SPEECH EMOTION RECOGNITION WITH SKEW-ROBUST NEURAL NETWORKS

Po-Yuan Shih and Chia-Ping Chen[*]

National Sun Yat-sen University
Computer Science and Engineering
Kaohsiung, Taiwan ROC

Hsin-Min Wang[†]

Academia Sinica
Institute of Information Science
Taipei, Taiwan, ROC

ABSTRACT

We propose a neural-network training algorithm that is robust to data imbalance in classification. In our proposed algorithm, weights are introduced to training examples, effectively modifying the trajectory traversed in the parameter space during the learning process. Furthermore, the proposed algorithm would reduce to the normal stochastic gradient decent learning if the data is balanced. On the FAU-Aibo database, which is known to be used in Interspeech Emotion Challenge, the proposed method achieves an unweighted average (UA) recall rate of 45.3% on the 5-class speech emotion recognition task. Within the static modeling framework, where each example is represented as a fixed-length vector, this performance is one of the best performance ever achieved on the 5-class task.

Index Terms— speech emotion recognition, data imbalance, neural networks

1. INTRODUCTION

A speech emotion recognition (SER) system takes a speech waveform as input, and outputs one of the emotional categories known to the classification system and hypothetically conveyed in the input speech. For examples of applications, SER can be incorporated in automatic speech recognition systems or spoken dialogue systems to improve recognition accuracy or user experience.

SER evaluation plans have been implemented to promote global researches. The Interspeech 2009 Emotion Challenge (henceforth referred to as the Challenge) is a large-scale evaluation project to advance the technology of speech emotion recognition [1]. Since the release of the Challenge, many methods have been proposed to raise the bar of SER [2, 3, 4, 5, 6, 7, 8].

It may be hard to believe, but sound performance levels for the SER tasks as defined in the Challenge have *not* been achieved as of today. The highest unweighted average (UA) recall rate among submissions to the Challenge was 41.7% according to a summarization paper [9]. Following the Challenge, modern approaches for classification have been proposed to improve the performance on the 5-class task. For the static modeling framework in which each speech chunk is represented by a fixed-size vector, the highest UA achieved is 44.0% with anchor models [4]. For the dynamic modeling framework in which each speech chunk is represented by a variable-length sequence of feature vectors, the highest UA achieved is 45.6% with hybrid HMM-DNN systems [6]. Admittedly, there is much room for improvement.

Two of the reasons for such difficulty, we believe, are the *issue of skewed database* and the *intrinsic ambiguity of emotion expression*. Both issues pose general challenges in machine learning. Using skewed data for system development is not an unusual scenario, as the data collection for a machine-learning system is often automated nowadays to reduce human factors. Further, as classification tasks move from areas of well-defined classes to uncharted territories, uncertainty in the labels is bound to happen either due to crowd-sourcing or the intrinsic ambiguity between target classes.

In this paper, we propose *skew-robust neural networks* which use data weighting to deal with skewed data. New training objective functions are incorporated into learning process, to emphasize the examples of the small emotional classes and de-emphasize the examples of the large emotional classes. Furthermore, as part of the ambiguity of emotion expression comes from the difference between speakers, we also apply *cross-speaker histogram equalization* method to ameliorate such differences.

The following sections of this paper are organized as follows. We introduce the basic ideas and describe the proposed methods in Section 2. Experiments and evaluation results are presented in Section 3. Finally, the concluding remarks are given in Section 4.

2. PROPOSED METHOD

FAU-Aibo database consists of highly skewed data set, which posts a serious challenge on emotion classification. The number of speech chunks in FAU-Aibo is shown in Figure 1. A naïve classifier that simply assigns each test chunk to the Neu-

[*]Thanks to the Ministry of Science and Technology for funding.
[†]Thanks to the Ministry of Science and Technology for funding.

