Integrating Emotion Recognition with Speech Recognition and Speaker Diarisation for Conversations

Wen Wu[1], Chao Zhang[2], Philip C. Woodland[1]

[1]Department of Engineering, University of Cambridge, Cambridge, UK
[2]Department of Electrical Engineering, Tsinghua University, Beijing, China

{ww368, pcw}@eng.cam.ac.uk; cz277@tsinghua.edu.cn

Abstract

Although automatic emotion recognition (AER) has recently drawn significant research interest, most current AER studies use manually segmented utterances, which are usually unavailable for dialogue systems. This paper proposes integrating AER with automatic speech recognition (ASR) and speaker diarisation (SD) in a jointly-trained system. Distinct output layers are built for four sub-tasks including AER, ASR, voice activity detection and speaker classification based on a shared encoder. Taking the audio of a conversation as input, the integrated system finds all speech segments and transcribes the corresponding emotion classes, word sequences, and speaker identities. Two metrics are proposed to evaluate AER performance with automatic segmentation based on time-weighted emotion and speaker classification errors. Results on the IEMOCAP dataset show that the proposed system consistently outperforms two baselines with separately trained single-task systems on AER, ASR and SD[1].

Index Terms: Automatic emotion recognition, automatic speech recognition, speaker diarisation, foundation model

1. Introduction

There has been much research work in automatic emotion recognition (AER) [1, 2, 3, 4, 5]. However, most AER systems operate on manually segmented utterances although manual segmentations are not generally available in practical use cases. Besides, automatic speech recognition systems trained on standard speech can give poor recognition performance on emotional speech [4, 6, 7].

This work proposes an integrated system for emotion recognition, speaker diarisation and speech recognition. Speaker diarisation is the process that detects speech regions of an audio recording and groups them into homogeneous segments according to the relative identity of the speaker. The outcome of speaker diarisation is referred to as segmentation in this paper. As illustrated in Figure 1, the system takes the audio recording of a dialogue as input. It automatically diarises the dialogue into segments associated with different speakers, transcribes the audio segments into text, and predicts the speaker's emotion state.

The model contains four downstream heads for voice-activity-detection (VAD), speaker indentity (SI) extraction, automatic speech recognition (ASR), automatic emotion recognition and an encoder shared by all downstream heads. Speaker

diarisation is achieved by the VAD head and the SI head. The VAD head classifies each frame into speech or non-speech. The SI head learns speaker embeddings that capture the characteristics of each speaker. Based on the predicted segmentation, the ASR head converts each speech segment into text and the AER head predicts the emotion states of the corresponding speaker. A speech foundation model [8] (i.e. WavLM [9]) is used as the shared encoder which takes raw speech waveform as input. Foundation models are pre-trained on large amount of unlabelled data and can help handle the data sparsity issue in AER tasks. The downstream heads take the weighted sum of intermediate hidden states from the shared encoder as input. Each head has an individual set of weights, which are trained jointly with the shared encoder and the downstream heads.

Furthermore, this work also proposes two metrics to evaluate the emotion recognition performance with automatic segmentation: the time-weighted emotion error rate (TEER) and the speaker-attributed time-weighted emotion error rate (sTEER). To the best of our knowledge, this is the first work that considers emotion recognition with automatic segmentation and integrates emotion recognition, speech recognition and speaker diarisation into a jointly-trained model.

The rest of the paper is organised as follows. Section 2 summarises related work. Section 3 introduces the proposed system and the TEER and sTEER metrics. Section 4 presents the experimental setup while the results and analysis are shown in Sections 5 and 6 respectively, followed by the conclusions.

2. Related work

Multi-task training and transfer learning have been investigated to improve the AER performance with ASR transcriptions. Feng et al. [10] used an attention-based encoder-decoder model for ASR and combined the states of the ASR decoder with the acoustic features for AER. Li et al. [11] encode the raw waveform with the Wav2Vec 2.0 model [12] for ASR and separately as MFCCs for AER. The ASR outputs were fused into the pipeline for joint training with AER. Zhou et al. [13] fine-tuned an ASR model on emotional speech for emotion recognition. Cai et al. [14] proposed a multi-task learning framework that fed the output of a Wav2Vec 2.0 model to an ASR head and an AER head each containing one fully-connected layer. Two heads were simultaneously trained while only the AER head was kept during inference. Heusser et al. [15] trained ASR from audio, AER from audio and from text independently and fine-tuned the combined sub-models. Ghriss et al. [16] pre-trained the AER model by ASR which is trained jointly with a sentiment classifier.

