpasses current works on speech-only, text-only, and multimodal emotion recognition.

Index Terms— Emotion recognition, disentanglement representation learning, deep learning, multimodality, wav2vec 2.0

1. INTRODUCTION

Correctly perceiving other people's emotion is one of the key components of good interpersonal communication. Emotions make conversation more natural. They can add or remove ambiguity, and they can change the meaning of what is being communicated altogether. Due to the importance of emotions in human-to-human conversation, automatic emotion recognition has been one of the main concerns of the Human-Computer Interaction (HCI) field for decades [1].

Many studies have proposed emotion recognition methods from para-linguistic features or from transcribed text data [2, 3, 4]. However, depending on paralanguage, the semantics of the linguistic communication may change and vice-versa. Thus, spoken text may have different interpretations depending on the speech intonation. Additionally, similar speech intonations may be used to convey different emotions, which can only be discerned by understanding the linguistic factor of communication. Therefore, relying solely on linguistic or para-linguistic information may not be enough to correctly recognize emotions in conversation.

One of the most fundamental challenges in emotion recognition is the definition of features that can capture emotion cues in the data. There is no agreement on the set of features that is the most powerful in distinguishing between emotions [5, 6, 7], and, in speech emotion recognition, this challenge is more aggravated, due to the acoustic variability introduced by speakers, speaking styles, and speaking rates.

Most current studies propose training deep learning models to extract those feature sets from the data [8, 9]. Although these approaches have yielded satisfactory results, two problems remain. First, the training may easily lead to overfitting, since these models are usually trained from scratch using low-level data representations, and emotion recognition datasets are known to have a limited amount of data. Second, it is known that deep learning architectures may learn from superficial cues [10], which makes us question if current models can actually capture emotion information in the data.

We address these issues by challenging commonly used low-level data representations in speech-based and text-based emotion recognition studies, and by incorporating high-level features to our method. We define as low-level the features obtained from feature engineering, and we define as high-level the generic features extracted from deep learning approaches.

We propose a cross-representation model for speech emotion recognition, in which we aim to reconstruct low-level mel-spectrogram speech representations from high-level wav2vec 2.0 ones, thus leveraging both representations. We choose wav2vec because they contain rich prosodic information [11]. Additionally, since we would like to capture a generic representation of emotion from speech, our model uses disentanglement representation learning techniques to eliminate speaker identity and phonetic variations in the data. In our method, we also define a CNN-based model for text-based emotion recognition on features acquired from Transformed-based models. We believe these features are better than the commonly used word2vec and GloVe features due to the Transformer's ability of modelling long contextual information in a sentence, which is necessary for emotion recognition. Finally, we combine the results from the speech-based and the text-based models to obtain the multimodal

emotion recognition results.

2. RELATED WORKS

2.1. Wav2vec 2.0

Wav2vec 2.0 [12] is a framework to obtain speech representations via self-supervision. The wav2vec model is trained on large amounts of unlabelled speech data, and then it is fine-tuned on labelled data for Automatic Speech Recognition (ASR). Wav2vec is composed of a feature encoder and a context network. The encoder takes a raw waveform as input, and outputs a sequence of features with stride of 20 ms and receptive field of 25 ms. These features encode the speech's local information, and they have a size of 768 and 1024 for the "base" and "large" versions of wav2vec, respectively. The feature sequence is then inputted to the Transformer-based context network, which outputs a contextualized representation of speech. In the "base" and "large" wav2vec, there are 12 and 24 Transformer blocks, respectively. Although the wav2vec 2.0 learned representations were originally applied to ASR, other tasks, such as speech emotion recognition, can also benefit from these representations [11, 13, 14].

2.2. Speech Emotion Recognition

During its early stage, most works on Speech Emotion Recognition (SER) proposed solutions based on Hidden Markov Models (HMM) [15], Support Vector Machines (SVM) [4, 16], or Gaussian Mixture Models (GMM) [5]. However, given the superior performance of deep learning on many speech-related tasks [17, 18], deep learning approaches for SER became predominant.