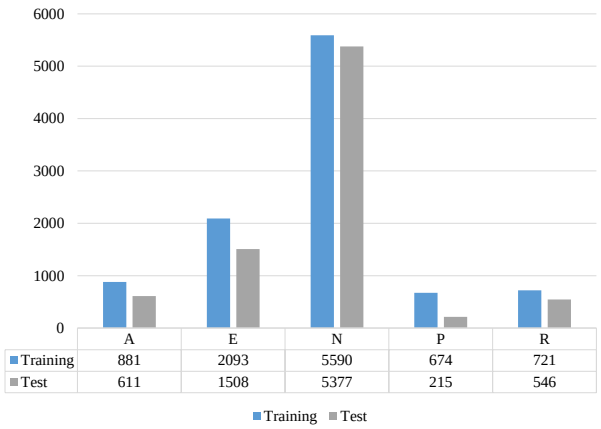

Fig. 1. Decomposition of class-wise speech chunks in FAU-Aibo corpus: anger (A), emphatic (E), Neutral (N), positive (P), and rest (R).

tral class would achieve a weighted average (WA) recall rate of 65%. Thus, WA is not a good measure of performance. The measure of performance adopted in the Challenge is the unweighted average (UA) recall rate, which is the average of the recall rates of different classes. The above classifier would achieve an UA of 20%, which would be more reasonable.

2.1. Skew-Robust Neural Networks

A classification system trained with unbalanced training data tends to predict a class that is populous in the training data. In this paper, we propose the following method, called *skew-robust neural networks*, to reduce such undesirable effects.

We first introduce the basic ideas. Let the classes be

$$\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K$$

where K is the number of classes. Let the number of training examples in these classes be

$$N_1, N_2, \ldots, N_K$$

respectively. In discriminative models such as neural networks, it is common practice to use one-hot (a.k.a. bit vector) representation for class label, and minimize the sum of cross-entropies[1] between targets and outputs, i.e.

$$\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \log y_{nk} \tag{1}$$

where t_{nk} is the target value of class k for example $\mathbf{x}_n$ and y_{nk} is the network output of class k, for training example $\mathbf{x}_n$. Parameters are often updated using error back propagation. The objective function Eq. (1) treats each example in the training set equally. Without further ado, the parameters of the network will be trained to favor the class with the most training examples.

In this paper, in order to make the most use of the precious data of small class and to reduce the swamping effect of large class, we propose to modify the objective function to

$$\sum_{n=1}^{N} \sum_{k=1}^{K} t'_{nk} \log y_{nk} \tag{2}$$

where

$$t'_{nk} = r_{nk} t_{nk} \tag{3}$$

and r_{nk} is the balance factor of class k for training example $\mathbf{x}_n$. More specifically, let $\mathcal{C}_{t(n)}$ be the class of $\mathbf{x}_n$, then we have

$$r_{nk} = \begin{cases} 0, & \text{if } k \neq t(n) \\ \alpha {N_k}^{-1}, & \text{if } k = t(n) \end{cases} \tag{4}$$

Being inversely proportional to N_k in Eq. (4), r_{nk} has the effects of emphasizing the impacts of the errors of small-class examples and of subduing the impacts of the errors of large-class examples during error back-propagation. Through r_{nk}'s, the relative contributions to the gradient of the objective function by examples of different classes are continuously adjusted, which increases the robustness of the training algorithm to skewed data.

2.2. Motivation and Benefits

The proposed objective function Eq. (2) and the original objective function Eq. (1) are different in the effects they create during *error back propagation* for parameter learning. Specifically, using r_{nk} as defined in Eq. (4), we get a better gradient that is emphasized (i.e. more aggressive error back propagation) in the direction to get the output prediction right for an example of a small class.

The proposed method renders off-line data-balancing methods, such as SMOTE [10] for up-sampling or spread sub-sample for down-sampling, unnecessary. *In principle, the proposed method is important as learning, whether by machine or by nature, is often based on highly skewed data sets.* In practice, it is also important as up-sampling and down-sampling methods are time-consuming and introduces random artifacts that are unexpected.

