INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Towards Discriminative Representations and Unbiased Predictions: Class-specific Angular Softmax for Speech Emotion Recognition

Zhixuan Li[1], *Liang He*[1,*], *Jingyang Li*[2,3], *Li Wang*[2,3], *Wei-Qiang Zhang*[1]

[1]Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University
[2]Institute of Forensic Science, Ministry of Public Security, China
[3]Key Laboratory of Intelligent Speech Technology, Ministry of Public Security, China

lizx17@mails.tsinghua.edu.cn, {heliang,wqzhang}@tsinghua.edu.cn,
{lijingyang,wangli}@cifs.gov.cn

Abstract

Speech emotion recognition (SER) is a challenging task: the complex emotional expressions make it difficult to discriminate different emotions; the unbalanced data misleads models to give biased predictions. In this work, we tackle these two problems by the angular softmax loss. First, we replace the vanilla softmax with angular softmax to learn emotional representations with strong discriminant power. Besides, inspired by its novel geometric interpretation, we establish a general calculation model and deduce a concise formula of decision domain. Based on these derivations, we propose our solution to data imbalance: class-specific angular softmax by which we can directly adjust decision domains of different emotion classes. Experimental results on the IEMOCAP corpus indicate significant improvements on two state-of-the-art models therefore demonstrate the effectiveness of our proposed methods.

Index Terms: speech emotion recognition, angular softmax loss, data imbalance, class-specific margin

1. Introduction

Speech emotion recognition (SER) has drawn increasing interest over the past two decades due to the surge of speech-based human-machine interaction situations. However, performances of current SER systems hardly live up to users' expectation.

The first performance-limiting factor is the complexity of emotional expressions. Therefore, discriminative representations are the enduring pursuit. Traditional SER systems often use hand-crafted features [1]. Motivated by the success of deep learning, multiple hybrid systems exploit powerful neural networks such as Deep Neural Network (DNN), Convolutional Neural Network (CNN) and Recurrent Neural Networks (RNN) to extract features and feed them into back-end classifiers[2] [3]. More recently, end-to-end training has dominated in SER due to its joint optimization of feature extractor and classifier. Additionally, intrinsic structures [4] [5] and efficient mechanism such as attention [6] [7] aim to refine emotional information in speech signal and produce more discriminative representations. From the perspective of loss function, however, fewer works are reported in SER despite there are successive state of the arts based on it in other domains [8] [9] [10].

The second hindrance to satisfactory SER system comes from the data. Data sparsity has long been an issue in SER task but we have relatively abundant methods[11][12][13][14]. By comparison, fewer works[15] try to alleviate data imbalance in SER using traditional methods[16] [17] [18] as in other fields [19] [20], let alone developing new ones.

Angular softmax loss [10] is first proposed in face verification and stands out due to its ease of use and superior ability to learn more discriminative representations. More importantly, it has clear and thought-provoking geometrical interpretation which contains possibilities of alleviating data imbalance.

Therefore, in this paper, we first use angular softmax loss to extract discriminative emotional representations. Besides, we propose "class-specific angular softmax" to alleviate data imbalance by directly adjusting decision domains of different emotion classes. Significant performance improvements on two popular baselines confirm the superiority of our idea.

The rest of this paper is organized as follows. Theories of original angular softmax and our class-specific angular softmax are introduced in Section 2. Experimental results on two baselines and detailed analysis are shown in Section 3. Finally, Section 4 presents the conclusions.

2. Angular Softmax Loss and Modifications

In the beginning, we will give a brief description of the original angular softmax loss [10]. Then class-specific angular softmax loss will be described in detail.

2.1. Angular softmax loss

For illustration purpose, we start from the original softmax and focus on two-class case later generalize to multi-class one.

