INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Speech Representation Learning for Emotion Recognition Using End-to-End ASR with Factorized Adaptation

Sung-Lin Yeh, Yun-Shao Lin, Chi-Chun Lee

Department of Electrical Engineering, National Tsing Hua University
MOST Joint Research Center for AI Technology and All Vista Healthcare
ff936tw@gapp.nthu.edu.tw, astanley18074@gmail.com, cclee@ee.nthu.edu.tw

Abstract

Developing robust speech emotion recognition (SER) systems is challenging due to small-scale of existing emotional speech datasets. However, previous works have mostly relied on handcrafted acoustic features to build SER models that are difficult to handle a wide range of acoustic variations. One way to alleviate this problem is by using speech representations learned from deep end-to-end models trained on large-scale speech database. Specifically, in this paper, we leverage an end-to-end ASR to extract ASR-based representations for speech emotion recognition. We further devise a factorized domain adaptation approach on the pre-trained ASR model to improve both the speech recognition rate and the emotion recognition accuracy on the target emotion corpus, and we also provide an analysis in the effectiveness of representations extracted from different ASR layers. Our experiments demonstrate the importance of ASR adaptation and layer depth for emotion recognition.

Index Terms: speech emotion recognition, end-to-end ASR, acoustic representation, domain adaptation

1. Introduction

Speech emotion recognition (SER) is an important module in human-centered applications such as the development of personalized agents [1] and mental health assessment [2]. Recognizing emotion via speech signals involves developing algorithms that could mathematically characterize affective acoustic properties that vary with speakers, e.g., change in pitch or loudness in conversations. A majority of SER models rely on training recognition models using low-level handcrafted acoustic features such as a combination of pitch, shimmer, loudness, and MFCCs as inputs. These acoustic features are shown repeatedly to carry substantial emotional cues [3, 4, 5, 6, 7]. However, models trained on these low-level features are often difficult to generalize well to a variety of domains due to the small amount of available emotional speech data; in addition, these features are not accessible and relevant to most other downstream tasks within a given technological solution.

Recently, researchers have started to develop representation learning approaches for SER under transfer learning setting. [8, 9] use unsupervised representation learning to extract robust speaker-invariant features. Ghosh et al. [10] utilize representations transferred from valence and activation regression task to 4-class emotion recognition. Another recent promising representation learning approach for SER task is through the use of automatic speech recognition (ASR) systems. Speech representations derived from ASR have been shown to preserve rich information that can be used in many other speech-based recognition tasks [11]. In comparison to the size of available data in SER domain, data in ASR domain is much larger in scale. This means that the speech representations learned from ASR models could be more robust to different variations and at the same time encode rich content, such as semantic, phonemic information, and also emotional expressions in speech. In this work, we also adopt ASR-derived speech representations for emotion classification.

Previous works often use ASR-derived speech representations to address the issue of small-scale emotional speech database [12, 13, 14]. Some notable works include: Tits et al. [13] propose a Wavenet-like ASR [15] and pre-train it with VCTK dataset [16]. The representations extracted from different stages of Wavenet are applied to estimate valence and arousal. Lakomkin et al. [12] use DeepSpeech-like model [17] and pre-train on three datasets, LibriSpeech, TED-LIUM v2, and VoxForge [18, 19]. During SER training, a concatenated vector of ASR and SER representations is fed into a softmax layer in a progressive network for SER. Lu et al. [14] use RNN-T model pre-trained on 125,000 hours Youtube videos [20] to extract ASR features for a sentimental decoder. While these works have demonstrated the usefulness of intermediated features of ASR models for emotion recognition by first pre-training them from high-resource ASR databases, these ASR models are used purely as a front-end extractor as its network layers are not adapted to the target emotion database at all. It is well-known that ASR systems are sensitive to the domain mismatch problem, the effect of mismatch may degrade the quality of speech representations used in these SER tasks. However, adaptation approach from ASR datasets to SER domain is not investigated in these works.

In this paper, we explore the strategy of using end-to-end (E2E) ASR system as representation extractor for SER applications. We first pre-train an attention-based Listen, Attend and Spell (LAS) ASR [21] on LibriSpeech subset (360 hours) [18] to extract frame-wise acoustic features. In order to obtain target emotion domain's speech representations, we perform a three-stage fine-tuning pipeline to adapt this pre-trained ASR to an emotional speech dataset, the IEMOCAP (12 hours) [22]. Additionally, we devise an adaptive model compression approach based on singular value decomposition (SVD). SVD-based model adaptation, i.e., a low-rank model adaptation approach, has been widely used to address conditions of acoustic mismatch [23, 24, 25]. In this work, we factorize a weight matrix into two sub-matrices using SVD in every fully-connected (FC) layer of Listener; this approach reduces the parameter size to learn a compact ASR model for a smaller target emotion corpus. Furthermore, in this work, we provide one of the first investigations in understanding the relation between ASR performances and SER accuracy when using ASR as a speech representation extractor.