Apart from combining AER and ASR, Velichko et al. [17] proposed a hierarchical framework to predict gender, emotion,
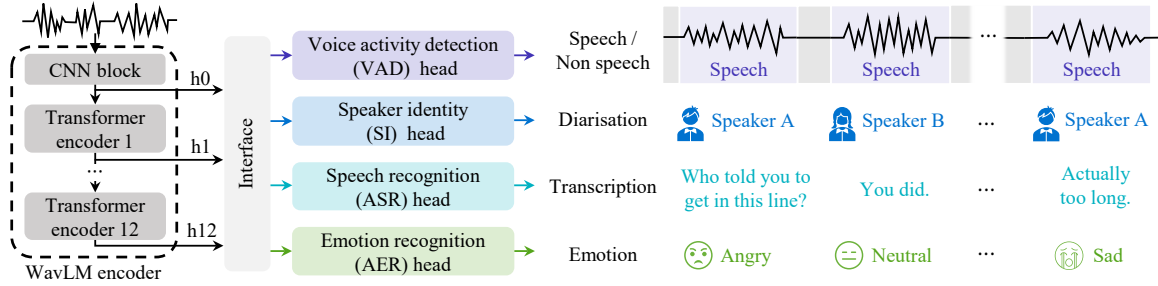
Figure 1: *Overview of the proposed integrated system. Taking a dialogue as input, the VAD and SI head perform automatic segmentation. The ASR and AER head recognise the text and emotion based on the predicted segments.*

and deception in a cascaded way. So far, combining speaker diarisation with emotion recognition and evaluating emotion recognition when using automatic segmentation hasn't been widely studied.

3. Proposed approach

The structure of the proposed system is shown in Figure 2, which consists of a shared encoder, an interface and four downstream heads.

3.1. Shared encoder and interface

The WavLM Base model [9] is used as the shared encoder in this paper, which takes the raw waveform as input. It contains a convolutional neural network (CNN) block as the feature extractor and 12 Transformer [18] encoder blocks. The output of the encoder is a frame sequence with a frame shift of 20 ms.

All the semantic and non-semantic information co-exist in the same speech signal. Research [9, 19, 20] has shown that intermediate representations of such foundation models contain different levels of information. The weighted sum of embeddings from the CNN block and each Transformer encoder block is used as the input to the downstream tasks to exploit this property. Each downstream head has its own set of weights which are trainable.

3.2. Downstream heads

The VAD head consists of 3 fully-connected (FC) layers with a hidden dimension of 256 and leaky ReLU activation and an output FC layer with Softmax activation which performs frame-level speech/non-speech detection. The SI head consists of an X-vector speaker embedding model [21], which generates one speaker embedding for each input sequence. The speaker embedding is fed into an FC layer with Softmax activation for speaker classification during training. During testing, spectral clustering is conducted based on speaker embeddings to produce speaker diarisation. The ASR head consists of 4 Bi-LSTM layers [22] with dimension of 256, followed by an FC layer for token prediction. A vocabulary of 29 graphemes is used including 26 letters in English plus a few punctuation characters. The AER head consists of 2 Transformer encoder layers of dimension 256. The representations are average-pooled along the time axis before being fed into an FC layer with Softmax activation for emotion classification. Six emotion classes are used: "happy", "sad", "angry", "neutral", "other", "no majority agreement (NMA)". "NMA" denotes that the human annotators don't have a majority agreed emotion class label for this utterance.
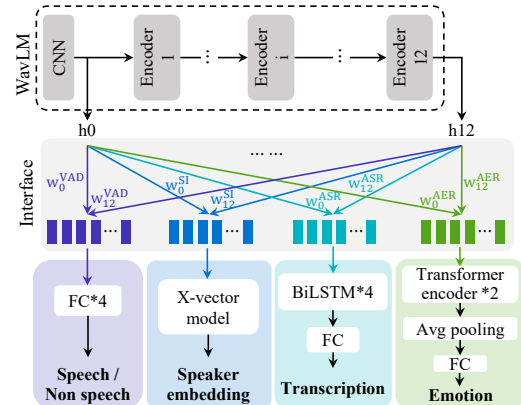