A problem characteristic to SER is the definition of appropriate features to represent emotion from speech [19]. Previous studies have attempted to extract emotion information from Mel-frequency cepstral coefficients (MFCC), pitch and energy [5, 6, 7]. However, recent studies showed that employing a weighted sum with learnable weights to combine the local and contextualized outputs of a pre-trained wav2vec 2.0 model yields better speech emotion recognition results [11].

2.3. Text Emotion Recognition

Current Text Emotion Recognition (TER) works use either features from an ASR model trained from scratch [9], or Word2Vec or GloVe [3] features. Such works yield good results, but, given the outstanding performance of Transformer-based models in various NLP tasks [20, 21, 22], it is natural to question commonly used representations for the TER task.

2.4. Disentanglement Representation Learning

Disentanglement Representation Learning aims to separate the underlying factors of variation in the data [23]. The idea is that, by disentangling these factors, we can discard the factors that are uninformative to the task that we would like to solve, while keeping the relevant factors. Disentanglement has been applied to image [24], video [25] and speech [26] applications. In speech-related works, it has been applied mainly to speech conversion and prosody transfer tasks [27, 28].

AutoVC [27] is an autoencoder that extracts a speaker-independent representation of speech content for speech conversion. A mel-spectrogram is inputted to the model's encoder, and the decoder reconstructs the spectrogram from the encoder's output and a speaker identity embedding. By controlling the encoder's bottleneck size, speaker identity information is eliminated at the encoder's bottleneck. SpeechFlow [26] builds upon AutoVC to disentangle speech into pitch, rhythm and content features.

Even though these works show impressive results in speech conversion, only few works attempt to disentangle speech for SER. [29] proposed an autoencoder to disentangle speech style and speaker identity from i-vectors and x-vectors, and they used the speech style embedding for SER. [30] used adversarial training to disentangle speech features into speaker identity and emotion features. These methods hold similarities with the SER method proposed in this paper, but our approach differs from previous works in that we explicitly eliminate speaker identity information from speech to obtain emotion features, and we perform experiments on the disentanglement property of these features.

3. METHODOLOGY

We propose a model to perform SER and a model to perform TER. The SER model takes as input wav2vec features, a mel spectrogram, speaker identity embeddings, and a phone sequence. All these features are extracted from the same speech segment, and the model outputs the probabilities for each emotion class, for the speech segment. The TER model takes as input text features extracted from an utterance's transcript, and outputs the probabilities for each emotion class, for the utterance. The SER and TER results are combined via score fusion to obtain the multimodal emotion class probabilities. Our proposed method, including the SER model, the TER model and the fusion approach, is depicted in Figure 1.

3.1. Speech Emotion Recognition

We propose an encoder-decoder model that takes wav2vec 2.0 features as input, and reconstructs the corresponding mel-frequency spectrogram. Our model has four main components: an encoder, a decoder, a phone encoder, and a classifier. The model is trained over speech segments of 96 frames, which is about 2 seconds long. These segments are randomly cropped from the speech utterances during training.

Our SER model is similar to AutoVC [27]. However, our model differs in three aspects. First, wav2vec features are the