2.3. Speaker Normalization

In order to eliminate the difference due to speakers and other non-emotion factors, we apply cross-speaker histogram equalization (CSHE) to normalize data [8]. The principle of CSHE is outlined as follows. Suppose we have the cumulative distribution function (CDF) $c_Y(y)$ of a random variable Y. Furthermore, suppose we have examples for another random variable X

$$\mathcal{D}_X = \{x_1, .., x_n\}\,. \tag{5}$$

[1] This function is the negative logarithm of the data-likelihood function, i.e. the error function.

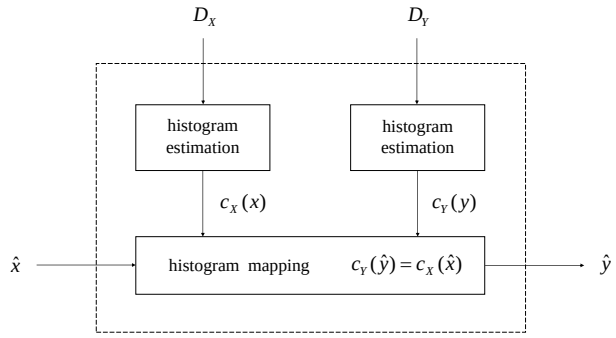

Fig. 2. Cross-speaker histogram equalization.

From $\mathcal{D}_X$, the CDF $c_X(x)$ can be estimated. HE transforms an instance x of X to the value y of Y such that

$$c_Y(y) = c_X(x), \tag{6}$$

that is

$$y = c_Y^{-1}(c_X(x)). \tag{7}$$

Thus, x and y are equalized in their CDF values and they correspond to the same bin in the respective histograms. The implementation of CSHE is illustrated in Figure 2.

FAU-Aibo consists of speech data from 26 speakers for the training set and 25 speakers for the test set. We take all the training data distribution as the target distribution of a virtual speaker, and convert the feature distribution of all speakers to this target distribution. Different versions of CSHE are further specified in Section 3.3.

3. EXPERIMENT

3.1. Data and Experimental Setting

3.1.1. Data

FAU-Aibo corpus contains recordings of spontaneous speech from 51 children as they are interacting with a SONY's pet robot Aibo. The data used in the Challenge consists of 9,959 chunks as training set and 8,257 chunks as test set. For the 5-class classification task, the speech chunks are labelled as anger (A), emphatic (E), Neutral (N), positive (P), and rest (R). Each speech chunk is emotionally-labeled by five persons, and the final label is decided by majority voting. As it turns out, the numbers of chunks of the classes are highly unbalanced: A (8.8%), E (21%), N (56.1%), P (6.8%), and R (7.2%).

3.1.2. Feature Extraction

For speech features, we use the feature set shown in Table 1, including features related to prosody, spectral shape, voice quality, and their derivatives. Specifically, the 16 low-level descriptors (LLDs) are zero-crossing rate (ZCR), root mean square (RMS), pitch frequency (normalized to 500 Hz), harmonics-to-noise ratio (HNR), and mel-frequency cepstral coefficients (MFCC). The delta of LLDs are extracted as well. The 12 functionals are mean, standard deviation, kurtosis, skewness, minimum and maximum values, relative position and range, and the coefficients and the mean squared error (MSE) of linear regression. Thus, the final feature vector contains $16 \times 2 \times 12 = 384$ attributes per speech chunk.

Table 1. Low-level descriptors (LLD) and functionals.

LLDs (16)	Functionals (12)
ZCR	mean
RMS Energy	standard deviation
F0	kurtosis, skewness
HNR	extremes, position, range
MFCC 1–12	regression coefficients, MSE

Table 2. Weights for skew-robust neural networks in the FAU-Aibo Emotional Challenge 5-class task. N_k is the number of chunks for class $\mathcal{C}_k$, and r_{nk} is the weight.