Figure 2: *Structure of the proposed integrated system.*

3.3. Multi-task training loss

The ASR head is trained using the CTC loss [23] and the other three heads were trained using the cross entropy loss. The shared encoder and four downstream heads are jointly trained using a multi-task loss:

$$\mathcal{L}_{\text{Total}} = \epsilon_{\text{VAD}}\mathcal{L}_{\text{VAD}} + \epsilon_{\text{SI}}\mathcal{L}_{\text{SI}} + \epsilon_{\text{ASR}}\mathcal{L}_{\text{ASR}} + \epsilon_{\text{AER}}\mathcal{L}_{\text{AER}} \quad (1)$$

where $\epsilon_{\text{VAD}}, \epsilon_{\text{SV}}, \epsilon_{\text{ASR}}, \epsilon_{\text{AER}}$ are coefficients that are set manually to keep the weighted loss of the three head in the same scale.

3.4. Training and testing procedures

Teacher forcing is applied during training using reference segmentations. The system takes segmented utterances as input. The encoder and downstream heads are trained jointly using the multi-task loss defined in Eqn 1. The segmented utterances can contain silence at the beginning, between words and at the end. The VAD head is trained based on the intra-utterance silence.

During testing, dialogues are input into the system. The system can benefit from knowing the context. A sliding window of 3 s length and 1 s overlap is applied to the dialogue. VAD is performed on each window. To avoid overlapped regions being counted twice at the output, the results of the middle second was kept for each window. This is equivalent to taking previous 1 s and future 1 s as context when making a prediction on the current 1 s of audio data. Post-processing is applied to the VAD predictions, so speech/non-speech regions shorter than 0.25 s are removed. A smaller sliding window of 1 s length and 0.5 s overlap is then applied to the detected speech regions. The SI head extracts a single speaker embedding from each window. Spectral clustering is used based on the speaker embeddings which groups segments from the same speaker together, thus

producing an automatic segmentation. Based on the automatic segmentation, the ASR and AER heads takes each segment as input and predict the text and emotion respectively.

3.5. Evaluating emotion classification performance with automatic segmentation

The segments predicted by the system can have different start and end times to the reference segments. Therefore, the classification accuracy is no longer sufficient to evaluate the performance of emotion classification in this case since it cannot handle the alignment between segments. This paper, therefore, proposes the time-weighted emotion error rate (TEER) in order to evaluate the AER performance given non-oracle segmentations. The TEER is computed as follows:

$$\text{TEER} = \frac{\text{MS} + \text{FA} + \text{CONF}_{\text{emo}}}{\text{TOTAL}} \qquad (2)$$

where missed speech (MS) is the duration of speech incorrectly classified as non-speech, false alarm speech (FA) is the duration of non-speech incorrectly classified as speech, confusion (CONF_{emo}) is the audio duration where emotion is wrongly classified, and TOTAL is the sum of the reference speech duration for all utterances.

Furthermore, the speaker-attributed TEER (sTEER) is proposed which expects the system to accurately predict both the speaker and their emotion. The sTEER is computed as follows:

$$\text{sTEER} = \frac{\text{MS} + \text{FA} + \text{CONF}_{\text{emo+spk}}}{\text{TOTAL}} \qquad (3)$$

CONF_{emo} in Eqn 2 is replaced by $\text{CONF}_{\text{emo+spk}}$, which is the duration where either speaker or emotion is wrong. sTEER reflects the overall performance of both speaker diarisation and emotion classification.

4. Experimental setup

4.1. Dataset

The benchmark IEMOCAP dataset [24] was used which consists of approximately 12 hours of English speech including 5 dyadic conversational sessions. There are in total 151 dialogues including 10,039 utterances. IEMOCAP provides the time-stamp of each utterance in a dialogue as well as word-level alignments of each utterance. The alignments show that 40% frames in the segmented utterances are silence. Each utterance was annotated by three human annotators. Sentences that don't have majority agreed emotion label from the annotators accounts for a 25% of the dataset.