We evaluate the performance of pre-trained and fine-tuned models on the IEMOCAP. Our results show that SER accuracy

http://dx.doi.org/10.21437/Interspeech.2020-2524

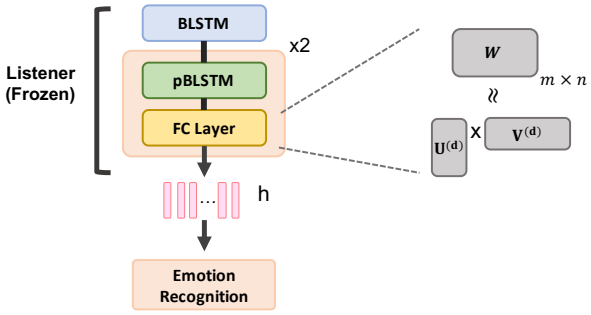

Figure 1: *The proposed framework. We adopt SVD-based model adaptation to FC layers in the encoder of LAS.*

is correlated with ASR performance on the target emotional corpus, i.e., the quality of ASR representations is critical. The fine-tuned LAS outperforms pre-trained LAS by 24.0% in word error rate (WER), 2.1% for emotion recognition accuracy. Finally, our experiments also show that ASR representations from a lower layer of Listener are more suitable for the speech emotion recognition task [11], i.e., achieving the best 66.0% emotion recognition rate.

2. Methodology

Figure 1 shows the proposed procedure for E2E ASR representation learning and emotion recognition task. First, we train a LAS model on LibriSpeech subset. Then, we perform domain adaptation approaches include full model fine-tuning and SVD-based model adaptation. Finally, we freeze the parameters of Listener in order to use it to extract frame-wise hidden representations as input to train the emotion recognition network. We also extract features from the first pBLSTM layer (pBLSTM+FC Layer) to see if representations from a lower layer are more informative. We analyze the emotion recognition performance using different ASR encoders: pre-trained LAS, fine-tuned LAS and structured LAS.

2.1. Datasets

LibriSpeech is a benchmark speech recognition dataset of read English speech publicly available for download [18]. We pre-train LAS model on the LibriSpeech train-clean-360 subset that contains 360 hours of audio samples with 921 speakers and evaluate our models with pre-defined validation and test sets.

IEMOCAP is a well-known conversational SER dataset that consists of five sessions, each session consists of different conversational scenarios between two speakers [22]. In total, the IEMOCAP contains 10 speakers and 12 hours of audio recordings. In this paper, we conduct four emotion classes classification: {anger, happiness+excitement, neutral, sadness}, where happiness and excitement are combined as happiness. The distributions of total 5531 utterances are: {19.9%, 29.5%, 30.8%, 19.5%}.

2.2. LAS Model

The LAS model [21] is a sequence-to-sequence network that consists of an encoder (i.e. acoustic model), a decoder (i.e. language model) with an attention layer between them that learns to align input acoustic signals to output character sequences. Given an input sequence of acoustic feature $X = \{x_1, ..., x_T\}$ with T timesteps, LAS outputs $p(Y|X) = p(y_1|X), ..., p(y_L|X)$, a sequence of posterior probability vectors of output characters, where L is the length of the output sequence, $L \leq T$. The encoder is a stack of two pyramidal BLSTM (pBLSTM) layers on top of a BLSTM layer, each pBLSTM layer is a combination of a FC layer after a pBLSTM shown in Figure 1. The encoder encodes input sequence X into high-level features $h = \{h_1, ..., h_U\}$ with $U \leq T$. The decoder consists of two LSTM layers and an attention layer [21] that maps h to character sequences Y [1]. In this work, we take the encoder to extract ASR representations h for SER training.

We pre-train LAS on LibriSpeech train-clean-360 subset [18], a relatively small amount of source data compared with dataset used in [12, 14], with MFCCs 39 (13+delta+accelerate) features. Moreover, we improve LAS by applying label smoothing and speed perturbation [27, 28]. For speed perturbation, we augment a speech signal by re-sampling it at speed factors 0.9 and 1.1. These techniques improve the generalization of LAS.