Suppose $\boldsymbol{x}$ is a feature vector, $\boldsymbol{w}_i, b_i$ are weight vector and bias corresponding to class i $(i = 1, 2)$ respectively. In original softmax, the decision boundary is $\boldsymbol{w}_1^T\boldsymbol{x} + b_1 = \boldsymbol{w}_2^T\boldsymbol{x} + b_2$ and equivalently $\|\boldsymbol{w}_1\|\,\|\boldsymbol{x}\|\,cos(\theta_1)+b_1 = \|\boldsymbol{w}_2\|\,\|\boldsymbol{x}\|\,cos(\theta_2)+b_2$ where θ_i is the angle between $\boldsymbol{w}_i$ and $\boldsymbol{x}$.

There are two steps of modification in angular softmax. Firstly, the bias is discarded: $b_1 = b_2 = 0$ and the weight vector is normalized: $\|\boldsymbol{w}_1\| = \|\boldsymbol{w}_2\| = 1$. Therefore the decision boundary becomes $cos(\theta_1) = cos(\theta_2)$ which is only decided by the angle θ_i. To learn more discriminative feature vectors, an integer angular margin m is introduced. Consequently, the decision boundary becomes $cos(m\theta_1) = cos(\theta_2)$ if the ground-truth label is class 1 and vice versa.

The definition of posterior probability of ground-truth class in multi-class case is as follows:

$$p^{(n)} = \frac{e^{\|\boldsymbol{x}^{(n)}\|cos(m\theta_{y_n}^{(n)})}}{e^{\|\boldsymbol{x}^{(n)}\|cos(m\theta_{y_n}^{(n)})} + \sum_{j\neq y_n} e^{\|\boldsymbol{x}^{(n)}\|cos(\theta_j^{(n)})}} \quad (1)$$

where $\{\boldsymbol{x}^{(n)}, y^{(n)}\}$ is a pair of feature vector and its ground-truth label, $\theta_j^{(n)}$ is the angle between $\boldsymbol{x}^{(n)}$ and the j-th weight

http://dx.doi.org/10.21437/Interspeech.2019-1683

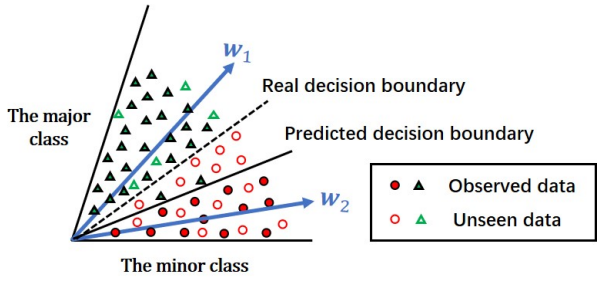

Figure 1: *Explanation of data imbalance*

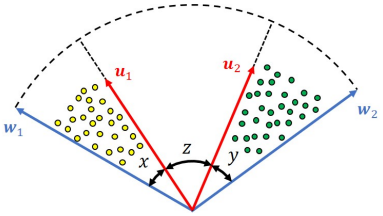

Figure 2: *Calculation model of decision domain*

vector $\boldsymbol{w}_j$. Notice that $\theta_{y_n}^{(n)}$ has to be in the range of $[0, \frac{\pi}{m}]$ since the cosine function only monotonically decreases in $[0, \pi]$. To get rid of this restriction, cosine is replaced with a monotonic function $\phi(\theta_{y_n}^{(n)}) = (-1)^k cos(m\theta_{y_n}^{(n)}) - 2k$, where $\theta_{y_n}^{(n)} \in [\frac{k\pi}{m}, \frac{(k+1)\pi}{m}]$ and $k \in [0, m-1]$. Finally the loss is:

$$L = \frac{1}{N}\sum_{n=1}^{N} -log\left(\frac{e^{\|\boldsymbol{x}^{(n)}\|\phi(m\theta_{y_n}^{(n)})}}{Z}\right) \tag{2}$$

$$Z = e^{\|\boldsymbol{x}^{(n)}\|\phi(m\theta_{y_n}^{(n)})} + \sum_{j\neq y_n} e^{\|\boldsymbol{x}^{(n)}\|cos(\theta_j^{(n)})} \tag{3}$$

2.2. Class-specific angular softmax loss

Data imbalance is a barrier in developing better SER systems. As Figure 1 shows, the real probability distribution of the minor class cannot be accurately modelled by the relatively scarce observed data, therefore, part of its decision domain is mistaken as the major class's. As a consequence, models trained with unbalanced data tend to predict the major class. Obviously, one reasonable solution is to shrink decision domain of the major class while expand the one of the minor class.