	N_k	r_{nk} for $k = t(n)$
Anger	831	1.1
Emphatic	2093	0.5
Neutral	5590	0.2
Positive	674	1.5
Rest	721	1.4

3.2. Skew-Robust Neural Networks

The weights we use for r_{nk} in Eq. (4) are listed in Table 2. These weights are decided by making them inversely proportional to N_k and having magnitudes near 1. Furthermore, it is guaranteed that errors are magnified in the classes of Positive, Rest, and Anger, which have smaller numbers of examples than Neutral and Emphatic.

The following settings are experimented to investigate the effectiveness of the proposed skew-robust methods.

- SVM: A support vector machine (SVM) system based on importance weights [11].
- NN: A multi-layer perceptron with a hidden layer.
- DNN: A feed-forward deep neural network with 2 hidden layers.

Here SVM is trained by the modified objective function of

$$\|\mathbf{w}\|^2 + C \sum_{n=1}^{N} r_{nt(n)} \xi_n \tag{8}$$

where $\xi_n \geq 0$ is the penalty incurred by the nth data example, $C > 0$ controls the trade-off between the hyperplane margin

Table 3. Unweighted average recall rates (UA) of the FAU-Aibo Emotional Challenge 5-class task with skew-robust neural networks. Please see text for details.

	SVM	NN	DNN
baseline	29.8	**30.1**	28.6
+SR	41.1	**39.6**	39.9
+CSHE+SR	42.6	**45.3**	44.9

and the penalty, and the same weights r_{nk} as listed in Table 2 are used. The topology of MLP and DNN are respectively

$$384 \times 30 \times 5, \quad 384 \times 20 \times 15 \times 5.$$

Experimental results are shown in Table 3. We have the following comments.

- With NN, the proposed algorithm leads to significant improvements from 30.1% UA to 39.6% UA. When combined with CSHE, the performance is 45.3%, which is the best performance we achieve in this work.
- The proposed method outperforms the off-line data-balancing method SMOTE (approximately 39% [1]).
- With SVM, the proposed method achieves an UA of 41.1%, and the combination with CSHE achieves 42.9%. The idea of giving more weights to examples of small classes also works with SVM.
- With DNN, the proposed method achieves an UA of 39.9%, and the combination with CSHE achieves 44.9%. In this case, i.e. adding an additional layer of hidden units to NN, DNN is not better than NN due to the limited amount of training data.

3.3. Speaker Normalization

The following settings are experimented to investigate CSHE in different scenarios. In `basic`, CSHE is not applied to train or test data. In `tr-only`, CSHE is applied to train data, but not to test data. In `one-spkr`, the test data is assumed to be from one virtual speaker when CSHE is applied. In `spkr-kNN`, a test example is assumed to be from the speaker recognized by kNN. In `all`, CSHE is applied to every speaker in training and test set, using ground-truth speaker information. Note that all the above versions do not require speaker information during testing phase, except for the case of `all`, which is applicable only when the speaker information is available, i.e., speaker-dependent speech emotion recognition. Versions of CSHE can be applied to speaker-independent speech emotion recognition. In the case of `spkr-kNN`, we implement a speaker recognition system based on kNN, where the number of voting nearest neighbors is set to $k = 90$. The cumulative distributions of test data in different scenarios of CSHE is shown in Figure 3. These scenarios lead to different empirical data distributions.

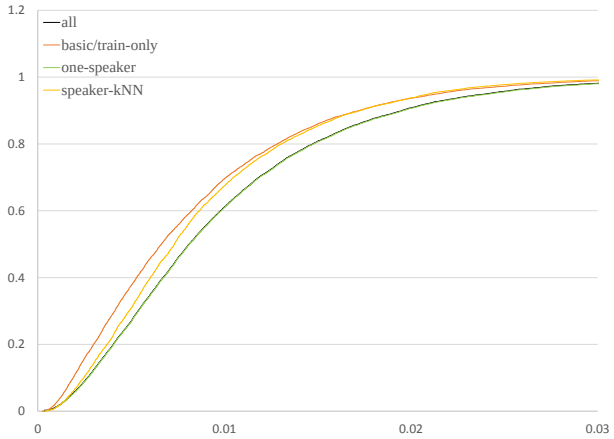

Fig. 3. Cumulative distributions of the "RMS energy mean" feature of the test set data for different scenarios of CSHE.