2.3. ASR Fine-tuning

Although several generalization techniques have been applied to prevent ASR models from over-fitting to source training data, ASR models remain sensitive to domain mismatch problems. Pre-trained ASR models often fail to consider idiosyncratic characteristics of speech samples from other domains, such as environmental noise, speaking speed and tone. Moreover, training customized ASR systems in SER domain is impractical due to the small scale of emotional speech datasets. We hypothesize that building a domain-adapted ASR can benefit the performance of the downstream task, i.e. speech emotion recognition. To adapt ASR models trained on an audio read dataset (LibriSpeech) to a conversational dataset that contains emotional expressions in speech signals (IEMOCAP), we fine-tune pre-trained ASR on the IEMOCAP with a three-stage fine-tuning pipeline as Figure 2 shows.

Initially, we fine-tune the encoder (acoustic model) and freeze the weights in the decoder. This stage aims at alleviating the acoustic mismatch between the LibriSpeech and the IEMCOAP. Also, stage-1 has the most effect on the IEMOCAP ASR performance. In the second stage, we freeze the fine-tuned encoder, and update the weights of the decoder to fit the lexical characteristics of the IEMOCAP. Finally, we fine-tune the full model weights. Compared with pre-training process, we fine-tune ASR models with a smaller learning rate and a lower probability of teacher forcing. During fine-tuning, speed perturbation and label smoothing are also applied. We use fine-tuned LAS to extract domain-aware speech representations for emotion recognition.

2.4. SVD-based Model Adaptation

SVD-based domain adaptation is a model adaptation technique that removes redundant parameters of a pre-trained model, it provides compact representations when we adapt pre-trained models to a smaller target domain [24, 23, 29]. We conduct low-rank matrix decomposition to FC layers in the encoder to reduce the number of adapted parameters and obtain domain-specific acoustic model. For an $m \times n$ weight matrix W in a FC layer, we approximate it with two low-rank matrices as shown in the right side of Figure 1. The factorized FC layer

[1] Our E2E ASR implementation in tensorflow: `https://github.com/30stomercury/Automatic-Speech-Recognition`.

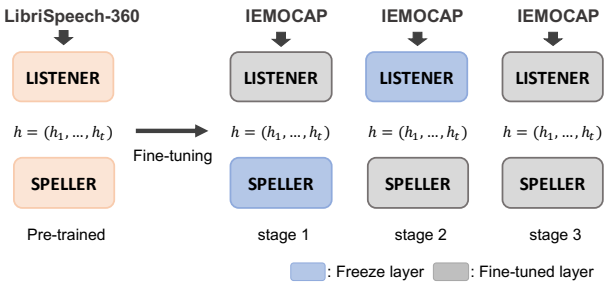

Figure 2: *The proposed three-stage fine-tuning, we fine-tune LAS trained on LibriSpeech with the IEMOCAP. Different parts of LAS are frozen in different stages.*

becomes:

$$\begin{aligned} FC(x) &= \tanh(xW + b) \\ &\approx \tanh(xU^{(d)}V^{(d)\mathsf{T}} + b) \end{aligned} \tag{1}$$

where $U^{(d)}$ is a matrix of size $m \times r$, $V^{(d)\mathsf{T}}$ is a matrix of size $r \times n$. Moreover, we conduct SVD approximation of W to initialize factorized weights. With the help of SVD initialization, adaptation of the encoder (i.e., acoustic model) achieves faster convergence [30, 24].

$$\begin{aligned} W_{n\times m} &= U_{m\times n}\Sigma_{n\times n}V^{\mathsf{T}}_{n\times n} \\ &\approx U_{m\times r}\Sigma_{r\times r}V^{\mathsf{T}}_{r\times n} \\ &\approx U_{m\times r}N_{r\times n} \end{aligned} \tag{2}$$

where $\Sigma_{n\times n}$ is a diagonal matrix with singular values of W in descending order, $r \leq n$. We then keep first r singular values of W and assign values of low-rank matrices $U_{m\times r}$ to $U^{(d)}$, $N_{r\times n}$ to $V^{(d)}$. The number of parameters are reduced from mn to $r(m+n)$. After decomposition, we apply the above three-stage fine-tuning to the structured LAS as well.

