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

A Lightweight Model Based on Separable Convolution for Speech Emotion Recognition

Ying Zhong[1,2], *Ying Hu*[1,2], *Hao Huang*[1,3], *Wushour Silamu*[1,3]

[1]School of Information Science and Engineering, Xinjiang University, Urumqi, China
[2]Key Laboratory of Signal Detection and Processing in Xinjiang Uygur Autonomous Region
[3]Key Laboratory of Multilingual Information Technology in Xinjiang Uygur Autonomous Region

zhongyingdl3@gmail.com, huying_75@sina.com, hwanghao@gmail.com

Abstract

One of the major challenges in Speech Emotion Recognition (SER) is to build a lightweight model with limited training data. In this paper, we propose a lightweight architecture with only fewer parameters which is based on separable convolution and inverted residuals. Speech samples are often annotated by multiple raters. While some sentences with clear emotional content are consistently annotated (easy samples), sentences with ambiguous emotional content present important disagreement between individual evaluations (hard samples). We assumed that samples hard for humans are also hard for computers. We address the problem by using focal loss, which focus on learning hard samples and down-weight easy samples. By combining attention mechanism, our proposed network can enhance the importing of emotion-salient information. Our proposed model achieves 71.72% and 90.1% of unweighted accuracy (UA) on the well-known corpora IEMOCAP and Emo-DB respectively. Comparing with the current model having fewest parameters as we know, its model size is almost 5 times of our proposed model.

Index Terms: Speech emotion recognition, lightweight, inverted residuals, focal loss

1. Introduction

Emotion plays an important role in daily human interactions, it helps us to contact with each other by expressing our feelings and providing feedback. Recognizing emotion from speech correctly can help intelligent spoken interaction system to understand the potential user's intention, and further improve the user's experience. SER is an important technology to understand human feelings. There has been a growing number of researches and applications in recent years.

Recently, deep learning has attracted increasing attention due to its outstanding performances for many tasks, more and more methods utilizing neural networks to extract valid features from raw data have emerged in the field of SER. Most of them focus on training strategy or modeling networks. Carlos *et al.* [1, 2, 3] explored a robust training strategy by establishing connections between the data. Dai *et al.* [4] proposed a model by cooperating softmax cross-entropy and center loss together to learn discriminative features. Ando *et al.* [5] proposed soft-target training to effectively handle both clear and ambiguous emotional utterances. Sahu *et al.* [6] compressed the high dimensional feature to low dimensionality for maximally capturing the difference between various emotion classes. Mirsamadi *et al.* [7] presented different recurrent neural networks (RNN) with local attention architectures for learning features in speech emotion recognition. Tao *et al.* [8] proposed a new variation of *Long short-term memory* (LSTM), *advanced LSTM* (A-LSTM), for better temporal context modeling for SER. Both of them proved that RNN is effective for sequence data. Chen *et al.* [9] employed attention-based convolutional RNN (ACRNN) network to extract high-level emotional feature representations from the log Mel-spectrogram. Their model showed better performance. The convolution neural networks (CNN) were used to learn affective salient features and manifested excellent performances on several benchmark datasets [4, 10, 11, 12]. Although CNN has been innovated to achieve better recognition performance, it needs a large training parameters. However, the training data for SER is extremely limited. Thus, the SER task is not suitable for models with large number of parameters. Chollet *et al.* [13] presented Xception model based on the depthwise separable convolution, which can learn richer feature representations with fewer parameters. This property provides a theoretical and experimental basis for us to build a lightweight model. Sandler *et al.* [14] proposed MobileNetV2 suggesting that using linear layers is crucial as it prevents non-linear operation from destroying too much information. Inspired by the Xception and MobileNetV2, we built a lightweight model based on separable convolution using inverted residuals for speech emotion recognition in this paper. And with the use of focal loss, the performance of proposed model is further improved.

The remaining of the paper is organized as follows. Section 2 reviews previous works. Section 3 introduces the proposed method. Section 4 describes the experiments and the results. Finally in section 5, we present conclusions.

