Emotion Profile Refinery for Speech Emotion Classification

Shuiyang Mao, P. C. Ching, Tan Lee

Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong

maoshuiyang@link.cuhk.edu.hk, pcching@ee.cuhk.edu.hk, tanlee@ee.cuhk.edu.hk

Abstract

Human emotions are inherently ambiguous and impure. When designing systems to anticipate human emotions based on speech, the lack of emotional purity must be considered. However, most of the current methods for speech emotion classification rest on the consensus, e. g., one single hard label for an utterance. This labeling principle imposes challenges for system performance considering emotional impurity. In this paper, we recommend the use of emotional profiles (EPs), which provides a time series of segment-level soft labels to capture the subtle blends of emotional cues present across a specific speech utterance. We further propose the emotion profile refinery (EPR), an iterative procedure to update EPs. The EPR method produces soft, dynamically-generated, multiple probabilistic class labels during successive stages of refinement, which results in significant improvements in the model accuracy. Experiments on three well-known emotion corpora show noticeable gain using the proposed method.

Index Terms: speech emotion classification, emotional impurity, emotional profiles, soft labeling, iterative learning

1. Introduction

Automatic detection of human emotion in natural expressions is non-trivial. This difficultly is in part due to emotional ambiguity and impurity [1]. However, conventional emotion classification systems rely on majority voting (i. e., one-hot hard label) from a set of annotators as the ground truth. This labeling principle imposes specific challenges on emotion classification tasks: 1) Incomplete Labeling: Human expressions involve a complex range of mixed emotional manifestations [2]. Emotion classification systems designed to output one emotion label per input speech utterance/segment may perform poorly if the expressions cannot be well captured by a single emotional label [1]. 2) Inter-category Dependency: Certain emotion classes are inherently ambiguous. For example, the emotion class of frustration has the potential to overlap with categories ranging from anger, to neutrality and to sadness [2, 3].

Soft labeling approaches have been recently developed to characterize blended emotional expressions. For instance, Lotfian et al. devised an innovative probabilistic method for soft labeling of emotions [4]. Ando et al. developed a *deep neural network* (DNN)-based model trained with soft emotion labels as ground truth, to better characterize the emotional ambiguity [5]. Kim et al. proposed to use cross entropy to directly compare human and machine emotion label distributions based on soft labeling [6].

While soft labeling provides better flexibility in characterizing the emotional impurity and ambiguity, in most of the existing work, the soft labels are assigned per utterance, which is termed static soft labeling. However, as is well known, emotions in natural human expressions do not follow a static mold. Instead, they vary temporally with speech [2, 7]. The static soft labeling thus fails to characterize the emotional fluctuation across the utterance. A natural solution to this problem is to perform segment-level soft labeling. As a first step toward this goal, this work adopts an emotion classification paradigm based on *emotion profiles* (EPs), which is a time series of segment-level soft labels across an utterance, with each dimension representing a classifier-derived probability of a possible emotion component.

EPs have been around within the community for a while. For instance, Mower et al. derived EPs using a set of binary *support vector machine* (SVM) outputs [1, 8]. Han et al. utilized a DNN-based model trained with stacked raw acoustic features to obtain deep-learned EPs [9]. Our previous work further extended EPs into an end-to-end approach using a *deep convolutional neural networks* (DCNN) [10]. While these EPs based studies have achieved impressive performance and provided more interpretable representations than traditional systems, one major shortcoming remains: the lack of segment-level ground truth labels. To circumvent this problem, most of the previous studies assigned the utterance-level one-hot label, which we call *pseudo one-hot label*, to all of the segments within the same utterance [9, 10], or trained the segment-level classifier with utterance-level dataset [8]. This may result in an inconsistency with the ground truth or impart a mismatch to the segment-level classifier.

To better train a segment-level classifier, we argue that several characteristics should apply to ideal segment-level labels: 1) Labels should be informative of the specific segment, meaning that they should not be identical for all the segments across a given utterance. Therefore, labels should be defined at the segment-level rather than merely inheriting the label of the whole utterance. 2) Determining an ideal label for each segment may require observing the entire data to establish intra- and inter-category relations, suggesting that labels should be collective across the whole dataset. To achieve this, we propose emotion profile refinery (EPR). This solution uses a neural network model and the data to dynamically update the segment-level labels during the successive stages of refinery, enabling to generate more informative and collective segment-level labels.