To achieve this goal, we need express the size of decision domain quantitatively. Figure 2 illustrates the calculation model which is simplified from two perspectives: firstly, since in angular softmax loss, prediction only depends on the angle between feature vector $\boldsymbol{x}$ and weight vector $\boldsymbol{w}_i$, it is natural to measure the decision domain by its angular span; secondly, only two classes are involved since the decision domain in multi-class case is constructed from results of multiple two-class pairs.

Formal derivations are as follows. Suppose $\boldsymbol{w}_1$, $\boldsymbol{w}_2$ are weight vectors and $\boldsymbol{u}_1$, $\boldsymbol{u}_2$ are feature vectors located on the angular margin boundaries. The angles x, y, z among $\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{u}_1, \boldsymbol{u}_2$ are clearly depicted on Figure 2. The angular margin parameter of class 1 is m_1 and class 2 is m_2. Generally speaking, they are not necessarily the same.

Notice $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ lie on the decision boundaries, therefore,

$$m_1\, x = z + y, \quad m_2\, y = z + x \tag{4}$$

By simplifying these equations we can get the relations among x, y and z:

$$x = \frac{m_2+1}{m_1+1}y, \quad z = \frac{m_1 m_2 - 1}{m_1+1}y \tag{5}$$

Because decision boundary is the bisector of angle between $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$, the angular span D of two decision domain is $D_1 = D_2 = \frac{1}{2}(x+y+z)$. By substituting (5) into,

$$D_1 = \frac{1}{2}(m_1+1)x, \quad D_2 = \frac{1}{2}(m_2+1)y \tag{6}$$

In original angular softmax, a class-agnostic angular margin m_0 is adopted: $m_1 = m_2 = m_0$, therefore,

$$D_1 = \frac{1}{2}(m_0+1)x, \quad D_2 = \frac{1}{2}(m_0+1)y \tag{7}$$

If we change from the base value m_0 to m_1 and m_2,

$$D_1^{'} = \frac{1}{2}(m_1+1)x^{'}, \quad D_2^{'} = \frac{1}{2}(m_2+1)y^{'} \tag{8}$$

Now we can embark on adjusting decision domains. Suppose class 1 is the major class and class 2 the minor class, the decision domain of class 1 should be compressed and that of class 2 should be expanded. Therefore, we expect $D_1 > D_1^{'}$ and $D_2 < D_2^{'}$. That is,

$$\frac{m_0+1}{m_1+1} > \frac{x^{'}}{x}, \quad \frac{m_0+1}{m_2+1} < \frac{y^{'}}{y} \tag{9}$$

It seems that the ratios: $\frac{x^{'}}{x}$ and $\frac{y^{'}}{y}$ obstruct further analysis. However, we argue that $x^{'} \approx x$ and $y^{'} \approx y$ for three reasons:

1. If we assign the base m_0 a relative large value (e.g., $m_0 = 5$, also the best class-agnostic margin as Table 1 shows) and control the deviations $|m_1 - m_0|$, $|m_2 - m_0|$, the ratios $\frac{x^{'}}{x}, \frac{y^{'}}{y}$ does not vary too much and approximate 1.
2. The above derivations are based on the ideal condition that training loss equals zero. However, it is impossible in practice thus the angles x and y vary not quite so much as margin changes from relative large base value m_0.
3. In the implementation of angular softmax loss, the final loss is a weighted sum of original softmax loss and angular softmax loss for stability and convergence[21]. This compromise leads to less stringent representations than expectation and moderate variation of x and y when margin changes.