Table 4. Unweighted average recall rates (UA) of the FAU-Aibo Emotional Challenge 5-class task with different versions of CSHE.

	SVM	SVM+SR	NN	NN+SR
`basic`	29.8	41.1	30.1	**39.6**
`all`	32.3	42.6	33.7	**45.3**
`tr-only`	29.7	40.3	35.4	**42.9**
`one-spkr`	31.1	41.7	29.5	**41.1**
`spkr-kNN`	28.0	37.5	27.9	**40.1**

The results of the FAU-Aibo 5-class task are shown in Table 4. For the proposed skew-robust neural networks, the best performance of UA is achieved in the case `all` at 45.3%. The performance of NN+SR in the case `basic` at 39.6% is improved by every version of CSHE, i.e. `one-spkr` at 41.1%, `tr-only` at 42.9%, and `spkr-kNN` at 40.1%. Thus, the speaker-dependent system is more accurate in emotion recognition than speaker-independent system, which can be improved by unsupervised speaker normalization.

4. CONCLUSION

We investigate a skew-robust parameter-learning method for neural networks. The main idea is to introduce weights to training examples, which effectively modifies the traversed trajectory in the parameter space during the learning process. In addition, we apply cross-speaker histogram equalization method to reduce the emotion expression of different speakers. Evaluated on the FAU-Aibo 5-class task, the proposed methods achieve an unweighted average recall rate of 45.3%. In light of the status quo as reviewed in Section 1, this is one of the best performance in the static modeling framework.

5. REFERENCES

[1] Björn Schuller, Stefan Steidl, and Anton Batliner, "The Interspeech 2009 emotion challenge," in *Proceedings of Interspeech 2009*, pp. 312–315.

[2] Norhaslinda Kamaruddin and Abdul Wahab, "Emulating human cognitive approach for speech emotion using MLP and GenSofNN," in *Proceedings of Information and Communication Technology for the Muslim World 2013*.

[3] Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9, pp. 1162–1171, 2011.

[4] Yazid Attabi and Pierre Dumouchel, "Anchor models for emotion recognition from speech," *Affective Computing, IEEE Transactions on*, vol. 4, no. 3, pp. 280–290, 2013.

[5] Cheng Zha, Ping Yang, Xinran Zhang, and Li Zhao, "Spontaneous speech emotion recognition via multiple kernel learning," in *Proceedings of international Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, 2016.

[6] Longfei Li, Yong Zhao, Dongmei Jiang, Yanning Zhang, Fengna Wang, Ivan Gonzalez, Emmanuel Valentin, and Hichem Sahli, "Hybrid deep neural network–hidden Markov model (DNN-HMM) based speech emotion recognition," in *Proceedings of Affective Computing and Intelligent Interaction 2013*, pp. 312–317.

[7] Vidhyasaharan Sethu, Eliathamby Ambikairajah, and Julien Epps, "Speaker normalisation for speech-based emotion detection," in *Proceedings of International Conference on Digital Signal Processing 2007*, pp. 611–614.

[8] Bo-Chang Chiou, "Cross-lingual automatic speech emotion recognition," M.S. thesis, National Sun Yat-sen University, 2014.

[9] Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011.

[10] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Artificial Intelligence Research*, pp. 321–357, 2002.

[11] Andrew Rosenberg, "Classifying skewed data: Importance weighting to optimize average recall," in *Proceedings of Interspeech 2012*, pp. 2242–2245.