Extensive experiments are conducted on three popular emotion corpora, namely, the CASIA corpus [11], the Emo-DB corpus [12] and the SAVEE database [13]. Experimental results show that the proposed method consistently improves the accuracy of models for speech emotion classification by a significant margin: the CASIA corpus from 93.10% to 94.83% (WA&UA), the Emo-DB corpus from 83.00% to 88.04% (WA) and 82.36% to 87.78% (UA), and the SAVEE database from 70.63% to 77.08% (WA) and 69.88% to 74.64% (UA). Our contributions include: 1) proposing the EPR framework for speech emotion classification task, 2) achieving the state-of-the-art accuracy on the three emotion corpora, and 3) demonstrating the ability of a network to improve accuracy by training from labels generated by another network of the same architecture.

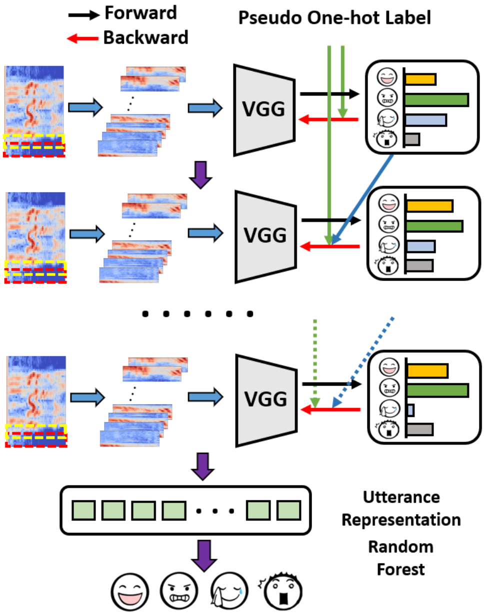


Figure 1: *Illustration of the proposed method*

2. Methods

Figure 1 illustrates a schematic approach of the proposed method. It comprises a series of VGG [14] networks trained to generate EPs from log-Mel filterbanks of individual segments. As the networks go through various stages of the refinery, the segment-level labels (and hence the EPs) are updated. The latest EPs are used for constructing utterance representations (i. e., extracting statistics across the EPs as in [10]). Finally, a *random forest* (RF) is employed to assign the utterance-level labels.

2.1. Emotion profiles (EPs)

Emotion profiles (EPs) were investigated and demonstrated to be useful for emotion classification tasks in [1, 8, 9, 10, 15, 16, 17]. Typically, EPs are time series of classifier-derived segment-level estimates of a set of the "basic" emotions (e. g., angry, happy, neutral, sad), with each EP component representing the probability of the corresponding emotion category.

2.1.1. Generating EPs

We generate the EPs using a VGG model trained on the 64-bin log Mel filterbanks of individual segments. The log Mel filterbanks are computed by *short-time Fourier transform* (STFT) with a window length of 25 ms, hop length of 10 ms, and FFT length of 512. Subsequently, 64-bin log Mel filterbank features are derived from each short-time frame, and the frame-level features are combined to form a time-frequency matrix representation of the segment. The trained VGG model aims to predict a probability distribution $\boldsymbol{P}_i$ for the i^{th} segment in Utterance $\boldsymbol{U}$:

$$\boldsymbol{P}_i = [p_i(e_1),\ p_i(e_2),\ \cdots,\ p_i(e_K)]^T \in \mathbb{R}^{K\times 1} \tag{1}$$

where, $e_1,\ e_2,\ \cdots,\ e_K$, represent the set of "basic" emotions, and K denotes the number of possible emotions. The EP for Utterance $\boldsymbol{U}$ can then be formed as a multi-dimensional signal:

$$\boldsymbol{U}_{EP} = [\boldsymbol{P}_1,\ \boldsymbol{P}_2,\ \cdots,\ \boldsymbol{P}_N] \in \mathbb{R}^{K\times N} \tag{2}$$

where N is the number of segments in the utterance.