Since the ratios $\frac{x^{'}}{x}$ and $\frac{y^{'}}{y}$ approximate 1, we can set $m_1 < m_0$ and $m_2 > m_0$ to make inequality (9) hold. In conclusion, by adopting smaller margins for the major classes and larger margins for the minor classes, we can compress decision domains of the major classes and expand those of the minor classes and consequently alleviate the data imbalance.

We name our idea as "class-specific angular softmax loss" and formulate it as follows. Suppose there are K classes and each has an integer margin parameter $m_i, i = 1, ..., K$. We just modify the definition of ϕ to $\phi^{'}(\theta_{y_n}^{(n)}) = (-1)^k cos(m_{y_n}\theta_{y_n}^{(n)}) - 2k$ where $\theta_{y_n}^{(n)} \in [\frac{k\pi}{m_{y_n}}, \frac{(k+1)\pi}{m_{y_n}}]$ and $k \in [0, m_{y_n} - 1]$. The class-specific angular softmax loss is as follows:

$$L = \frac{1}{N}\sum_{n=1}^{N} -log\left(\frac{e^{\|\boldsymbol{x}^{(n)}\|\phi^{'}(\theta_{y_n}^{(n)})}}{Z^{'}}\right) \tag{10}$$

$$Z' = e^{\|\boldsymbol{x}^{(n)}\|\phi'(\theta_{y_n}^{(n)})} + \sum_{j \neq y_n} e^{\|\boldsymbol{x}^{(n)}\| cos(\theta_j^{(n)})} \quad (11)$$

Someone may complain the search scope of angular margins expands dramatically. However, it can be avoided in practice:

1. The number of classes in SER is much smaller than that in face or speaker verification;
2. Furthermore, it is not necessary to assign one margin for one class. Instead, we can group classes with similar sizes and assign one margin for one group;
3. Class-specific margins can be selected based on the optimal value of class-agnostic margin m_0.

3. Experiments

3.1. Baselines

Angular softmax can be a substitution where original softmax appears. To verify its effectiveness in SER, we conduct experiments on the following two baselines.

The first baseline is Convolutional Recurrent Neural Network (CRNN) [4] in Figure 3. Figure 4 shows the second which is based on one state-of-the-art model [6]. The initial time-frequency (TF) convolution is the essence which brings significant performance promotion. However we don't adopt the proposed attention pooling since there is marginal relative improvement (only 0.5% on accuracy approximately) compared with simple global average pooling (GAP), as reported in this work. For simplicity, we name the two baselines "CRNN" and "CNN-TF-GAP" respectively in the following experiments.

3.2. Experiment settings

We use the Interactive Emotional Dyadic Motion Capture dataset (IEMOCAP) [22] for all experiments. This dataset is divided into five sessions and each contains utterances from one female and one male. Following previous works [4] [6], we just choose 4 emotion types (angry, happy, neural and sad) from the improvised speech with unambiguous label.

Considering the relatively scarce data, we split one utterance into multiple fixed-length segments with overlap. We choose 2 seconds segments, 1 second overlap in training and 1.6 seconds overlap in evaluation, according to [6]. In training, all segments from one utterance share the same emotion label; in evaluation, logits or angles of segments from one utterance are averaged to produce a single label.

The input of baselines is spectrogram calculated as follows. A sequence of 40 ms Hamming windows is applied to a fixed-length segment with window shift of 10 ms. For each frame we calculate a DFT of length 1600 (for 10 Hz grid resolution) and only use a frequency range of 0 to 4000 Hz. Finally we obtain a matrix of $M \times N$ where $M = 400$ according to the frequency resolution and $N = 200$ according to the segment length.