2. Related Work

Depthwise separable convolution [13], factorizing a standard convolution into a depthwise convolution followed by a pointwise convolution (*i.e.*, 1×1 convolution), drastically reduces computational complexity. Specifically, the depthwise convolution performs a spatial convolution independently for each input channels, while the pointwise convolution is employed to combine the outputs from the depthwise convolution. Results reported on Xception, which is based on depthwise separable convolution, showed that the absence of any non-linearity leads to both faster convergence and better final performance. MobileNets [15] is based on a streamlined architecture that uses depthwise separable convolutions to build lightweight deep neural network. Subsequently, the MobileNetV2 [14] presented inverted residuals and linear bottlenecks. It is also based on the depthwise separable convolution. The authors found that it's important to remove non-linearities in the narrow layers in order to maintain representational power. Inspired by Xception and MobileNetV2, we build a model based on separable convo-

http://dx.doi.org/10.21437/Interspeech.2020-2408

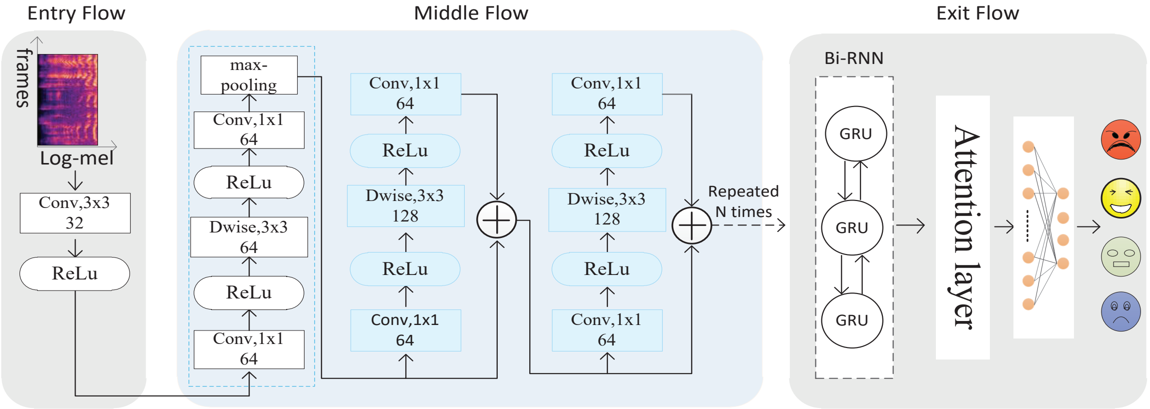


Figure 1: *The proposed lightweight model. The entry flow maps the log-Mels of an utterance to a high-dimensional representation and the middle flow extracts richer information. The exit flow outputs predicted class label.*

lution combining with inverted residuals.

Not all parts of an utterance including emotion in real scene, thus, the attention mechanism is applied to learn emotion relevant regions for utterance-level SER [7, 9, 12]. Conventional SER methods rely on adopting majority votes from multiple annotators as the ground truth. However, the inconsistency of annotations leads to the difficulty of training directly. Lotfian *et al.* presented curriculum learning method to learn hard samples from crowdsourced labels [1]. Chou *et al.* applied hard and soft labels to address the inconsistency of annotations [16]. Focal loss [17] was proposed to address the one-stage object detection scenario in which there is an extreme imbalance between foreground and background classes. In our work, we employ the focal loss to handle classes imbalanced and hard examples.

In this paper, we propose a lightweight model to extract discriminative features for SER from utterances with variable length. We evaluated proposed model on the IEMOCAP and Emo-DB corpus.

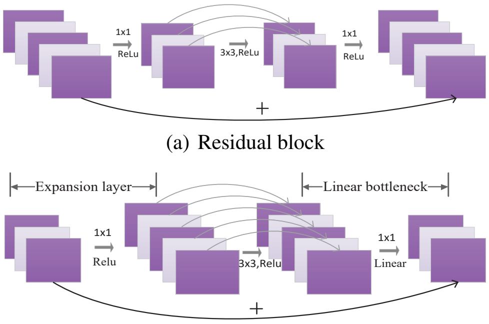
3. The Proposed Method

In this section, we describe the proposed lightweight model as shown in Figure 1. The input features first go through the entry flow which using a 2-D convolution layer with stride of 2, then through the middle flow which is used to automatically extract discriminative feature representations. Finally, these feature representations further produce higher level features for SER through the exit flow.