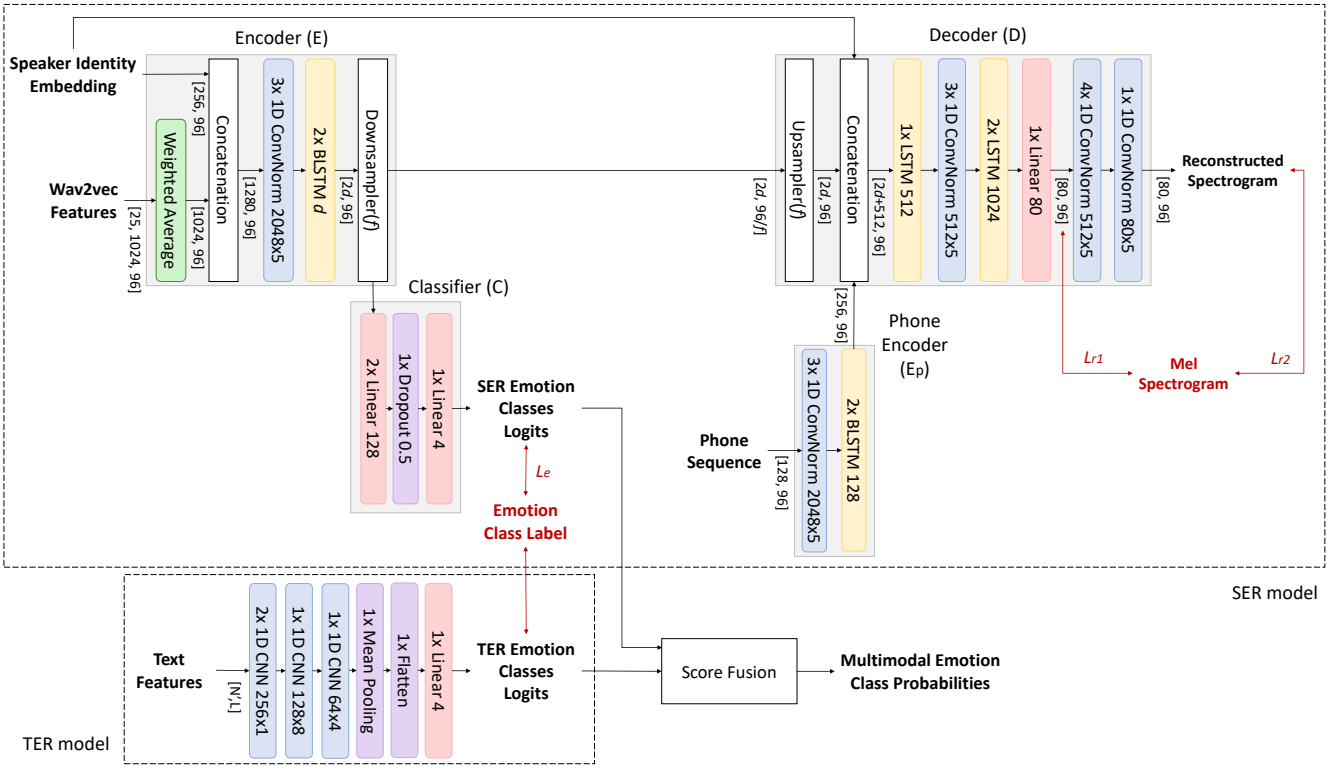

Fig. 1. Proposed method depicting the SER and TER models and the fusion approach. 1D ConvNorm layers are defined as a 1D CNN layer followed by batch normalization, and BLSTM layers are bidirectional LSTMs. The number of layers is shown as each block name's prefix. The number of filters F and kernels K in each ConvNorm and CNN layers are shown as the block name's suffix FxK. The number of neurons in LSTM, BLSTM, and Linear layers is shown as the blocks names' suffix.

acoustic input to our encoder. Second, we include an emotion classifier and an emotion loss to our method. Third, we define a phone encoder, whose output is inputted to the decoder.

3.1.1. *Wav2vec 2.0 Feature Extraction*

We extract the wav2vec features from a "large" wav2vec 2.0 model pre-trained on 60k hours of unlabelled speech data from the LibriVox dataset[1]. We take the features from the feature encoder's output and from the output of all the 24 Transformer layers in the context network. Thus, for each speech frame, there are 25 1024-dimensional wav2vec features.

3.1.2. *Speaker Identity Feature Extraction*

We extract speaker identity features with Resemblyzer [31][2], which is pre-trained on LibriSpeech [32], VoxCeleb1 [33] and VoxCeleb2 [34]. For each utterance, a 256-dimensional embedding is obtained to represent the speaker identity. For each speaker, we extract speaker identity features from 100 randomly selected utterances, and we take their average as the final identity embedding to represent the speaker.