In training, in order to keep consistent with baselines [4] [6], for "CRNN" we use the Adam optimizer. Initial learning rate is 0.01 and decays in every epoch with rate of 0.9. For "CNN-TF-GAP" we use the SGD with Nesterov momentum of 0.9. Initial learning rate is 0.1 and decays at the 21, 41 and 61 epochs with rate of 0.1. For both baselines, we use a weight decay of 0.001 and train over 70 epochs with batch size of 64.

In evaluation, we perform 5-fold cross validation. In each fold four sessions are used for training and the remaining is split, one speaker for validation and the other for evaluation. Performance is evaluated using Weight Accuracy (WA) and Unweighted Accuracy (UA). Each evaluation is repeated with different seeds and the average and standard deviation are reported.

Table 1: *Results of different class-agnostic margins on **CRNN** baseline*

Model	WA(%)	UA(%)
softmax	$66.46_{\pm 0.70}$	$58.29_{\pm 1.29}$
$m = 3$	$66.27_{\pm 1.80}$	$60.12_{\pm 0.39}$
$m = 4$	$65.68_{\pm 0.97}$	$59.88_{\pm 1.07}$
$m = 5$	$\mathbf{68.21}_{\pm 0.96}$	$\mathbf{61.35}_{\pm 0.88}$
$m = 6$	$67.34_{\pm 1.68}$	$60.56_{\pm 1.89}$

Table 2: *Results of different class-agnostic margins on **CNN-TF-GAP** baseline*

Model	WA(%)	UA(%)
softmax	$68.97_{\pm 0.47}$	$60.89_{\pm 1.09}$
$m = 3$	$70.84_{\pm 1.02}$	$63.54_{\pm 1.28}$
$m = 4$	$71.29_{\pm 1.21}$	$\mathbf{63.78}_{\pm 1.25}$
$m = 5$	$\mathbf{71.34}_{\pm 1.58}$	$63.01_{\pm 0.74}$
$m = 6$	$70.07_{\pm 1.19}$	$61.23_{\pm 0.30}$

3.3. Experiments of angular softmax

Firstly, we explore the effectiveness of angular softmax loss on two baselines with different class-agnostic margins. Results are summarized in Table 1 and 2. We can observe that:

1. There are consistent improvements on both baselines. Because the only modification is from original softmax to angular softmax, it is convincing to attribute improvements to angular softmax.
2. The appropriate value range of angular margin is much wider than that in face [10] and speaker [23] verification where best performances are at margin of 3 or 4. Here it degrades until margin of 6. It can be explained that the number of classes in SER is much smaller than that in face and speaker verification.
3. Ideally larger margin could bring stronger discriminative power thus better performance. However, this does not always hold in practice. There are two possible reasons. First, the larger angular margin is, the harder model converges. Second, model with excessively larger margin tends to overfit since datasets in SER are much smaller than those in face and speaker verification.

3.4. Experiments of class-specific angular softmax

In this part, we conduct experiments of "class-specific angular softmax" which aims to alleviate data imbalance.

After segmentation described in Section 3.2, the proportions of "angry", "happy", "neural" and "sad" in training set are 13.14%,11.67%,44.26% and 30.93% respectively. Obviously "neural" and "sad" predominate, moreover, "angry" and "happy" have similar number of segments, the same as "neural" and "sad". Therefore, we first make two groups: one includes "angry" and "happy", the other "neural" and "sad". Therefore only two different class-specific margins are adopted. Then we