3.1. Inverted Residual

Figure 1 describes a complete architecture of proposed model. The entry flow extracts shallow information from the features with variable length. The middle flow containing several blocks is used to extract richer information. Especially, the first block with blue dashed box is different from other blocks that it does not adopt residual connection. While the other blocks in middle flow are all inverted residual blocks [14]. The inverted residual block is different from conventional residual block as shown in Fig. 2. The inverted residual block takes an input a low-dimensional feature representation. Then through an expansion operation by a 1×1 convolution layer followed with *Relu* activation operation, the features are expanded to high dimension. And those high-dimension features are further filtered with a depthwise convolution for obtaining richer information. Features are subsequently projected back to low-dimensional representations through a linear convolution without activation operation. Finally, the input and output of each inverted residual blocks are added as the input of next block. We only perform max pooling operation in the end of the first block with the pooling size of 3×3 and stride of 2.

Assuming taking an $h \times w \times d_m$ input tensor X_m, a standard 2D convolution attempts to learn filters in a 3D space with two spatial dimensions and one channel dimension and by applying convolutional kernel $K \in R^{k \times k \times d_m \times d_n}$, to produce an $h \times w \times d_n$ output tensor X_n. Where $k \times k$ is the filter size, and d_m and d_n are the number of input channels and of output channels respectively. The parameters number of standard convolution is calculated by $k \times k \times d_m \times d_n$. That of separable convolution is $k \times k \times d_m + d_m \times d_n$. Compared with standard convolution, the parameters number of depthwise separable convolution is reduced by 5 times approximately.



(a) Residual block

(b) Inverted residual block

Figure 2: *The difference between residual block and inverted residual block. (a) is residual block [18], (b) is inverted residual block [14].*

3.2. Attention Layer

In the exit flow, attention mechanism is applied after a Bi-RNN which compresses variable length sequences produced by middle flow to a fixed-length vector. Each direction of Bi-RNN contains 128 Gated Recurrent Units (GRUs)[4]. Then we can obtain a sequence of 256-dimensional high-level features by concatenating the outputs of two directions. The attention layer is employed to focus on emotion relevant parts and produce discriminative utterance-level representations for SER [9]. As shown in (1), the weight α_t is first computed by a softmax function, where h_t is the Bi-RNN output, then the utterance-level representations c are calculated by performing a weighted sum on h_t according to the weights, as shown in (2).

$$\alpha_t = \frac{exp(W \cdot h_t)}{\sum_{t=1}^{T} exp(W \cdot h_t)} \tag{1}$$

$$c = \sum_{t=1}^{T} \alpha_t h_t \tag{2}$$

Finally, the utterance-level representations are passed into a fully connection layer with 64 output units, then followed with *PReLU* [4] activation function and use one softmax layer to calculate the probability of per emotion.

3.3. Focal Loss

The training of a deep network is based on updating the network parameters to minimize a loss function that expresses the divergence between the predictions and the ground truth labels [19]. For SER, each sentence is often annotated by multiple raters, which are aggregated with methods such as majority vote rules. The inconsistency of evaluations of emotional content may lead to that emotion recognition becomes more difficult. In addition, the imbalance of categories in training data also makes SER more difficult. A common method for addressing class imbalance is to introduce weighting factors [4, 17]. We assigned weights to cross entropy(CE) loss, shown in Eq.(3), the weight w_i is in inverse proportion to the sample number of the class in training set, $\hat{y_i}$ is the *i*-th element of network predictions. The weighted CE loss as follows:

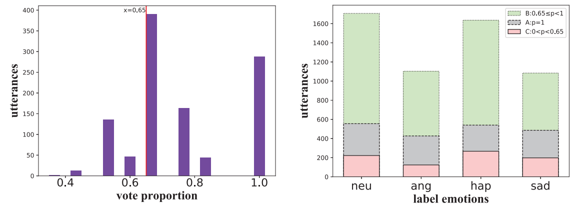
$$CE_w = -\sum_{i=1}^{m} w_i y_i log(\hat{y_i}) \tag{3}$$