AER in this paper uses a six-way classification setup. Emotion class "excited" is merged with "happy". All sentences with emotion label other than " happy", "sad", "angry", "neutral" are grouped into class "others". Sentences that don't have a majority agreed emotion label from the annotators are grouped into the sixth class "NMA". Speaker exclusive leave-one-session-out five-fold cross validation (CV) are performed and the average results are reported. Speakers in the test set are unseen in the training and validation set and it is ensured that utterances from the same dialogue are either all in the training set or all in the validation set.

4.2. Evaluation metrics

The false alarm rate (FAR) and missed rate (MSR) were used to evaluate the VAD performance. FAR computes the ratio of

the number of non-speech frames mispredicted as speech to the total number of speech frames. MSR computes the ratio of the number of speech frames mispredicted as non-speech to the total number of speech frames. Diarisation error rate (DER) was used to evaluate the performance of diarisation which maps the predicted relative speaker identity to the true speaker identity and measures the fraction of time not attributed correctly to a speaker or to non-speech. Overlapped speech was considered when computing DER. Since manual annotations cannot be precise at the audio sample level, it is common to remove from evaluation a forgiveness collar around each segment boundary. Unless otherwise mentioned, a collar of $0.25\,\text{s}$ was applied when evaluating with automatic segmentation.

Word error rate (WER) and classification accuracy (Acc_{emo}) were used to evaluate the performance of speech recognition and emotion classification respectively with oracle segmentation. With automatic segmentation, the concatenated minimum-permutation word error rate (cpWER) [25] was used to evaluate the ASR system performance which concatenates utterances of the same speaker and computes the WER. The sTEER and TEER were used to evaluate the AER system which have been defined in Section 3.5.

4.3. Training specification

The shared encoder was initialised with the publicly available WavLM Base+ model[2]. It was fine-tuned jointly with the downstream head while the CNN feature extractor of the WavLM model was frozen during fine-tuning. Speed perturbation was applied to the ASR and SI heads. For each epoch, the speed of each waveform was randomly adjusted to 0.95 or 1.05 of the original speech or remain unchanged. Speed perturbation was not applied to AER since speed is an important clue for emotion detection. The model was implemented using the SpeechBrain toolkit [26]. The system was trained using Adadelta optimiser with the Newbob learning rate scheduler. Scaling coefficients ϵ_{VAD} and ϵ_{SI} were set to 1.2 while the other two were set to 1. For each fold in 5-CV, training took around 5 hours on an NVIDIA A100 GPU and five checkpoints with the lowest validation loss were averaged after training for testing.

5. Results

The performance of the SI, ASR and AER heads were first evaluated with oracle segmentations in Section 5.1. The complete system was then evaluated with automatic segmentation in Section 5.2. The proposed system was compared to two baseline models:

- "Baseline-ref": Reference models which have been trained on other large datasets. The reference ASR model was pre-trained using 100 hours of LibriSpeech [27] training data, which has a WER of 5.64% on "test-clean" set and 12.15% on "test-other" set. The reference speaker embedding model[3] was pre-trained on Voxceleb 1.0 [28]. A VAD module [29] pre-trained on the augmented multi-party interaction (AMI) meeting corpus [30] was used as the reference baseline for VAD, which has 2.1% FAR and 4.7% MSR on the AMI eval set. No reference model was used for AER since we are evaluating on the emotion dataset. The reference system is a cascade of the reference models in the order of VAD, SI, ASR.

- "Baseline-frzn": The shared encoder was frozen during train-

---

[2]https://huggingface.co/microsoft/wavlm-base-plus
[3]https://huggingface.co/microsoft/wavlm-base-plus-sv

Table 1: *Results with reference segmentation. Collar was set to 0 when computing DER since oracle segmentation was assumed. '↑' denotes the higher the better, '↓' denotes the lower the better. Best results of each column are shown in bold.*

	%Acc$_{emo}$↑	%WER↓	%DER↓
Baseline-ref	/	33.29	1.10
Baseline-frzn	44.44	31.43	0.40
Proposed	**49.49**	**24.61**	**0.30**

Table 2: *VAD and speaker diarisation results.*

	%FAR↓	%MSR↓	%DER↓
Baseline-ref	5.15	1.30	8.20
Baseline-frzn	3.14	1.16	7.04
Proposed	**2.91**	**1.06**	**6.87**

ing. In this case, the four downstream heads were independent of each other and the multi-task loss becomes equivalent to training each head separately.

