SPEECH EMOTION RECOGNITION WITH ENSEMBLE LEARNING METHODS

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

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

Chung-Hsien Wu[†]

National Chung Kung Univerisity
Computer Science and Information Engineering
Tainan, Taiwan, ROC

ABSTRACT

In this paper, we propose to apply ensemble learning methods on neural networks to improve the performance of speech emotion recognition tasks. The basic idea is to first divide unbalanced data set into balanced subsets and then combine the predictions of the models trained on these subsets. Several methods regarding the decomposition of data and the exploitation of model predictions are investigated in this study. On the public-domain FAU-Aibo database, which is used in Interspeech Emotion Challenge evaluation, the best performance we achieve is an unweighted average (UA) recall rate of 45.5% for the 5-class classification task. Furthermore, such performance is achieved with a feature space of 40-dimension. Compared to the baseline system with 384-dimension feature vector per example and an UA of 38.9%, such a performance is very impressive. Indeed, this is one of the best performances on FAU-Aibo within the static modeling framework.

Index Terms— speech emotion recognition, ensemble learning, representation learning

1. INTRODUCTION

As speech technology advances, the recognition of all sorts of patterns conveyed in speech signals that help to identify the hidden states of the speakers becomes more and more useful. Thus, in addition to linguistic units such as phonemes and words, research efforts have been made for the recognition of information regarding speakers, languages, and emotions in recent years. In this paper, we focus on the application of speech emotion recognition (SER). An SER system takes a speech signal as input, and outputs one of the emotional categories known to the system and hypothetically conveyed in the speech. SER is essentially a *classification* problem, which is fundamental in machine learning. Global evaluation plans have been implemented to promote researches on SER, e.g. the Interspeech 2009 Emotion Challenge (henceforth referred to as the Challenge) is a large-scale evaluation plan to advance the technology of speech emotion recognition using the FAU-Aibo database [1]. Following the Challenge, a comparison of the systems submitted to the Challenge is presented in [2].

Many approaches to improve SER performance on FAU-Aibo have been proposed since the Challenge. To name a few, there have been techniques such as multi-layer perceptrons (MLP), feature ranking and segregation, support vector machines (SVM), anchor models, multiple kernel methods, hierarchical frameworks, hybrid systems, and histogram equalization [3, 4, 5, 6, 7, 8].

In spite of endeavors by the community of SER researchers, the performance levels for the tasks as defined in the Challenge are still far from decent. The highest unweighted average (UA) recall rate among submissions to the Challenge was 41.7% [2]. Following the Challenge, more advanced methods have also been proposed to improve the performance for the 5-class task. As far as we know, the highest UA achieved is 44.0% [5] for the static modeling framework in which each speech chunk is represented by a fixed-size vector, and 45.6% [7] for the dynamic modeling framework in which each speech chunk is represented by variable-length feature vectors. It is fair to say that there is still much room for improvement for this task.

Two important reasons for such difficulty, we believe, are the *issue of skewed database* and the *intrinsic ambiguity of emotion expression*. Both issues are quite worthy of research efforts as they are *general challenges in machine learning*. Skewed data issue is not an unusual scenario nowadays, as data collection processes for machine-learning systems are often automated that hopefully require as little human intervention as possible, so the collected data is bound to be unbalanced. The uncertainty of data labels is also commonplace nowadays. As classification tasks move from areas of well-defined classes to uncharted territories, such uncertainty in the labels is bound to happen either due to crowd-sourcing or the intrinsic ambiguity among target classes. Thus, systematic approaches to deal with skewed data or labeling uncertainty do have great utility.

In this paper, we propose to *apply ensemble-learning methods* to ameliorate the issue of skewed data. The idea is to divide the set of examples of the largest emotional class into

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

smaller sets, so the training data for a classifier is balanced. This is *different* from common methods, such as SMOTE [1], that increase the number of data points to balance data sets. Furthermore, as part of the ambiguity of emotion expression comes from the difference between speakers, we also apply *cross-speaker histogram equalization* [8] to reduce such difference.

The rest of this paper is 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. Concluding remarks are given in Section 4.

2. PROPOSED METHOD

Naturally collected databases often have issues of highly skewed data. The distribution of data examples in FAU-Aibo is shown in Figure 1. In fact, a naïve classifier that simply assigns each test example to the Neutral 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 recall rate of each class. The above naïve classifier without any training would achieve an accuracy of 20% for UA, which is more reasonable than WA.

2.1. Ensemble Learning

With FAU-Aibo, a classifier trained with the skewed training data would inaccurately favor the high-population Neutral class. To deal with the skewed data issue, we propose to use ensemble learning method. Ensemble learning is a feasible and often-working method to improve the performance of a machine-learning system: train different systems and combine their results. The proposed ensemble-learning system consists of several component classifiers, each of which is trained with balanced train data. During testing, the outputs of the component classifiers are combined for final prediction. Different combination methods are further detailed in Section 3.

