SOFT-TARGET TRAINING WITH AMBIGUOUS EMOTIONAL UTTERANCES FOR DNN-BASED SPEECH EMOTION CLASSIFICATION

Atsushi Ando, Satoshi Kobashikawa, Hosana Kamiyama, Ryo Masumura, Yusuke Ijima and Yushi Aono

NTT Media Intelligence Laboratories, NTT Corporation, Japan

ABSTRACT

This paper presents a novel emotion classification method for natural speech. One of the problems in the state-of-the-art method based on Deep Neural Network (DNN) is the paucity of the training data compared to model complexity. To solve this problem, this paper utilizes the *ambiguous* emotional utterances, utterances that have no dominant target emotion label. While previous work ignored *ambiguous* emotional utterances for training, the proposed method leverages all annotated labels via soft-target training. In addition, this paper modifies the soft-target training in order to effectively handle both *clear* and *ambiguous* emotional utterances. Experiments show that the proposed method yields performance improvements in terms of both weighted and unweighted accuracies.

Index Terms— Speech emotion recognition, LSTM with attention, soft-target, ambiguous emotional utterance

1. INTRODUCTION

Speech emotion recognition is an important technology to understand natural human conversations because it helps to convey actual messages. It has many applications such as effective voice-of-customer analysis in contact center calls [1] and better understanding of human requests in spoken dialog systems [2]. Though there are two types of challenges, classification of categorical emotions and regression of dimensional emotions [3], the aim of this paper is emotion classification from acoustic information for natural speech.

Many studies on speech emotion classification have been published. The most traditional methods are based on heuristic features such as utterance mean of fundamental frequency and loudness [4]. However, it is difficult to find optimal features for classifying target emotions, which restricts the performance of heuristic feature-based methods. Recently, several researchers have been attempting to acquire optimal features automatically by Deep Neural Network (DNN) [5–8]. The first approach estimates emotions frame-by-frame by DNNs, then integrates the frame-level results to get an utterance-level emotion [5]. Bidirectional Long Short-Term Recurrent Neural Networks (BLSTM-RNNs) are used to utilize much longer contexts for frame-level emotion estimation [6]. The state-of-the-art methods use an attention mechanism with BLSTM-RNN to achieve both frame-level feature extraction and feature-integration for utterance emotion classification at the same time [7, 8]. These approaches can classify emotions by utilizing the specific parts of an utterance that strongly suggest emotional characteristics.

Though DNN-based methods offer great improvement, there is a common problem: paucity of the training data. The model of the conventional methods has so many parameters that large training data sets are required, but the training data for speech emotion classification is usually small. This data limitation decreases the generalization performance, which degrades estimation performance.

One reason for the data limitation is that conventional methods use only *clear* emotional utterances, those that have a dominant target emotion label, for training. However, natural human conversations contain a considerable amount of *ambiguous* emotional utterances in which none of the target emotion labels are dominant. In this paper, we utilize these *ambiguous* emotional utterances to solve the data paucity problem. We assume that not only *clear* but also *ambiguous* emotional utterances express some characteristics of the target emotions. This paper is an initial work of employing *ambiguous* emotional utterances for emotion classification.

In order to utilize *ambiguous* emotional utterances, we focus on soft-target training of DNNs. While the soft-target training was only applied to *clear* emotional utterances [9, 10], this paper applies it to *ambiguous* ones. In addition, this paper modifies the soft-target training in order to handle both *clear* and *ambiguous* utterances effectively. The proposed method with only *ambiguous* emotional utterances yields accuracy approaching that of the conventional method with *clear* emotional utterances. Furthermore, the proposed method with both *clear* and *ambiguous* emotional utterances attains greatly improved classification accuracy.

2. DNN-BASED EMOTION CLASSIFICATION

In this section, we describe the state-of-the-art emotion classification based on BLSTM-RNNs with attention mechanism [8]. It estimates the posterior probabilities of emotions

by utilizing some parts of the characteristics of an utterance that strongly suggest emotional characteristics.

2.1. Overiview

Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ be the feature sequence of an utterance and $c \in \{c_1, \cdots, c_K\}$ is the correct emotion of the utterance; T is the length of the sequence and K is the total number of target emotion classes. The estimated emotion, $\hat{c}$, is obtained by directly evaluating the posterior probabilities of the emotions;

$$\hat{c} = \arg\max_{c_k} p(c_k|\mathbf{X}, \theta), \tag{1}$$

where θ is the set of the parameters in the emotion classification model. A model structure of the BLSTM-RNNs with attention is shown in Fig. 1. α_t means the attention value and $\mathbf{u}$ is the context vector used to calculate α_t. The output vector $\mathbf{y}$ shows the posterior probabilities $[p(c_1|\mathbf{X}, \theta), \ldots, p(c_K|\mathbf{X}, \theta)]$.

The model parameters are updated by the loss function L based on softmax cross entropy,

$$L(\theta; \mathbf{X}, c) \quad \equiv \quad -\sum_{k=1}^{K} q(c_k) \log p(c_k|\mathbf{X}, \theta), \tag{2}$$

where $q(c_k)$ is the reference class distribution.