Focal loss was proposed in [17] to address class imbalance and hard examples by focusing on learning hard examples and down-weight easy examples. As shown in Eq.(4), adding a factor $(1 - \hat{y_i})^{\lambda}$ to the weighted cross entropy, where λ is hyper-parameter adjusting the rate at which easy examples are down-weighted. Setting $\lambda > 0$ reduces the relatives loss for well-classified examples, putting more focus on hard and misclassified examples. When λ = 0, the model is trained using only weighted cross entropy loss. As $\hat{y_i}$ closes to 1, the factor goes to 0 and the loss for well-classified examples is down-weighted.

$$Focal\_loss = -\sum_{i=1}^{m} w_i (1 - \hat{y_i})^{\lambda} y_i log(\hat{y_i}) \tag{4}$$

Table 1: *Exemplary complete annotations of utterance of Ses01F_impro03_F025.*

Annotators	Annotations		Label	Vote proportion
C-E1	Happiness		hap	4/6
C-E2	Happiness			
C-E4	Happiness;	Excited		
C-F1	Happiness;	Excited		



(a) (b)

Figure 3: *(a)The distribution of vote proportion of label and (b) samples distribution in IEMOCAP database.*

4. Experiments and Results

4.1. Datasets

The IEMOCAP [20] database containing 12 hours English conversations is employed for performance assessment. They are segmented and categorized into utterances with 9 emotion classes. We conducted the classification task only on the same 5 emotion classes as [4, 10, 21]. Same to the reported procedure, utterances in *exciting* class are combined to the *happy* class in evaluation, to form a four-class database labeled with $\{$*happy, angry, sad, neutral*$\}$, each class contains $\{1636, 1103, 1084, 1708\}$ utterances respectively. Each utterance is labeled by three or four annotators and the classification label is the majority label among the annotations [22]. As shown in Table 1, there are six annotations, four of which are *happiness*, thus the vote proportion of label *hap* is 4/6. The distribution of vote proportion of labels is shown in Fig.3(a). We define the samples with vote proportion of 1 as *easy* samples (A), between 1 and 0.65 as *medium difficulty* samples (B), and less than 0.65 as the *difficulty* samples as shown in Fig.3(b). Emo-DB [23] consists of 535 utterances that displayed by ten professional actors, covering seven emotions $\{$*angry, bored, disgust, fear, happy, sadness, neutral*$\}$. The number of each class is $\{127, 81, 46, 69, 71, 62, 79\}$ and all seven emotions are used for our tasks. The sample rate of IEMOCAP database is 16kHz. The Emo-DB database sampled at 44.1kHz, and later downsampled to 16kHz. Because each sample in Emo-DB has only one annotation, so we just apply focal loss on IEMOCAP dataset. The code is available at [1].

Table 2: *The validation of the effect of attention layer.*

	IEMOCAP			Emo-DB		
	UA(%)	WA(%)	F1(%)	UA(%)	WA(%)	F1(%)
No Attention	68.21	66.73	67.25	80.96	84.03	82.83
Attention	70.51	69.63	69.67	90.10	91.81	90.67

Table 3: *Average results (%) of three kind of losses.*

Losses		UA	WA	F1-score
Soft_loss	- -	69.95	69.05	69.29
Lq_loss	- -	68.74	68.51	68.75
Focal_loss	$\lambda_B = 0, \lambda_C = 0$	70.51	69.63	69.67
	$\lambda_B = 1, \lambda_C = 2$	**71.72**	**70.37**	**70.37**
	$\lambda_B = 1.5, \lambda_C = 2$	71.05	70.07	70.72