2.2. Speaker Normalization

In order to eliminate the differences between speakers and other non-emotion factors, we apply cross-speaker histogram equalization (CSHE) to normalize data. 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 a set of examples for another random variable X

$$\mathcal{D}_X = \{x_1, .., x_n\}. \quad (1)$$

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

$$c_Y(y) = c_X(x), \quad (2)$$

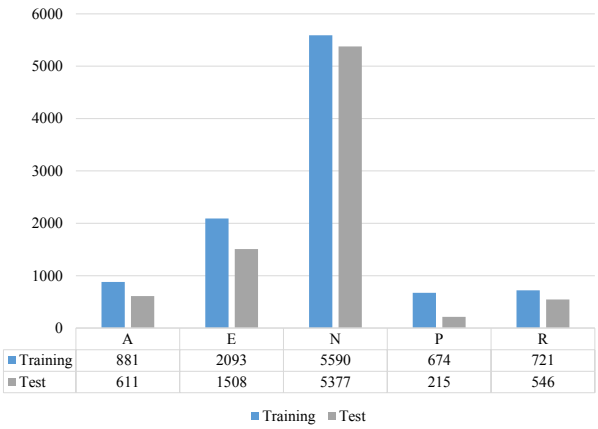

Fig. 1. Decomposition of class-by-class examples in FAU-Aibo corpus: anger (A), emphatic (E), Neutral (N), positive (P), and rest (R).

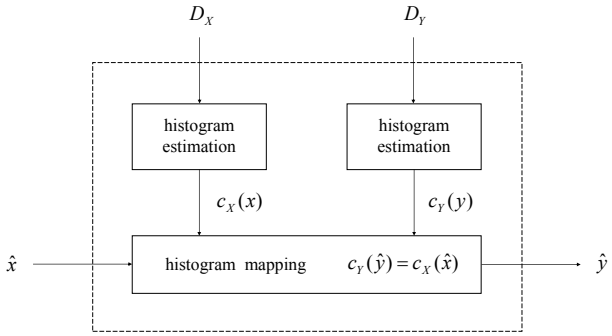

Fig. 2. Cross-speaker histogram equalization.

that is

$$y = c_Y^{-1}(c_X(x)). \quad (3)$$

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.

2.3. Representation Learning

In this work, we use neural networks to learn proper representation for data, and then use an additional neural network to learn prediction function based on the learned representation. A good example in natural language processing of representation learning is to use Skip-gram model to learn word-embedding vectors, on which recurrent neural networks for language model are built [9].

In this paper, the predictions of the component classifiers are concatenated and then used as features to train a classifier. That is, instead of using the maximum or the average of output probabilities of the component classifiers to decide

the class of a data point, we treat the ensemble learning system as a pre-processing stage to learn to extract better feature representation.

The proposed representation-learning system is *two-pass* and *supervised*. It is two-pass because the feature-extractor learning and the classifier learning modules are not jointly optimized. Thus, it is *different* from deep learning systems, such as neural networks with hidden layers or graphical models with hidden variables, which use one-pass learning methods. It is also *different* from the *pre-train* methods, which use unsupervised learning methods.

3. EXPERIMENT

3.1. Data and Experimental Setting

The FAU-Aibo corpus contains recordings of spontaneous speech in German from 51 children interacting with 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 tasks, the emotion labels are Anger (A), Emphatic (E), Neutral (N), Positive (P), and Rest (R). Class-wise, the data is highly unbalanced: A (8.8%), E (21%), N (56.1%), P (6.8%), and R (7.2%).

Following the same experimental setting specified by the Challenge, we use the baseline feature set shown in Table 1, including common 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) energy, pitch frequency (normalized to 500 Hz), harmonics-to-noise ratio (HNR), and mel-frequency cepstral coefficients (MFCC). For these LLDs, the delta coefficients are additionally computed. The 12 functionals are mean, standard deviation, kurtosis, skewness, minimum and maximum values, relative position and range, as well as the coefficients and the mean squared errors (MSE) of linear regression. Thus, the feature vector contains a fixed size of

$$16 \times 2 \times 12 = 384$$

elements per speech chunk.

We use the OpenSMILE [10] toolkit for feature extraction. The features out of the feature extractors are referred to as the basic features. They may be converted to CSHE features via CSHE. We use the deep-learning software tool Theano [11] for classifiers based on artificial neural networks with network topology

$$384 \times 30 \times 5$$

and the software tool Weka [12] for SVM.

3.2. Combination of Model Predictions

The following methods for the combination of neural network outputs (predictions) have been experimented in this paper.

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. Unweighted average (UA) recall rates for the FAU-Aibo 5-class tasks of the proposed ensemble-learning methods. Here case `basic` means CSHE is not applied to the features. The other cases are stated in the text.

	`basic`	CSHE
`baseline`	30.1	**33.7**
SMOTE	40.7	43.2
sub-sample	41.3	44.1
ensemble-Max	40.1	44.2
ensemble-Avg	42.2	**45.0**
ensemble-Log	42.7	44.9

- ensemble-Max: For each class, the maximum of the output probabilities of the component classifiers is the ensemble output.
- ensemble-Avg: For each class, the average of the output probabilities of the component classifiers is the ensemble output.
- ensemble-Log: For each class, the average of the logarithm of the output probabilities of the component classifiers is the ensemble output.