2.2. Emotion profile refinery (EPR)

Simply assigning the utterance-level emotion label to all of its segments as the ground truth may not be accurate. We address this problem by passing the dataset through multiple EPs refiners (i. e., a series of VGG networks). The first refinery network C_1 is trained over the dataset, where each training segment is assigned the pseudo one-hot hard label that inherited from its utterance. The second refinery network C_2 is trained over the same dataset but uses soft labels generated by C_1 (maybe combined with the original pseudo one-hot hard labels to mitigate an overfitting problem caused by the refinery process, which will be discussed in Section 4). Once C_2 is trained, we can similarly use the updated EPs to train a subsequent network C_3, and so on. The latest EPs are used as the ground truth EPs to construct the utterance representations for further classification.

2.2.1. Loss

We train the first refinery VGG network C_1 using the cross-entropy loss against the pseudo one-hot labels. We train each of the subsequent refinery networks C_t for $t > 1$ by minimizing the KL-divergence between its output and the soft label (maybe combined with the original pseudo one-hot hard label) generated by the previous refinery network C_{t-1}. Letting $p^t(e_k)$ be the probability assigned to class e_k in the output of model C_t, our loss function for training model C_t is:

$$\begin{aligned}\mathcal{L}_t &= -\sum_k p^{t-1}(e_k)\ \log\frac{p^t(e_k)}{p^{t-1}(e_k)} \\ &= -\sum_k p^{t-1}(e_k)\ \log p^t(e_k) + \sum_k p^{t-1}(e_k)\ \log p^{t-1}(e_k)\end{aligned} \tag{3}$$

The second term is constant with respect to C_t. We can remove it and instead minimize the cross-entropy loss:

$$\hat{\mathcal{L}}_t = -\sum_k p^{t-1}(e_k)\ \log p^t(e_k) \tag{4}$$

3. Emotion Corpora

Three different emotion corpora are used to evaluate the validity and universality of our method, namely, a Chinese emotion corpus (CASIA) [11], a German emotion corpus (Emo-DB) [12] and an English emotional database (SAVEE) [13], which are summarized in Table 1. All of the emotion categories are selected for each of the three stated emotion corpora, respectively.

Specifically, the CASIA corpus [11] contains $9,600$ utterances that are simulated by four subjects (two males and two females) in six different emotional states, i. e., angry, fear, happy, neutral, sad, and surprise. In our experiments, we only use $7,200$ utterances that correspond to 300 linguistically neutral sentences with the same statements.

The Berlin Emo-DB German corpus (Emo-DB) [12] was collected by the Institute of Communication Science at the Technical University of Berlin. Ten professional actors (five males and five females) each produced ten utterances in German to simulate seven different emotions. The number of spoken utterances for these seven emotions is not equally distributed: 126 anger, 81 boredom, 47 disgust, 69 fear, 71 joy, 79 neutral, and 62 sadness.

The Surrey audio-visual expressed emotion database (SAVEE) [13] consists of recordings from four male actors in seven different emotions: anger, disgust, fear, happy, sad, surprise, and neutral. Each speaker produced 120 utterances. The sentences were chosen from the standard TIMIT corpus and phonetically-balanced for each emotion.

Table 1: *Overview of the selected emotion corpora. (#Utterances: number of utterances used, #Subjects: number of subjects, and #Emotions: number of emotions involved.)*

Corpora	Language	#Utterances	#Subjects	#Emotions
CASIA	Chinese	7,200	4 (2 female)	6
Emo-DB	German	535	10 (5 female)	7
SAVEE	English	480	4 (0 female)	7

4. Experiments

We evaluate the proposed method on the three mentioned emotion corpora. We first explore the effect of EPR without combining the original pseudo one-hot hard label, which we call *standard EPR* (sEPR). We then present some ablation studies and analyses to investigate the source of the improvements using the sEPR method. Finally, the original pseudo one-hot hard label is combined with the soft label generated by an iterative EPR process, which we call *pseudo one-hot hard label assisted EPR* (pEPR). The pEPR method achieves the best results.