[1] https://github.com/zhong-ying-china/A-lightweight-network-for-SER

Table 4: *SER average preformance and parameters(MB) of different inverted residual blocks between weighted CE loss and focal loss on IEMOCAP and Emo-DB in terms of UA(%), WA(%) F1-score(%). Note that all convolution layers without batch normalization. All depthwise convolution layers use a depth multiplier of 1.*

blocks	**Parameters**	**Weighted CE Loss**						**Focal Loss**		
		IEMOCAP			**Emo-DB**			**IEMOCAP**		
		UA	**WA**	**F1-score**	**UA**	**WA**	**F1-score**	**UA**	**WA**	**F1-score**
1	0.83M	70.02	68.50	69.03	87.68	89.37	88.47	70.48	69.25	69.74
2	0.85M	**70.57**	69.11	69.61	**90.57**	91.42	90.63	70.61	69.46	69.92
3	0.87M	70.18	68.49	68.92	88.08	90.64	89.25	70.40	68.73	69.06
4	0.88M	69.74	68.64	69.14	90.10	**91.81**	**90.67**	70.31	68.77	69.12
5	0.90M	70.51	**69.63**	**69.97**	88.79	90.83	89.80	**71.72**	**70.39**	**70.85**
6	0.92M	69.49	68.05	68.58	86.90	88.18	87.48	70.66	69.57	70.01

4.2. Experimental Settings

We used 128-dimensional log scale Mel-spectrogram (log-Mel) as input features [4, 12, 24]. The spectrogram is extracted using 1024-point short-time Fourier transformation (STFT) with 25% overlap. Neumann *et al.* found that 7s long utterance contains enough emotional information [12]. So if the utterance is longer than 7s, only the middle part with the length of 7s was calculated.

We employed TensorFlow to implement the proposed method. Adam is used as optimizer. Learning rate was set with 0.0003 and batch size 64. Both of datasets were divided into 10 subsets randomly keeping the emotion distribution, 8 subsets were used for training, one for validation and one for testing (fixed test set). The experimental results are the average of 9 times cross validation. We use three metrics, Unweighted Accuracy (UA), Weighted Accuracy (WA) and F1-score, to evaluate proposed method.

4.3. Results and Discussions

In our first set of experiments, we evaluate the effect of attention layer in Exit Flow (Fig.1) using IEMOCAP and Emo-DB datasets. The system adapted weighted CE loss (Eq.3) for training. Table 2 shows the effect of attention mechanism. *No Attention* in Table 2 indicates that the outputs of Bi-RNN are fed into fully connection layer directly but not pass the attention layer.

Then, we conducted experiments on IEMOCAP adopting focal loss (Eq.4) and comparing with two kinds of losses[19]. For *easy* samples (A), λ was set with 0, denoted as $\lambda_A = 0$ (the loss of easy samples was calculated by Eq.3). λ_B and λ_C denote the values of λ for *medium difficulty* samples (B) and *difficulty* samples (C). Table 3 shows that when $\lambda_B = 1, \lambda_C = 2$, UA, WA and F1-score can achieve the best performance. $\lambda_B = 0, \lambda_C = 0$ means that the model was trained by using weighted CE loss. When λ_B was increased to 1.5 and λ_C remained 2, the performance decreases slightly. The *medium difficulty* samples accounts for nearly 64% of training data. So increasing λ_B slight means that it will reduce the contribution of *difficulty* samples relatively. The followed experiments all adopted the settings of $\lambda_B = 1$ and $\lambda_C = 2$.

Followed, we explored the performance of the system with various number of inverted residual blocks. As shown in table 4, the model achieves the best performance on IEMOCAP when the number of blocks is 5, and on Emo-DB when the number of blocks is 4. Using focal loss, the performance of model increased by 1.7%, 1.08% and 1.3% respectively for UA, WA and F1-score compared with weighted loss (with blocks are 5). It further proves that focal loss can facilitate the generalization of network. Although the network become deeper, the amount of parameters doesn't increase obviously.