[1]https://huggingface.co/facebook/wav2vec2-large-lv60

[2]We use the code in: https://github.com/resemble-ai/Resemblyzer

3.1.3. *Encoder*

A weighted average $\mathbf{h}_{\text{avg}}$, with trainable weights α, over the 25 wav2vec 2.0 features $\mathbf{h_i}$, is computed as described in [11]:

$$\mathbf{h}_{\text{avg}} = \frac{\sum_{i=1}^{25} \alpha_i \mathbf{h_i}}{\sum_{i=1}^{25} \alpha_i}. \quad (1)$$

$\mathbf{h}_{\text{avg}}$ is then concatenated with the 256-dimensional speaker identity embedding frame by frame.

The BLSTM layers in the encoder have d neurons, and their output have a size of $[2d, 96]$ since we concatenate the layers' output in both forward and backward directions. d determines the size of the bottleneck, as it reduces the size of the features in the channel dimension. The downsampler operation [27] takes as input an array of size $2d$ for each speech frame, and returns the arrays taken at every f frames. Thus, this operation reduces the temporal dimension of the feature array, by the downsampling factor f. The encoder outputs a feature array of size $[2d, 96/f]$, in which d and f control the bottleneck dimension. By controlling the size of the bottleneck, we would like to obtain a disentangled speech representation, that contains emotion information, but that does not contain speaker identity or phonetic information.

3.1.4. Decoder

At the decoder, the encoder's features are upsampled so that their size is the same as before the downsampling operation, by repeating each feature in the temporal dimension f times. Since the encoder's features contain only emotion information, the decoder takes as input not only the encoder's output, but also speaker identity embeddings and phone sequence embeddings to be able to reconstruct the spectrogram.

The output from the decoder's linear layer is a feature array of size 80 for each speech frame, which represents a mel-spectrogram of the speech segment. These features are compared with the ground-truth mel-spectrogram by means of a reconstruction loss L_{r1}, which is used to update all the model's parameters. We also compute the reconstruction loss L_{r2} between the decoder's output and the same ground-truth mel-spectrogram. L_{r1} and L_{r2} are computed as

$$L_r = \frac{1}{M}\sum_{k=1}^{M}(x_k - y_k)^2, \tag{2}$$

in which M is the training batch size, x_k is the k-th feature element in the batch outputted by the model, and y_k is the corresponding ground-truth for x_k.

3.1.5. Phone Encoder

The phone encoder takes as input a sequence of phone embeddings, and outputs a representation for the whole phone sequence. We follow two steps to obtain these phone embeddings. First, for each utterance, we extract the phone alignment information from the speech signal and its corresponding text transcript, by using the Gentle aligner[3]. Second, we obtain the phone sequence from the phone alignment information, by determining the longest phone for each frame. We define an id number to each phone, and we also assign ids to silence, not-identified phones, and to each special token in the dataset (e.g. "[LAUGHTER]"), totalling 128 distinct phone ids represented as one-hot embeddings.

3.1.6. Classifier

The classifier encourages the encoder's output to contain emotion information. We compute the cross-entropy loss L_e between the emotion label c and the softmax of the logits z outputted by the classifier as

$$L_e(z, c) = -\log\left(\frac{\exp(z[c])}{\sum_j \exp(z_j)}\right). \tag{3}$$

3.1.7. Training and inference

The objective function to be minimized during training is given as the sum between L_{r1}, L_{r2} and L_e.

[3]https://github.com/lowerquality/gentle

During inference, the softmax function is applied to the model's outputted emotion classes logits array to obtain the emotion class probabilities. The class with the highest probability is selected as the final classification result. The model takes as input features from a speech segment of 96 frames. Thus, to obtain an utterance-level prediction, we first compute the emotion class probabilities every 96 consecutive frames in an utterance (without overlap), zero-padding as necessary, and we take the average of these segment-level probabilities as the final utterance-level probability.