4.1. Setup

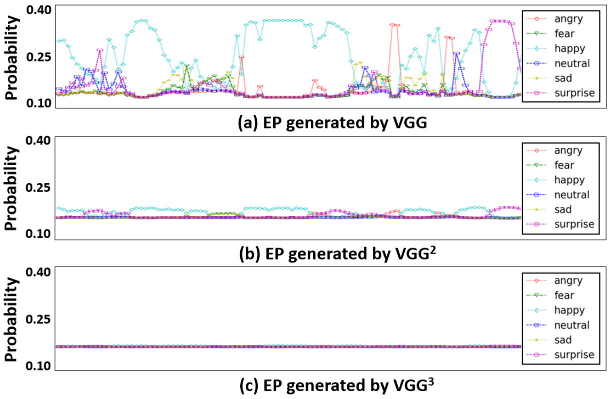
The size of each speech segment is set to 32 frames, i.e., the total length of a segment is 10 ms × 32 + (25 - 10) ms = 335 ms. For the CASIA corpus, the segment hop length is set to 30 ms, whilst it is set to 10 ms for the Emo-DB corpus and the SAVEE database. In this way, we collected 418,722 segments for the CASIA corpus, 131,053 segments for the Emo-DB corpus, and 51,027 segments for the SAVEE database, to train the VGG network, respectively.

For the VGG network, the architecture of the convolutional layers is based on the configurations (i. e., configuration E) in the original paper [14]. A tweak is made to the number of units in the last softmax layer in order to make it suitable for our tasks. In the training stage, ADAM [18] optimizer with default setting in Tensorflow [19] was used, with an initial learning rate of 0.001 and an exponential decay scheme with a rate of 0.8 every 2 epochs. The batch size was set to 128. Early stopping with patience of 3 epochs was utilized to mitigate an overfitting problem. Maximum number of epochs was set to 20.

The EPs were generated using ten-fold cross-validation. A *random forest* (RF) with default setting in Scikit-learn [20] was then employed to make the utterance-level decision, where another ten-fold cross-validation was performed. The results were presented in terms of unweighted accuracy (UA) and weighted accuracy (WA), respectively. It is worth noting that the UA and WA are the same for the CASIA corpus as the CASIA corpus is (perfectly) balanced concerning the emotion category.

4.2. Standard EPR (sEPR)

We first investigated the effect of sEPR. Table 2 shows the experimental results on the three mentioned emotion corpora. Each row represents a randomly-initialized instance of VGG network trained with labels refined by the network directly one row above it in the table. As can be observed: 1) All VGG networks achieved the best performance after one single round of sEPR process, after which performance diminished significantly. 2) The performance gain was only minor. To explain these observations, we looked into the EPs generated during each sEPR iteration. Figure 2 shows an example of EP evolution during two successive stages of refinement for the audio file "Happy_liuchanhg_440.wav" from the CASIA corpus. It is obvious that the sEPR method tends to flatten and collapse the EPs iteratively, and each output dimension of VGG^2 is close to 0.16, i. e., the value obtained by a random guess for the CASIA corpus. We argue that this is because the model tends to minimize the cross-entropy progressively, and the refined EPs contain information that it has memorized from the previous round of training examples. Therefore, a severe overfitting problem happened. We further argue that there is a trade-off between the minimization of segment-level cross-entropy and the maximization of utterance-level accuracy. To address this problem, the pEPR method was proposed and experimented. This is discussed further in Section 4.4.



(a) EP generated by VGG

(b) EP generated by VGG^2

(c) EP generated by VGG^3

Figure 2: *An example of EP evolution for the audio file "Happy_liuchanhg_440.wav" from the CASIA corpus. The sEPR method was applied.*

Table 2: *Results using the sEPR method on the three stated emotion corpora. Each model is trained using labels refined by the model right above it. That is,* VGG^2 *is trained by the labels refined by* VGG*, and so on. The first row networks are trained using the original pseudo one-hot hard labels.*

	CASIA		Emo-DB		SAVEE	
Model	WA	UA	WA	UA	WA	UA
VGG	93.10	93.10	83.00	82.36	70.63	69.88
VGG^2	**93.67**	**93.67**	**83.74**	**83.96**	**71.88**	**70.64**
VGG^3	90.07	90.07	69.91	67.92	26.04	21.07

4.3. Dynamic labels vs. soft labels

In the very beginning, we posit that the benefits of using sEPR are twofold: 1) Each segment is dynamically re-labeled with a more accurate label, and 2) the introduction of soft labeling. To assess the improvement from dynamic labeling alone, we performed label refinement with hard dynamic labels. Specifically, we passed each segment to the VGG network, and the one-hot label was assigned by choosing the most-likely category from the network output. To observe the improvement from soft labeling alone, we investigated the soft static labels. To compute the soft static label for a given segment, we passed all segments within the same utterance to the VGG network, and the