The results for ensemble-learning are summarized in Table 2. In this table, `baseline` refers to the systems trained with the original training data, which is unbalanced. SMOTE refers to the systems trained with balanced training data in which the size of the small classes are increased with SMOTE. Case "sub-sample" refers to systems trained with balanced training data in which the size of the large classes are reduced by ignoring randomly selected examples. From this table, we can see that the ensemble-learning method via random selection improves the UA from 33.7% to 45.0% when CSHE is applied. Furthermore, the ensemble-learning methods outperform the common data-balancing methods of SMOTE (43.2%) and spread sub-sample (44.1%).

3.3. Decomposition of Data for Ensemble Learning

The following methods for the separation of Neutral-class data (5,590 examples) into subsets have been experimented.

- random: Random selection of examples without replacement. This is repeated for each subset until each

Table 3. Unweighted average (UA) recall rates for the FAU-Aibo 5-class tasks of the proposed ensemble-learning using different grouping methods.

	random	k-means	k-means-B
ensemble-Max	44.2	41.3	42.0
ensemble-Avg	**45.0**	43.0	43.1
ensemble-Log	44.9	42.9	42.9

subset has a pre-determined number of examples.[1]

- k-means: Cluster data into k subsets first, followed by random selection. Note that some subsets may have fewer data points than a pre-determined number, if the corresponding clusters are small.
- k-means-B: To compensate for the small subsets, we re-assign surplus examples in large subsets to small subsets. This is a *twisted* k-means to guarantee the number of examples in each subset meets a pre-determined number.

For the other classes, the numbers of examples are reduced to be the same as the number of examples in Positive-class (674) by random selection.

The results of different grouping methods are summarized in Table 3. We can see that the random sub-sample grouping scheme has better performance than the others for ensemble-learning systems (45.0%). The difference between random sub-sample and k-means clustering is that the resultant groups of Neutral data by random sub-sample are more diverse than clustered groups. Thus, even in the scenario of ensemble of classifiers, it appears to be important to have diverse data points in each component classifier.

3.4. Experiments on Representation Learning

In this section, we describe how we achieve our best performance for 5-class classification tasks through a representation-learning technique. We turn the above ensemble-learning system with random sub-sample grouping into a representation-learning system by using the outputs of the component neural networks as features. These features are then fed into a neural network for classification. For C component classifiers with K classes, the learned representation is a feature vector of KC dimensions. In the current situation we have $K = 5$ and $C = 8$, so the learned representation is 40-dimension, which is smaller than the dimensionality of 384 in the original representation.

In the following experiments, the grouping of Neutral-class data for the component classifiers are based on random sub-sample. After the component classifiers are trained, every data point is passed through the component classifiers to create a 40-dimension feature vector, which is then used as an input vector to MLP and SVM. The number of examples in each class are not balanced, as class Neutral has 8 times the amount of data of the other classes. To deal with data imbalance, we apply a label-dependent weighting scheme which is inversely proportional to the number of examples of a class

$$r_{nk} \propto N_k^{-1}$$

during network training. These weighting factors effectively magnify the errors of the rare examples for error back-propagation.

[1]Note that this is the same setting as used in the experiments whose results are summarized in Table 2.

Table 4. Unweighted average (UA) recall rates for the FAU-Aibo 5-class tasks with representation-learning methods.

	`basic`	CSHE
SVM	40.2	**45.5**
MLP	41.3	44.5

The results are summarized in Table 4. The SVM using the learned features achieves 45.5% UA, which is our best performance. Note the following facts.

- This performance is one of the best results achieved in the static modeling framework as is reviewed in Section 1.
- The MLP achieves 44.5% UA, which is still better than sub-sample or SMOTE (44.1% and 43.2% as shown in Table 2), showing the benefits of ensemble learning.

4. CONCLUSION

In this paper, we apply ensemble learning methods to improve the performance of speech emotion recognition. The success is due to using small balanced data subsets in multiple component classifiers, instead of using a large unbalanced data set in a single classifier. Using ensemble-learning method alone, the best UA is 45.0% for the 5-class classification task on FAU-Aibo database. When we use the outputs of the component classifiers as features, the performance is improved to 45.5%, which is one of the best performances on FAU-Aibo within the static modeling framework. These results clearly show that the proposed methods work well for speech emotion recognition.

For future work, we believe learning methods which are robust to label errors should be investigated. Instead of giving a hard (deterministic) target for each training example, giving a soft (probabilistic) target to accommodate the uncertainty of label would surely be more flexible. Such methods would be most useful when the signals are noisy or when the categories are themselves ambiguous.

5. REFERENCES

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

[2] 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.

[3] 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*.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.

[10] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: the Munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, 2010, pp. 1459–1462.

[11] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," May 2016.

[12] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.