3.2. Text Emotion Recognition

[35] shows that the TER task can benefit from processing embeddings of all text tokens before performing the emotion classification. Inspired by these results, we propose a CNN-based model to process all the token's embeddings in an utterance, extracted with Transformer-based models.

We extract a text representation of shape $[N, L]$ for each utterance with pre-trained Transformer-based models, in which N is the number of tokens in the utterance excluding special tokens, and L is the size of each token's feature. We zero-pad the text representation so that the input to the TER model have size $[N', L]$, in which N' is the maximum number of tokens found in an utterance of the dataset. These text features are processed with the TER model illustrated in Figure 1, which is trained on the cross-entropy loss defined in Equation (3). Similar to our SER model, during inference, the softmax function is applied to the TER model's logits array to obtain the emotion class probabilities. However, differently from the SER model, the output from the TER model represents the utterance-level emotion classification result.

3.3. Multimodal Emotion Recognition

The speech-based utterance-level probabilities p_s and the probabilities outputted from the text model p_t for the same utterance are combined as

$$p_f = w_1 \cdot p_s + w_2 \cdot p_t, \tag{4}$$

in which p_f is the fused probability, and w_1 and w_2 are fixed weights assigned to the speech and text modalities, respectively. The weights determine the degree of contribution of each data modality to the fused probability, and the emotion classification result for an utterance corresponds to the emotion class with the highest fused probability.

4. DATASET

We utilize the Interactive Emotional Dyadic Motion Capture (IEMOCAP) [36] dataset to evaluate our method. Given the amount of data and the phonetic and semantic diversity of its utterances, the IEMOCAP dataset is considered well-suited for speech-based and text-based emotion recognition.

Table 1. UA (%) results for the SER task on a model with a small bottleneck and a model with a large bottleneck.

Model	Fold					Avg ± std
	1	2	3	4	5	
Small	67.9	71.7	67.3	72.2	71.6	$\mathbf{70.1 \pm 2.3}$
Large	59.6	73.6	66.0	71.3	70.0	68.1 ± 5.5

There are 10 actors in this dataset, whose interactions are organized in 5 dyadic sessions, each with an unique pair of male and female actors. The dataset contains approximately 12 hours of audiovisual data, which is segmented into speech turns (or utterances). Each utterance is labelled by three annotators.

Following previous works [3, 8, 9], we consider only the utterances which are given the same label by at least two annotators, and we merge the utterances labelled as "Happy" and "Excited" into the "Happy" category. We further select only the utterances with the labels "Angry", "Neutral", "Sad" and "Happy", resulting in 5,531 utterances, which is approximately 7 hours of data. We utilize only the speech data, the transcripts, and the labels.

5. TRAINING CONFIGURATION

We perform a leave-one-session-out cross-validation in all our experiments. We report our results in terms of Weighted Accuracy (WA) and Unweighted Accuracy (UA). WA is equivalent to the average recall over all the emotion classes and UA is the fraction of samples correctly classified.

All models are implemented in PyTorch, and, in every training experiment, we use the Adam optimizer with learning rate 10^{-4}, and with the default exponential decay rate of the moment estimates. The SER models are trained with a batch size of 2 for 1 million iterations. The TER models are trained with a batch size of 4 for 412,800 iterations.

6. SPEECH EMOTION RECOGNITION

6.1. Emotion Recognition Experiments

We perform SER with two bottleneck configurations for the encoder, "Small" and "Large", which have respective bottleneck dimension d equal to 8 and 128, and respective downsampling factor f set as 48 and 2. The utterance-level SER results are presented in Table 1.

Table 1 indicates that the "Small" configuration performs better in the SER task when compared to the "Large" model. We compare the results obtained with the "Small" model with the current state-of-the-art in Table 2.

We further evaluate whether inputting the wav2vec embeddings is advantageous to SER. We train our model with the same parameters as the "Small" configuration, but with a mel-spectrogram as input instead of the wav2vec features.

Table 2. Comparison of our SER results with current works in terms of UA(%) and WA (%).