5.1. Oracle segmentation

In this section, utterances based on the reference segmentation were used as input to the SI, ASR and AER heads. The results are shown in Table 1. The ASR head with a frozen encoder reduced the WER on IEMOCAP to 31.43% and the SI head with frozen encoder reduced the DER to 0.40%. The proposed system with shared encoder jointly fine-tuned with downstream heads further reduced the WER and DER to 24.61% and 0.30% respectively. The 6-way emotion classification accuracy increased from 44.44% to 49.49%. The proposed integrated system outperforms the baselines for all three heads. Fine-tuning the pre-trained encoder on emotion data helps to adapt it to the specific domain, while sharing the encoder between the four downstream heads helps to capture general information relevant to the domain and avoids overfitting to trivial patterns, especially given the scarcity of data.

5.2. Automatic segmentation

The VAD performance and diarisation results based on VAD predictions are summarised in Table 2. The proposed system produced the best results on both VAD and diarisation.

ASR and AER were conducted using the diarisation output. As shown in Table 3, the proposed integrated system reduced cpWER by relative 12% compared to the baselines. sTEER is slightly higher than TEER as it takes speaker prediction error into account. The proposed system outperforms the single AER head in both emotion metrics, showing its superior performance for emotion recognition with automatic segmentation.

Table 3: *Speaker-attributed ASR and AER performance under automatic segmentation. Best results shown in bold.*

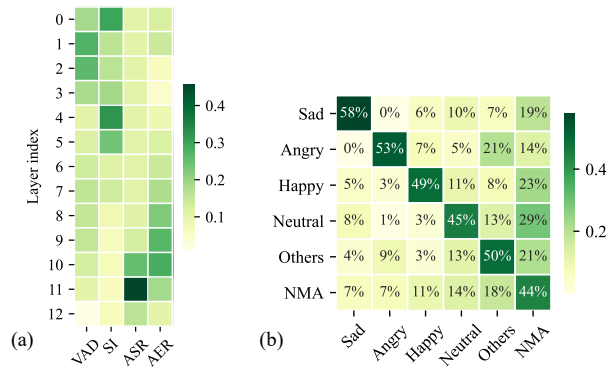|	%cpWER↓	%sTEER↓	%TEER↓
Baseline-ref	43.76	/	/
Baseline-frzn	41.19	69.49	68.70
Proposed	**36.20**	**66.03**	**65.17**

Figure 3: *(a) Weights of the interface for different downstream heads. (b) Confusion matrix of 6-way emotion recognition.*

6. Discussion and analysis

6.1. Trainable weights of the interface

The trainable weights for the four downstream heads are plotted in Figure 3(a). As can be seen, layer 0 and layer 4 are particularly useful for extracting speaker information. Layer 8-10 are more effective for AER and layer 11 contains most text information. This shows a similar pattern to previous findings [9, 19] that block-wise evolution of intermediate representations of a foundation model follows an acoustic-linguistic hierarchy, where the lower layers encode speaker-related information and higher layers encode phonetic/semantic information.

6.2. Confusion matrix of 6-way emotion classification

Based on the 6-way predictions, 4-way classification accuracy considering "happy", "sad", "angry", "neutral" is 73.96%, which is better than the results of Wav2Vec 2.0 Base (63.43%) and WavLM Base+ (68.65%) from the SUPERB leaderboards [20]. The class "NMA" is relatively easily confused, as shown by the 6-way confusion matrix of in Figure 3(b). For utterances classified as "NMA", the human annotators gave different emotion class labels and didn't reach majority agreement. These utterances may contain ambiguous emotions, mixed emotions, or emotions that tend to confuse the annotators. Among the other five classes, "angry" is the least likely to be confused with "NMA" probably because "angry" is relatively less ambiguous. By contrast, "neutral" is more likely to be wrongly predicted as "NMA", possibly because neutral emotions are relatively weak and human annotators are likely to disagree due to subjective perception.