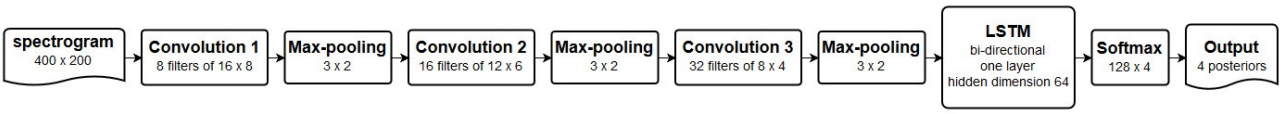

Figure 3: *CRNN baseline*

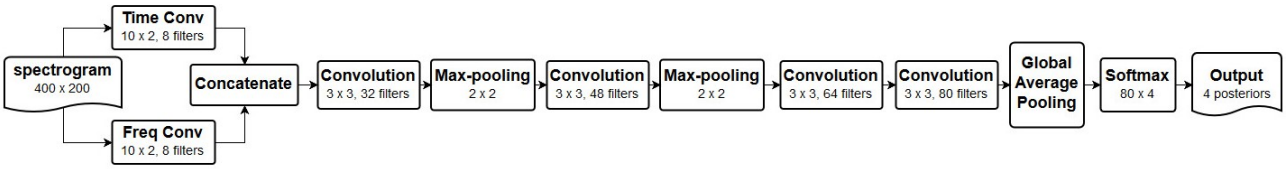

Figure 4: *CNN-TF-GAP baseline*

Table 3: *Results of class-specific margin on **CRNN** baseline*

Model	WA(%)	UA(%)
softmax	$66.46_{\pm 0.70}$	$58.29_{\pm 1.29}$
class-agnostic best	$68.21_{\pm 0.96}$	$61.35_{\pm 0.88}$
$m_{ah}=6, m_{ns}=4$	$68.74_{\pm 0.85}$	$\mathbf{64.16}_{\pm 1.33}$
$m_{ah}=6, m_{ns}=3$	$68.30_{\pm 0.98}$	$62.65_{\pm 1.10}$
$m_{ah}=5, m_{ns}=3$	$\mathbf{68.78}_{\pm 0.80}$	$63.24_{\pm 0.42}$

Table 4: *Results of class-specific margin on **CNN-TF-GAP** baseline*

Model	WA(%)	UA(%)
softmax	$68.97_{\pm 0.47}$	$60.89_{\pm 1.09}$
class-agnostic best	$71.34_{\pm 1.58}$	$63.78_{\pm 1.25}$
$m_{ah}=6, m_{ns}=4$	$\mathbf{73.33}_{\pm 1.34}$	$64.29_{\pm 1.58}$
$m_{ah}=6, m_{ns}=3$	$71.60_{\pm 0.70}$	$64.09_{\pm 0.74}$
$m_{ah}=5, m_{ns}=3$	$72.43_{\pm 1.83}$	$\mathbf{64.80}_{\pm 1.31}$

select the optimal value of class-agnostic margin and assign a smaller margin for "neural" and "sad" and a larger one for "angry" and "happy", according to derivations in Section 2.2. Corresponding results are displayed in Table 3 and 4. In these tables, "softmax" means results of original softmax; "class-agnostic best" means the best results of class-agnostic margin (from Table 1 and 2); "m_{ah}" means margin of "angry" and "happy" classes; "m_{ns}" means margin of "neural" and "sad".

We can observe that models with class-specific angular margins outperform the best with class-agnostic margin. The relative improvements of WA/UA on two baselines are $0.83\%/4.58\%$ and $2.79\%/1.60\%$ over global margin, $3.49\%/10.07\%$ and $6.32\%/6.42\%$ over original softmax.

We also test popular solutions to data imbalance including weighted sampling and weighted loss. In both methods, weight is inversely proportional to a class's segments number, so a normalized weight vector: $[0.35, 0.39, 0.10, 0.15]$ is adopted. We incorporate each method into the best models of class-agnostic margin (5 for "CRNN" and 4 for "CNN-TF-GAP" as Table 1 and 2 show) and report results in Table 5 and 6 where results of "class-agnostic best" and "class-specific best" are copied from Table 1, 2 and Table 3, 4 for convenient comparison.