2.5. SER Model

Speech emotion recognition in the IEMOCAP is our target recognition task. The goal of SER model is to predict the 4-class emotion. In this paper, we employ IAAN[2] as our emotion classifier [31] that predicts emotion by leveraging contextual information from the current speaker and his/her interlocutor. IAAN includes two GRUs to model contextual information. An interaction-aware attention network is utilized to incorporate contextual information into the target utterance emotion modeling. To lower the computational cost, we use a unidirectional GRU in this work instead of a bidirectional GRU used in [31].

As Figure 1 shown, we use encoder from pre-trained or fine-tuned LAS to extract ASR representations of the target utterance, the previous utterance of the current speaker and the previous utterance of the interlocutor. The details of the IAAN model architecture can be found in [31]. Note that the parameters of the encoder are frozen during the training stage of emotion classification.

[2]Our SER implementation in tensorflow: `https://github.com/30stomercury/Interaction-aware-Attention-Network`.

3. Experimental Setup and Results

3.1. Experimental Setup

3.1.1. Model and Training Details

ASR pre-training is conducted on LibriSpeech train-clean-360 subset. After pre-training, we fine-tune LAS on the IEMOCAP. Special tokens in the IEMOCAP transcriptions like [LAUGHTER], [BREATHING] and [GARBAGE] are removed since LibriSpeech have no such annotations. We compute 39 dimensional MFCC features (13-dimensional Mel-frequency cepstral coefficients with Δ and $\Delta\Delta$) every 10 ms. We use MFCC-39 with per-utterance CMVN as ASR input. We employ a BLSTM and two pBLSTMs with cell units 256 for each direction in Listener. For the Speller, we use two LSTMs with 512 cell units. The size of weights in Listener, i.e. $m \times n$, is 1024×512. During pre-training, the learning rate is set as 1e-4 and a mini-batch size is set as 32. While fine-tuning, we use a smaller learning rate 1e-5 and 10% dropout rate. Also, we apply teacher forcing with probability 1 in the first two fine-tuning stages, 0.8 of teacher forcing probability in the final stage of fine-tuning. While performing SVD-based domain adaptation, we pick $r = 256$. In total, we compressed 10% of the original parameters. For the emotion classification network, we use three GRUs with 1024 cell units. We use the learning rate 1e-4 and a mini-batch size 64. All IAAN models are trained on representations extracted from the ASR encoder.

3.1.2. Baselines

We report word error rate (WER) of LAS on both datasets. For emotion recognition, we present unweighted accuracy (UA) and weighted accuracy (WA), UA is the average of accuracies of each category, WA is the percentage samples correctly classified. We compare different model baselines to examine the effectiveness of domain adaptation on both tasks. In all of our experiments, we use a leave-one-session-out cross validation instead of leave-one-speaker-out cross-validation used in related work [13, 14]. Leave-one-speaker-out cross-validation tends to give a higher performance due to the natural interlocutor dependency. This assures our evaluation is truly speaker-independent and resembles real-world scenarios.

Table 1 presents our experiments of transfer learning and model adaptation baselines.

Pre-trained LAS: The baseline LAS [21] pre-trained on LibriSpeech train-clean-360.
Baseline LAS: The LAS model directly trained on the IEMOCAP without pre-training.
Fine-tuned LAS: The LAS model fine-tuned on the IEMOCAP. We perform three-stage fine-tuning pipeline and fine-tune full model.
Structured LAS: The structured LAS fine-tuned on the IEMOCAP. The weights of FC layers in Listener network is approximated by low-rank matrices with SVD-based initialization.

Table 2 summarizes the results of emotion recognition with representations extracted from different layer depths. A pBLSTM Layer is a stack of a pBLSTM and a FC layer in Figure 1. We experiment using representations derived from the first and the second pBLSTM layer. We denote pBLSTM-Layer$_i$ as the i^{th} pBLSTM layer.

Dataset	Model	WER	WA	UA
LibriSpeech	Pre-trained LAS	26.2	-	-
IEMOCAP	Baseline LAS	97.7	27.9	28.1
	Pre-trained LAS	80.4	61.3	62.3
	Fine-tuned LAS	58.2	62.5	64.4
	Structured LAS	**56.4**	**63.1**	**64.4**

Table 1: *The performance of ASR and SER tasks on both datasets.*

Model	Layers	WA	UA
Fine-tuned LAS	pBLSTM-Layer$_1$	**64.7**	65.8
	pBLSTM-Layer$_2$	62.5	64.4
Structured LAS	pBLSTM-Layer$_1$	64.0	**66.0**
	pBLSTM-Layer$_2$	63.1	64.4