Model	UA	WA
GRU+Context [37]	68.3	66.9
Self-Attn+LSTM [3]	55.6	-
BLSTM+Self-Attn [9]	57.0	55.7
Transformer [38]	-	64.9
CNN+Feat-Attn [39]	66.7	-
wav2vec+CNN [11]	-	67.9
Ours (Small)	**70.1**	**70.7**

Table 3. UA (%) results for the speaker identification task on features extracted with the "Small" and "Large" models.

Model	Fold					Avg ± std
	1	2	3	4	5	
Small	18.4	20.6	19.4	15.9	13.6	17.6 ± 2.8
Large	25.6	19.8	24.5	23.4	21.4	22.9 ± 2.3

Our results achieved an UA of 50.4% on the 5-fold cross-validation, which is 19.7% worse than of the model with wav2vec features as input, in terms of absolute accuracy. Therefore, we can conclude that the learned weighted average of the wav2vec embeddings is a better representation of speech for SER on the IEMOCAP dataset when compared to the traditional mel-frequency spectrograms.

6.2. Disentanglement Experiments

We train 4-linear-layer (with, from the input to the output, 2048, 1024, 1024, and 8 neurons) classifiers on the obtained speech representations to solve the speaker identification task. Our goal is to see if the obtained emotion features contain speaker identity information. Ideally, we would like our features to be speaker-independent, and to hold a generic emotion representation that could be used across speakers.

We train the classifiers on a 5-fold cross-validation, but we define the folds differently for this experiment. We randomly separate 80% of each speaker's data for training, and the remaining 20% for test. The folds have speaker dependent train and test sets, and each of them contains the data of only 4 sessions. We train the classifiers on a cross-entropy loss. Table 3 summarizes the speaker identity recognition results.

Table 3 suggests that the features extracted with the "Large" model contain more information about speaker identity than the ones from the "Small" model. Overall, from the results in Tables 3 and 1, we can see that the features extracted with the "Small" model can achieve a better SER accuracy and a worse speaker identity accuracy when compared to the features extracted with the "Large" model. This result suggests that the bottleneck size can lead to a disentanglement of factors in speech, which makes the SER task easier.

Table 4. 5-fold cross-validation UA (%) results for TER on input features extracted from different models. (c = "cased", u = "uncased", uwm = "uncased with whole word masking")

Model	Avg $\pm$ std
ALBERT	62.3 ± 2.3
BERT$_c$	65.5 ± 3.3
BERT$_u$	$\mathbf{66.1 \pm 2.1}$
BERT$_{uwm}$	65.8 ± 2.6
ELECTRA	56.6 ± 3.8
RoBERTa	64.1 ± 3.5
XLNet$_c$	58.1 ± 3.6

Table 5. Comparison of our TER results with current works in terms of UA(%) and WA (%).

Model	UA	WA
BERT+Attn+Context [40]	71.9	71.2
BERT+Attn [40]	64.8	62.9
BLSTM+Self-Attn [9]	63.6	63.7
Self-Attn+LSTM [3]	65.9	-
Ours (BERT uncased)	**66.1**	**67.0**

6.3. Discussion

We believe that our SER model achieves better results when compared to previous methods due to three factors. First, we use high-level speech representations as the input to our model, which, apart from our work, is only done by [11] and [38]. Second, we are careful in analyzing the type of information encoded in the features obtained by our model, which makes the features have a certain level of disentanglement as shown in Section 6.2. Third, our model can leverage both high-level and low-level features since it is trained to reconstruct spectrograms from wav2vec features.

7. TEXT EMOTION RECOGNITION

We train the TER model with input features extracted from different Transformer-based models[4]. We use the "large" version of all these models, which output a 1024-dimensional feature for each token. The TER results are shown in Table 4 for different feature extractors, and we compare our best results with the current state-of-the-art in Table 5.