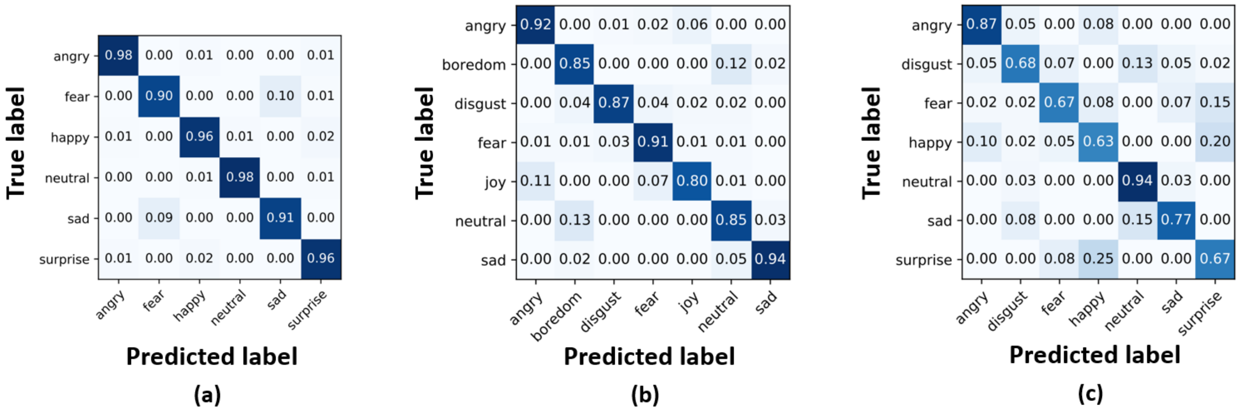
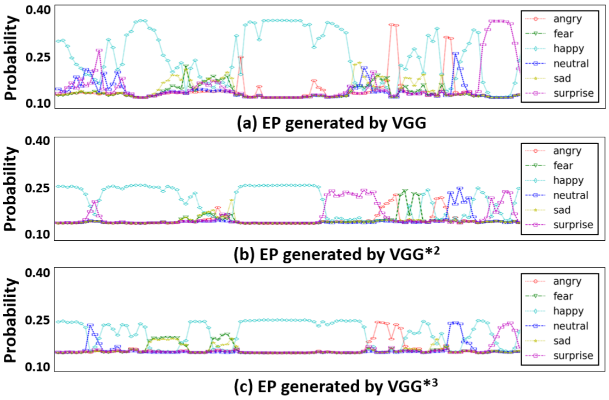


Figure 3: *Confusion matrices obtained using the pEPR method on (a) the CASIA corpus; (b) the Emo-DB corpus; (c) the SAVEE database.*



(a) EP generated by VGG

(b) EP generated by VGG*2

(c) EP generated by VGG*3

Figure 4: *An example of EP evolution for the audio file "Happy_liuchanhg_440.wav" from the CASIA corpus. The pEPR method was applied.*

soft static label was computed by averaging the network outputs across the utterance. Table 3 shows the results. As can be seen, the hard dynamic labeling consistently improved the accuracy of the network for the three emotion corpora, while it was not the case for the soft static labeling. However, when they were combined we observed an additional improvement, suggesting that they address different issues with labels in the dataset.

Table 3: *Comparison of experimental results for hard dynamic labels and soft static labels.*

	CASIA		Emo-DB		SAVEE	
Model	WA	UA	WA	UA	WA	UA
No Refinery	93.10	93.10	83.00	82.36	70.63	69.88
Soft Static	91.36	91.36	79.64	78.77	67.71	66.48
Hard Dynamic	93.21	93.21	83.18	83.13	71.04	70.00
Soft Dynamic	**93.67**	**93.67**	**83.74**	**83.96**	**71.88**	**70.64**