7. Conclusions

This paper proposes a system that integrates emotion recognition with speech recognition and speaker diarisation in a jointly-trained model. The system investigates emotion recognition with automatic segmentation to address the issue of lacking manual segmentation in practical applications. The system also improves recognition performance on emotional speech by 12% reduction in relative word error rate with automatic segmentation. Time-weighted emotion error rate and speaker-attributed time-weighted emotion error rate were proposed to evaluate emotion classification performance when segmentation is non-oracle. Although the benchmark dataset used in this paper contains only dyadic conversations, the proposed method can also be applied to multi-party conversations.

8. References

[1] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *Proc. ICASSP*, Vancouver, Canada, 2013.

[2] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," *IEEE Intelligent Systems*, vol. 33, no. 6, pp. 17–25, 2018.

[3] P. Liu, K. Li, and H. Meng, "Group gated fusion on attention-based bidirectional alignment for multimodal emotion recognition," in *Proc. Interspeech*, Shanghai, China, 2020.

[4] W. Wu, C. Zhang, and P. C. Woodland, "Emotion recognition by fusing time synchronous and time asynchronous representations," in *Proc. ICASSP*, Toronto, Canada, 2021.

[5] W. Wu, C. Zhang, and P. C. Woodland, "Distribution-based emotion recognition in conversation," in *Proc. SLT*, Doha, Qatar, 2022.

[6] S. Sahu, V. Mitra, N. Seneviratne, and C. Y. Espy-Wilson, "Multimodal learning for speech emotion recognition: An analysis and comparison of asr outputs with ground truth transcription." in *Proc. Interspeech*, Graz, Austria, 2019.

[7] R. Fernandez, "A computational model for the automatic recognition of affect in speech," Ph.D. dissertation, Massachusetts Institute of Technology, 2004.

[8] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[9] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[10] H. Feng, S. Ueno, and T. Kawahara, "End-to-end speech emotion recognition combined with acoustic-to-word asr model." in *Proc. Interspeech*, Shanghai, China, 2020.

[11] Y. Li, P. Bell, and C. Lai, "Fusing ASR outputs in joint training for speech emotion recognition," in *Proc. ICASSP*, Singapore, 2022.

[12] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2Vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, Conference held virtually, 2020.

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

[14] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech Emotion Recognition with Multi-Task Learning," in *Proc. Interspeech*, Brno, Czech, 2021.

[15] V. Heusser, N. Freymuth, S. Constantin, and A. Waibel, "Bimodal speech emotion recognition using pre-trained language models," *arXiv preprint arXiv:1912.02610*, 2019.

[16] A. Ghriss, B. Yang, V. Rozgic, E. Shriberg, and C. Wang, "Sentiment-aware automatic speech recognition pre-training for enhanced speech emotion recognition," in *Proc. ICASSP*, Singapore, 2022.

[17] A. Velichko, M. Markitantov, H. Kaya, A. Karpov *et al.*, "Complex paralinguistic analysis of speech: Predicting gender, emotions and deception in a hierarchical framework," *Proc. Interspeech*, 2022.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, Long Beach, US, 2017.

[19] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *Proc. ASRU*, Cartagena, Colombia, 2021.

[20] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech*, Brno, Czech, 2021.

[21] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*, Hyderabad, India, 2018.

[22] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, pp. 2673 – 2681, 12 1997.

[23] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proc. ICML*, Pittsburgh, US, 2006.

[24] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Provost, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[25] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, and D. Snyder, "CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *Proc. CHiME*, San Francisco, US, 2020.

[26] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, and J. Chou, "SpeechBrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[27] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, South Brisbane, Australia, 2015.

[28] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[29] G. Sun, C. Zhang, and P. C. Woodland, "Combination of deep speaker embeddings for diarisation," *Neural Networks*, vol. 141, pp. 372–384, 2021.

[30] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, and G. Lathoud, "The AMI meeting corpus: A pre-announcement," in *Proc. Machine Learning for Multimodal Interaction*, Edinburgh, UK, 2005.