From Table 5, we can see that our method achieves better results than previous works, except for [40], which uses context information (i.e., features from succeeding and preceding utterances). Our model differs from previous works in that we do not use recurrent neural networks or self-attention, and we benefit from the text representations learned by Transformer-based models trained on large text corpora. We believe our model achieved good results due to BERT's deep features, and to the ability of our model's 1D CNN layers to extract temporal information from the sequence of token's features.

[4]The trained models can be found at https://huggingface.co/models

Table 6. Comparison of our multimodal results with current works in terms of UA(%) and WA (%).

Model	UA	WA
BERT+Attn+Context [40]	76.1	77.4
LAS-ASR [8]	66.0	64.0
ASR-SER [9]	69.7	68.6
CMA+Raw waveform [3]	72.8	-
Ours ($w_1 = 0.6$, $w_2 = 1$)	**73.0**	**73.5**

8. MULTIMODAL EMOTION RECOGNITION

We combine the results from our best speech and text models by experimenting with different weight values w_1 and w_2. Our best multimodal results are acquired when $w_1 = 0.6$ and $w_2 = 1$, and they are reported in Table 6.

This result shows that, when combining the speech and the text results, it is better to give less importance to the speech model's result, even though its accuracy is higher than the text model's. We believe this may be related to the confidence in which the TER and the SER models obtain their scores, but further investigation is required.

By comparing our results in Tables 1, 4, and 6, we can conclude that the solution to the emotion recognition task benefits from combining different types of data, since our multimodal result is better than our speech-only and text-only results. Our multimodal approach gives better results than current works except for [40], which uses the context information. We attribute the reason of our good results to the fact that our unimodal models outperform other unimodal models, and not to our choice of fusion method. We believe we could achieve better results with a more sophisticated fusion approach or by jointly training the speech and text modalities.

9. CONCLUSION

We proposed a cross-representation encoder-decoder model inspired in disentanglement representation learning to perform SER. Our model leverages both high-level wav2vec features and low-level mel-frequency spectrograms, and it achieves an accuracy of 70.1% on the IEMOCAP dataset. We also used a CNN-based model that processes token's embeddings extracted with pre-trained Transformer-based models to perform TER, achieving an accuracy of 66.1% on the same dataset. We further combined the speech-based and the text-based results via score fusion, achieving an accuracy of 73.0%. Our speech-only, text-only and multimodal results surpassed current works', showing that emotion recognition can benefit from disentanglement representation learning, high-level data representations, and multimodalities.

10. REFERENCES

[1] Stephen E Levinson, "Continuously variable duration hidden markov models for automatic speech recognition," *Computer Speech & Language*, vol. 1, no. 1, pp. 29–45, 1986.

[2] Yequan Wang, Aixin Sun, Jialong Han, Ying Liu, and Xiaoyan Zhu, "Sentiment analysis by capsules," in *Proceedings of the 2018 world wide web conference*, 2018, pp. 1165–1174.

[3] DN Krishna and Ankita Patil, "Multimodal emotion recognition using cross-modal attention and 1d convolutional neural networks," *Proc. Interspeech 2020*, pp. 4243–4247, 2020.

[4] Yixiong Pan, Peipei Shen, and Liping Shen, "Speech emotion recognition using support vector machine," *International Journal of Smart Home*, vol. 6, no. 2, pp. 101–108, 2012.

[5] Aditya Bihar Kandali, Aurobinda Routray, and Tapan Kumar Basu, "Emotion recognition from assamese speeches using mfcc features and gmm classifier," in *TENCON 2008-2008 IEEE region 10 conference*. IEEE, 2008, pp. 1–5.

[6] Kun Han, Dong Yu, and Ivan Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Fifteenth annual conference of the international speech communication association*, 2014.

[7] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[8] Sung-Lin Yeh, Yun-Shao Lin, and Chi-Chun Lee, "Speech representation learning for emotion recognition using end-to-end asr with factorized adaptation," *Proc. Interspeech 2020*, pp. 536–540, 2020.

[9] Han Feng, Sei Ueno, and Tatsuya Kawahara, "End-to-end speech emotion recognition combined with acoustic-to-word asr model," *Proc. Interspeech 2020*, pp. 501–505, 2020.