Table 2: *Comparisons of representations from different depths of ASR models.*

3.2. Results and Analysis

3.2.1. ASR Performance

The ASR performance of different baselines are shown in Table 1. Pre-trained LAS obtains 24.9% and 26.2% in WER on LibriSpeech dev and test sets respectively, we only present the result on test set. Pre-trained LAS obtains 80.4% WER on the IEMOCAP, 17.3% relative improvement over Baseline LAS. With three-stage fine-tuning pipeline, Fine-tuned LAS reaches 58.2% with relative 22.2% improvement over Pre-trained LAS. Furthermore, by applying SVD-based adaptation on the encoder, Structured LAS achieves 1.8% relative improvement in WER over Fine-tuned LAS.

3.2.2. SER Performance

Regarding the performance using ASR representations on the 4-class emotion classification task, we compared representations from different LAS models. We find that the performance of ASR models are correlated to emotion recognition accuracy. From Table 1, Baseline LAS only obtains 28.1% in UA. With the help of transfer learning, Pre-trained LAS obtains 62.3% in UA with a relative 34.2% improvement over Baseline LAS. After fine-tuning, Fine-tuned LAS obtains 64.6% in UA, outperforming Pre-trained LAS 2.3% relative. While Structured LAS obtains improvement over Fine-tuned LAS in ASR performance, its performance on SER task, 64.6%, 63.1% in UA and WA, shows only little improvement over Fine-tuned LAS. Finally, from Table 2, representations extracted from the lower layer of Fine-tuned LAS achieves the best 64.7% WA, from the lower of layer Structured LAS achieves the best 66.0% in UA.

3.2.3. Effect of Adaptation

In this section, we analyze the effect on the quality of different adapted ASR representations for SER. Observed from Table 1, Baseline LAS shows poor performance in both tasks, showing the challenge of training a target-only E2E ASR on a small-scale dataset. On the other hand, Pre-trained LAS outperforms Baseline LAS due to its better characterization of speech signals obtained from a larger scale ASR database (Pre-trained LAS obtains 26.2% WER in LibriSpeech while WER of Pre-trained LAS on the IEMOCAP is still high). This indicates that ASR model pre-trained on a relatively sufficient amount of source speech data (360 hours) is capable of capturing rich emotional cues that can be used as input for emotion recognition.

However, the problem of acoustic mismatch still remains. From Table 1, with the help of three-stage fine-tuning, the ASR performance of Fine-tuned LAS on the IEMOCAP is significantly improved. The fine-tuning pipeline enables both acoustic model and language model to better adapt to the IEMOCAP. Consequently, Fine-tuned LAS can extract domain-specific information based on the pre-trained LAS, achieving 64.4% UA in the downstream emotion recognition task along with a relative improvement of 22.2% in WER. Moreover, a compact model adaptation technique, SVD-based model adaptation, is experimented to investigate its contribution to ASR representation. Despite the better performance of Structured LAS in ASR task, i.e., 1.8% improvement over Fine-tuned LAS, Structured LAS shows no improvement in SER task. We hypothesize that the reason is the choice of rank parameter used in our low-rank approximation. In this work, we set the rank $r = 256$ for matrix decomposition. However, pruning too many singular values may lose useful information that benefits downstream tasks such as SER. While Structured LAS exhibits no improvement in SER task, it achieves higher ASR and similar SER performance with much fewer parameters compared with LAS fine-tuned on full model.

3.2.4. Effect of Lower-Layer Representations

Instead of using the ASR representations after the two pBLSTM layers, we investigate representations extracted from a lower layer of LAS, i.e., the first pBLSTM layer. From Table 1, pBLSTM-Layer$_1$ outperforms pBLSTM-Layer$_2$ in Fine-tuned LAS and Structured LAS. Representations from a deeper layer degrade the emotion recognition accuracy, which suggests that deeper layers of ASR may contain less emotional information. This result may be attributed to the mismatch between the transcript of LibriSpeech and the IEMOCAP. For example, annotations in the IEMOCAP such as [LAUGHTER] and [BREATH] are not labeled in LibriSpeech. The conversational language is naturally different from the read speech collected in LibriSpeech, which makes the layers adjacent to the decoder carry less emotional content. Lower layers, however, offer better representations for emotion recognition potentially as they target less on directly mapping to character sequences, i.e., the acoustic properties such as emotion characteristics are preserved.