Table 5: *Comparions of model size and performance in term of UA with other systems on IEMOCAP and Emo-DB.*

Approaches	**Para**	**IEMOCAP**	**Emo-DB**
3D-ACRNN[9]	323.46M	64.74%	82.82%
DRN [11]	9.9M	67.4%	-
BCRNN[10]	4.34M	61.9%	79.7%
Proposed Model	**0.9M**	**71.72%**	**90.1%**

Finally, we compared our system with several baseline on IEMOCAP and Emo-DB datasets. As shown in table 5, the proposed model has achieved significant improvement comparing to state-of-the-art models with their reported results especially on Emo-DB dataset. 3D-ACRNN has the largest number of model parameters which are almost 32 times of DRN and 73 times of the BCRNN. Among the compared models, BCRNN has the least number of model parameters, however, its model size is almost 5 times of our proposed model. Meanwhile, our proposed model achieves better performance with only fewer parameters.

5. Conclusions

To facilitate the SER application to real-time system, we proposed a lightweight model based on separable convolution network and inverted residuals. By employing attention layer, the model can focus on the parts of emotion relevant. The model also use focal loss to address the problem of class imbalance and difficult samples, and to help the network focus on learning hard examples. IEMOCAP and Emo-DB databases are used to evaluate the performance of the model in terms of UA, WA and F1-score. Results indicate that our proposed model can yield better results compared with state-of-the-art models with fewer parameters.

6. Acknowledgements

This work is supported by National Natural Science Foundation of China (NSFC) (U1903213, 61761041, 61663044), Natural Science Foundation of the Xinjiang (2016D01C061) and University Scientific Research Project of Xinjiang (XJEDU2017T002).

7. References

[1] R. Lotfian and C. Busso, "Curriculum learning for speech emotion recognition from crowdsourced labels," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 815–826, 2019.

[2] M. Abdelwahab and C. Busso, "Incremental adaptation using active learning for acoustic emotion recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5160–5164.

[3] R. Lotfian and C. Busso, "Over-sampling emotional speech data based on subjective evaluations provided by multiple individuals," *IEEE Transactions on Affective Computing*, 2019.

[4] D. Dai, Z. Wu, R. Li, X. Wu, J. Jia, and H. Meng, "Learning discriminative features from spectrograms using center loss for speech emotion recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7405–7409.

[5] A. Ando, S. Kobashikawa, H. Kamiyama, R. Masumura, Y. Ijima, and Y. Aono, "Soft-target training with ambiguous emotional utterances for dnn-based speech emotion classification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4964–4968.

[6] S. Sahu, R. Gupta, G. Sivaraman, W. AbdAlmageed, and C. Espy-Wilson, "Adversarial auto-encoders for speech based emotion recognition," in *Proc. Interspeech 2017*, 2017, pp. 1243–1247.

[7] 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.

[8] F. Tao and G. Liu, "Advanced lstm: A study about better time dependency modeling in emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2906–2910.

[9] M. Chen, X. He, J. Yang, and H. Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[10] H. Zhao, Y. Xiao, J. Han, and Z. Zhang, "Compact convolutional recurrent neural networks via binarization for speech emotion recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6690–6694.

[11] R. Li, Z. Wu, J. Jia, S. Zhao, and H. Meng, "Dilated residual network with multi-head self-attention for speech emotion recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6675–6679.

[12] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," *Proc. Interspeech 2017*, pp. 1263–1267, 2017.

[13] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[14] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR)*, 2018, pp. 4510–4520.

[15] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. A. Mobilenets, "Efficient convolutional neural networks for mobile vision applications," *arXiv preprint ArXiv:1704.0486*, 2017.

[16] H.-C. Chou and C.-C. Lee, "Every rating matters: Joint learning of subjective labels and individual annotators for speech emotion classification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5886–5890.

[17] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision(ICCV)*, 2017, pp. 2980–2988.

[18] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition(CVPR)*, 2017, pp. 1492–1500.

[19] E. Fonseca, M. Plakal, D. P. Ellis, F. Font, X. Favory, and X. Serra, "Learning sound event classifiers from web audio with noisy labels," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 21–25.

[20] 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.

[21] Y. Xu, H. Xu, and J. Zou, "Hgfm: A hierarchical grained and feature model for acoustic emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6499–6503.

[22] L. Tarantino, P. N. Garner, and A. Lazaridis, "Self-attention for speech emotion recognition," *Proc. Interspeech 2019*, pp. 2578–2582, 2019.

[23] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[24] 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.