[10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[11] Leonardo Pepino, Pablo Riera, and Luciana Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *Proc. Interspeech 2021*, pp. 3400–3404, 2021.

[12] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[13] Shamane Siriwardhana, Andrew Reis, Rivindu Weerasekera, and Suranga Nanayakkara, "Jointly Fine-Tuning "BERT-Like" Self Supervised Models to Improve Multimodal Speech Emotion Recognition," in *Proc. Interspeech 2020*, 2020, pp. 3755–3759.

[14] Manon Macary, Marie Tahon, Yannick Estève, and Anthony Rousseau, "On the use of self-supervised pre-trained acoustic and linguistic features for continuous speech emotion recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 373–380.

[15] Tin Lay Nwe, Say Wei Foo, and Liyanage C De Silva, "Speech emotion recognition using hidden markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.

[16] Yi-Lin Lin and Gang Wei, "Speech emotion recognition based on hmm and svm," in *2005 international conference on machine learning and cybernetics*. IEEE, 2005, vol. 8, pp. 4898–4901.

[17] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[18] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," in *9th ISCA Speech Synthesis Workshop*, 2016, pp. 125–125.

[19] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[21] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le, "Xlnet:

Generalized autoregressive pretraining for language understanding," *Advances in Neural Information Processing Systems*, vol. 32, pp. 5753–5763, 2019.

[22] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," in *International Conference on Learning Representations*, 2020.

[23] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*, MIT Press, 2016.

[24] Yu Liu, Fangyin Wei, Jing Shao, Lu Sheng, Junjie Yan, and Xiaogang Wang, "Exploring disentangled feature representation beyond face identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2080–2089.

[25] Emily L Denton et al., "Unsupervised learning of disentangled representations from video," in *Advances in neural information processing systems*, 2017, pp. 4414–4423.

[26] Kaizhi Qian, Yang Zhang, Shiyu Chang, Mark Hasegawa-Johnson, and David Cox, "Unsupervised speech decomposition via triple information bottleneck," in *International Conference on Machine Learning*. PMLR, 2020, pp. 7836–7846.

[27] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," *International Conference on Machine Learning*, 2019.

[28] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J Weiss, Rob Clark, and Rif A Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," *Proeedings of ICML*, 2018.

[29] Jennifer Williams and Simon King, "Disentangling style factors from speaker representations.," *Proc. Interspeech 2019*, pp. 3945–3949, 2019.

[30] Haoqi Li, Ming Tu, Jing Huang, Shrikanth Narayanan, and Panayiotis Georgiou, "Speaker-invariant affective representation learning via adversarial training," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7144–7148.

[31] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[32] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[33] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *Proc. Interspeech 2017*, 2017.

[34] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *Proc. Interspeech 2018*, 2018.

[35] Leonardo Pepino, Pablo Riera, Luciana Ferrer, and Agustín Gravano, "Fusion approaches for emotion recognition from speech using acoustic and text-based features," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6484–6488.

[36] 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, 2008.

[37] Srividya Tirunellai Rajamani, Kumar T Rajamani, Adria Mallol-Ragolta, Shuo Liu, and Björn Schuller, "A novel attention-based gated recurrent unit and its efficacy in speech emotion recognition," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6294–6298.

[38] Ruixiong Zhang, Haiwei Wu, Wubo Li, Dongwei Jiang, Wei Zou, and Xiangang Li, "Transformer based unsupervised pre-training for acoustic representation learning," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6933–6937.

[39] Shuiyang Mao, PC Ching, C-C Jay Kuo, and Tan Lee, "Advancing multiple instance learning with attention modeling for categorical speech emotion recognition," *Proc. Interspeech 2020*, pp. 2357–2361, 2020.

[40] Wen Wu, Chao Zhang, and Philip C. Woodland, "Emotion recognition by fusing time synchronous and time asynchronous representations," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6269–6273.