4. Conclusions

In this paper, we present an ASR-based representation that can be used for speech emotion recognition. We improve the pre-trained ASR encoder and show the contribution of domain adaptation approaches to both ASR and SER performance in the IEMOCAP. LAS with three-stage fine-tuning and SVD-based adaptation present a significant improvement over pre-trained model in both tasks. Besides, we demonstrate that pre-trained ASR can achieve comparable performance on SER task without the need to train on thousands of hours of ASR data. In the future, we would like to investigate the influence of decomposition rank on both ASR and SER tasks. Also, we are interested in devising other adaptation approaches to better improve the ASR and SER performance. Another immediate step is to investigate the effect that the quantity and the type of ASR data could have on SER tasks.

5. References

[1] B. Indyk, I. A. Podgorny, and R. Chan, "Personalized support routing based on paralinguistic information," Sep. 10 2019, uS Patent 10,412,223.

[2] D. Bone, C.-C. Lee, T. Chaspari, J. Gibson, and S. Narayanan, "Signal processing and machine learning for mental health research and clinical applications [perspectives]," *IEEE Signal Processing Magazine*, vol. 34, no. 5, pp. 196–195, 2017.

[3] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[4] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 5200–5204.

[5] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2227–2231.

[6] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, "The interspeech 2010 paralinguistic challenge," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[7] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[8] S. E. Eskimez, Z. Duan, and W. Heinzelman, "Unsupervised learning approach to feature analysis for automatic speech emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5099–5103.

[9] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7390–7394.

[10] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation learning for speech emotion recognition." in *Interspeech*, 2016, pp. 3603–3607.

[11] Y. Belinkov and J. Glass, "Analyzing hidden representations in end-to-end automatic speech recognition systems," in *Advances in Neural Information Processing Systems*, 2017, pp. 2441–2451.

[12] E. Lakomkin, C. Weber, S. Magg, and S. Wermter, "Reusing neural speech representations for auditory emotion recognition," *arXiv preprint arXiv:1803.11508*, 2018.

[13] N. Tits, K. E. Haddad, and T. Dutoit, "Asr-based features for emotion recognition: A transfer learning approach," *arXiv preprint arXiv:1805.09197*, 2018.

[14] Z. Lu, L. Cao, Y. Zhang, C.-C. Chiu, and J. Fan, "Speech sentiment analysis via pre-trained features from end-to-end asr models," *arXiv preprint arXiv:1911.09762*, 2019.

[15] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[16] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.

[17] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[18] V. Panayotov, G. Chen, D. Povey, and S. 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.

[19] A. Rousseau, P. Deléglise, and Y. Esteve, "Enhancing the ted-lium corpus with selected data for language modeling and more ted talks." in *LREC*, 2014, pp. 3935–3939.

[20] H. Soltau, H. Liao, and H. Sak, "Neural speech recognizer: Acoustic-to-word lstm model for large vocabulary speech recognition," *arXiv preprint arXiv:1610.09975*, 2016.

[21] W. Chan, N. Jaitly, Q. Le, and O. 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.

[22] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[23] M. Masana, J. van de Weijer, L. Herranz, A. D. Bagdanov, and J. M. Alvarez, "Domain-adaptive deep network compression," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4289–4297.

[24] J. Xue, J. Li, D. Yu, M. Seltzer, and Y. Gong, "Singular value decomposition based low-footprint speaker adaptation and personalization for deep neural network," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 6359–6363.

[25] Y. Zhao, J. Li, and Y. Gong, "Low-rank plus diagonal adaptation for deep neural networks," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5005–5009.

[26] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[27] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4774–4778.

[28] Y. Yin, R. Prieto, B. Wang, J. Zhou, Y. Gu, Y. Liu, and H. Lin, "Attention-based sequence-to-sequence model for speech recognition: development of state-of-the-art system on librispeech and its application to non-native english," *arXiv preprint arXiv:1810.13088*, 2018.

[29] J. Xue, J. Li, and Y. Gong, "Restructuring of deep neural network acoustic models with singular value decomposition." in *Interspeech*, 2013, pp. 2365–2369.

[30] K. C. Sim, A. Narayanan, A. Misra, A. Tripathi, G. Pundak, T. N. Sainath, P. Haghani, B. Li, and M. Bacchiani, "Domain adaptation using factorized hidden layer for robust automatic speech recognition." in *Interspeech*, 2018, pp. 892–896.

[31] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee, "An interaction-aware attention network for speech emotion recognition in spoken dialogs," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6685–6689.