4.4. Pseudo one-hot hard label assisted EPR (pEPR)

In this section, we aimed at mitigating the overfitting problem reported in Section 4.2. We handled this issue by combining the generated soft labels with the original pseudo one-hot hard labels. Specifically, the network output (e. g., [0.6, 0.1, 0.1, 0.2]) of a certain segment and its original pseudo one-hot hard label (e. g., [1, 0, 0, 0]) were added and normalized (i. e., [0.8, 0.05, 0.05, 0.1]), which was then used as the refined label to train the next network. The intuition of this operation is only natural. Since there exists a trade-off between the minimization of the segment-level cross-entropy and the optimization of the utterance-level performance (refer to Section 4.2), we conjecture that the combination of the original pseudo one-hot hard labels might offer an advantage in regularizing the segment-level network training and adding a strong bias towards utterance-level accuracy. Figure 4 shows an example of EP evolution generated using the pEPR method for the same audio file as in Section 4.2. It can be observed that the serve EPs flattening and collapse encountered using sEPR method (see Figure 2) disappeared. Table 4 shows the results. A significant improvement can be observed compared to the sEPR method, which corroborated our previous conjecture. Figure 3 shows the corresponding confusion matrices obtained using the pEPR method on the three mentioned emotion corpora, respectively.

Table 4: *Results using the pEPR method on the three stated emotion corpora.*

	CASIA		Emo-DB		SAVEE	
Model	WA	UA	WA	UA	WA	UA
VGG	93.10	93.10	83.00	82.36	70.63	69.88
VGG^{*2}	**94.83**	**94.83**	87.10	86.78	73.96	71.67
VGG^{*3}	94.54	94.54	86.92	86.42	76.67	74.33
VGG^{*4}	94.60	94.60	85.23	85.07	**77.08**	**74.64**
VGG^{*5}	94.24	94.24	**88.04**	**87.78**	74.58	73.10

5. Conclusions

In this paper, we addressed the problem of emotional impurity encountered in speech emotion classification task using emotion profile refinery (EPR). This method allows us to dynamically label the speech segments with soft targets, which characterizes the probability distributions of the underlying mixture of emotions at segment level. Two EPR method, namely, the standard EPR (sEPR) and the pseudo one-hot hard label assisted EPR (pEPR), were proposed and investigated, and the latter significantly outperformed the former. We achieved the state-of-the-art results on three well-known emotion corpora, respectively.

6. References

[1] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1057–1070, 2010.

[2] E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S. Narayanan, "Interpreting ambiguous emotional expressions," in *Proc. ACII*, 2009, pp. 1–8.

[3] 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.

[4] R. Lotfian and C. Busso, "Formulating emotion perception as a probabilistic model with application to categorical emotion classification," in *Proc. ACII*, 2017, pp. 415–420.

[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 *Proc. ICASSP*, 2018, pp. 4964–4968.

[6] Y. Kim and J. Kim, "Human-like emotion recognition: Multi-label learning from noisy labeled audio-visual expressive speech," in *Proc. ICASSP*, 2018, pp. 5104–5108.

[7] C. Busso and S. S. Narayanan, "Interrelation between speech and facial gestures in emotional utterances: a single subject study," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2331–2347, 2007.

[8] E. M. Provost and S. Narayanan, "Simplifying emotion classification through emotion distillation," in *Proc. APSIPA*, 2012, pp. 1–4.

[9] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. INTERSPEECH*, 2014, pp. 223–227.

[10] S. Mao, P. C. Ching, and T. Lee, "Deep learning of segment-level feature representation with multiple instance learning for utterance-level speech emotion recognition," in *Proc. INTERSPEECH*, 2019, pp. 1686–1690.

[11] J. Tao, F. Liu, M. Zhang, and H. Jia, "Design of speech corpus for mandarin text to speech," in *Proc. the 4th Workshop on Blizzard Challenge*, 2005.

[12] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proc. INTERSPEECH*, 2005, pp. 1517–1520.

[13] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (savee) database," *University of Surrey: Guildford, UK*, 2014.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[15] S. Mao and P. C. Ching, "An effective discriminative learning approach for emotion-specific features using deep neural networks," in *Proc. ICONIP*, 2018, pp. 50–61.

[16] Y. Shangguan and E. M. Provost, "Emoshapelets: Capturing local dynamics of audio-visual affective speech," in *Proc. ACII*. IEEE, 2015, pp. 229–235.

[17] Y. Kim and E. M. Provost, "Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions," in *Proc. ICASSP*, 2013, pp. 3677–3681.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proc. OSDI*, 2016, pp. 265–283.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.