We can observe that weighted sampling and weight loss did improve UA of two baselines but at the cost of WA. By contrast,

Table 5: *Comparison with other methods on **CRNN** baseline*

Method	WA(%)	UA(%)
class-agnostic best	$68.21_{\pm 0.96}$	$61.35_{\pm 0.88}$
weight sampling	$66.38_{\pm 0.52}$	$62.26_{\pm 1.39}$
weight loss	$66.30_{\pm 1.02}$	$61.47_{\pm 0.27}$
class-specific best	$\mathbf{68.78}_{\pm 0.80}$	$\mathbf{64.16}_{\pm 1.33}$

Table 6: *Comparison with other methods on **CNN-TF-GAP** baseline*

Method	WA(%)	UA(%)
class-agnostic best	$71.34_{\pm 1.58}$	$63.78_{\pm 1.25}$
weighted sampling	$71.11_{\pm 0.31}$	$64.08_{\pm 0.71}$
weighted loss	$70.61_{\pm 1.51}$	$64.36_{\pm 1.18}$
class-specific best	$\mathbf{73.33}_{\pm 1.34}$	$\mathbf{64.80}_{\pm 1.31}$

the appropriate class-specific margins can promote both criteria.

4. Conclusions

In this paper, we first introduce the angular softmax into SER task to learn representations with strong discriminant power. Furthermore, We propose "class-specific angular softmax" to alleviate the severe data imbalance. Specifically, we adopt larger margins for the minor classes while smaller margins for the major classes. The original angular softmax loss brings $2.63\%/5.25\%$ and $3.43\%/4.75\%$ relative improvements of WA/UA on two baselines compared with the original softmax. The class-specific angular softmax brings $3.49\%/10.07\%$ and $6.32\%/6.42\%$ relative improvements. The proposed class-specific angular softmax also bring consistent improvements of WA and UA compared with other solutions to data imbalance. All these results confirm the effectiveness of our idea. Future work will apply the proposed class-specific angular softmax to other state-of-the-art SER models, popular datasets and real scenarios such as noise environment.

5. Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No.U1836219.

6. References

[1] F. Eyben, M. Wöllmer, and B. W. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010*, 2010, pp. 1459–1462.

[2] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, 2014, pp. 223–227.

[3] L. Guo, L. Wang, J. Dang, L. Zhang, and H. Guan, "A feature fusion method based on extreme learning machine for speech emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, 2018, pp. 2666–2670.

[4] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, 2017, pp. 1089–1093.

[5] 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 2018, Calgary, AB, Canada, April 15-20, 2018*, 2018, pp. 2906–2910.

[6] P. Li, Y. Song, I. V. McLoughlin, W. Guo, and L. Dai, "An attention pooling based representation learning method for speech emotion recognition," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018.*, 2018, pp. 3087–3091.

[7] P. Hsiao and C. Chen, "Effective attention mechanism in dynamic models for speech emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, 2018, pp. 2526–2530.

[8] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015, pp. 815–823.

[9] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII*, 2016, pp. 499–515.

[10] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 6738–6746.

[11] Y. Zhang, Y. Liu, F. Weninger, and B. W. Schuller, "Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, 2017, pp. 4990–4994.

[12] Z. Aldeneh and E. M. Provost, "Using regional saliency for speech emotion recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, 2017, pp. 2741–2745.

[13] S. E. Eskimez, Z. Duan, and W. B. Heinzelman, "Unsupervised learning approach to feature analysis for automatic speech emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, 2018, pp. 5099–5103.

[14] 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 2018, Calgary, AB, Canada, April 15-20, 2018*, 2018, pp. 4964–4968.

[15] P. Shih, C. Chen, and H. Wang, "Speech emotion recognition with skew-robust neural networks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, 2017, pp. 2751–2755.

[16] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009.

[17] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[18] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018.

[19] C. Huang, Y. Li, C. Change Loy, and X. Tang, "Learning deep representation for imbalanced classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5375–5384.

[20] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 8, pp. 3573–3587, 2017.

[21] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks," in *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, 2016, pp. 507–516.

[22] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[23] Z. Huang, S. Wang, and K. Yu, "Angular softmax for short-duration text-independent speaker verification," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018.*, 2018, pp. 3623–3627.