2.2. Hard-target training

The conventional method regards the majority of the annotated emotion labels as the ground truth. In this case, the reference class distribution is represented as,

$$q(c_k) = \begin{cases} 1 & \text{if} \quad k = \arg\max_k \frac{\sum_n h_k^{(n)}}{\sum_{k'} \sum_n h_{k'}^{(n)}} \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

where $h_k^{(n)}$ is the binary label-existence which is 1 if the n-th annotator gives class label c_k, otherwise 0. Note that the utterances that have no majority target emotion are excluded from the training data.

2.3. Problems

Though the conventional method offers good performance, one problem remains; the paucity of the training data relative to model complexity. BLSTM-RNN structure uses many parameters to describe complex contextual information. However, the size of the training data is usually limited in speech emotion classification tasks. This discrepancy decreases classification performance.

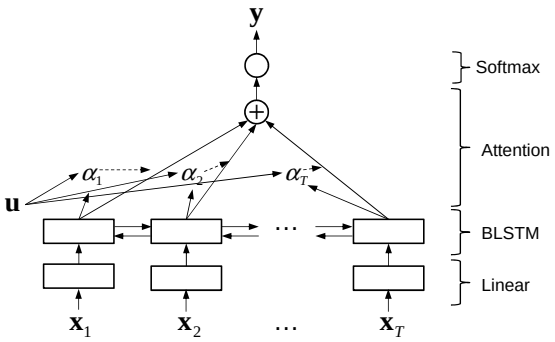

Fig. 1. An example of the structure of the emotion classification model based on BLSTM-RNNs with attention [8].

3. SOFT-TARGET TRAINING WITH AMBIGUOUS EMOTIONAL UTTERANCES

3.1. Approach

For the following explanation, we describe emotional utterances as two types: *clear* or *ambiguous*. *clear* emotional utterances mean the utterances in which more than 50% of the annotated emotion labels are same. *ambiguous* emotional utterances are those that have at least one of the target emotion labels but are not *clear*.

One reason for the paucity of the training data is that conventional methods use only *clear* utterances of the target emotions. Due to that emotion labels given by annotators usually vary greatly in natural speech, ground truth emotions of utterances have to be defined for emotion classification. There are two types of definitions. The first, used in most emotion classification methods, is that individual utterances have only one ground truth emotion. In this case, *ambiguous* emotional utterances are regarded as no ground truth and excluded from the training dataset. The second are that utterances contain one or more of ground truth emotions [9–11]. They regard the frequencies of emotion labels given by all annotators as indicative of the ground truth, and are implemented as a soft-targets. However, all of the previous works of the soft-targets also eliminates *ambiguous* emotional utterances for training. We consider that they implicitly regard that there are mislabeled emotions which are not suitable for training. In either cases, there are the utterances which are annotated but eliminated from training, which decreases the amount of the training data.

However, natural speech contains a lot of *ambiguous* emotional utterances. A small analysis is shown to quantify this understanding; we used the human dialogue dataset named Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [12]. It holds approximately 12 hours of dyadic interactions by 10 speakers with emotional expressions. Each utterance was given one or more of 10 emotional labels by

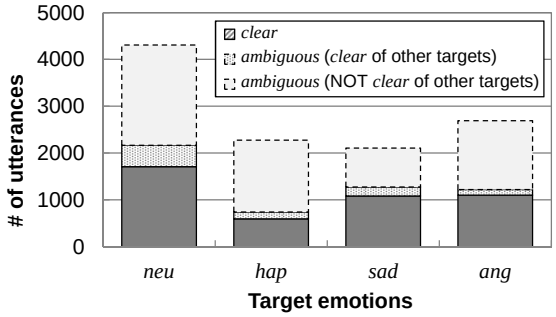

Fig. 2. Number of *clear* and *ambiguous* emotional utterances in IEMOCAP corpus [12].

each of three annotators. Total number of utterances was 10039 and we selected *neutral, happy, sad, angry* as the target emotions. The numbers of *clear* and *ambiguous* emotional utterances are shown in Fig. 2. The broken line parts, the utterances which include at least one specific emotion labels but eliminated from the training data, is almost the same amount as *clear* emotional utterances. Therefore there are many utterances which are not utilized in the conventional training dataset.

In this paper, we have a new ground truth hypothesis which mitigates the data paucity problem. Our definition is the same as those of the soft-target approaches except that there are no mislabeled emotions. From this viewpoint, not only *clear* but also *ambiguous* emotional utterances express emotional characteristics and so are valuable for training.

In order to utilize *ambiguous* utterances, we apply the soft-target training. In addition, this paper modifies it in order to handle both types of emotional utterances effectively.

3.2. Conventional soft-targets

Soft-targets can describe the reference intensities of the target classes, and hence are suitable to represent *ambiguous* emotions. This is also used in distillation [13] which is a famous technique for DNN.

In emotion classifications, soft-targets are calculated by annotated labels [9, 10],

$$q(c_k) = \frac{\sum_n h_k^{(n)}}{\sum_{k'} \sum_n h_{k'}^{(n)}}. \quad (4)$$

The soft-targets are used in the same loss function Eq. (2) to update parameters.

3.3. Modified soft-targets

There is a problem in applying conventional soft-targets to *ambiguous* emotional utterances. They allocate the same reference distributions for *ambiguous* and *clear* emotional utterance when they have one and the same kind of the target emotion labels, as shown in Table 1.

Table 1. An example of the hard-target, soft-targets and modified soft-targets. The leftmost column means the emotion labels given by three annotators and the rest show the reference distributions of the target emotions [*neu, hap, sad, ang*]. In the modified soft-targets, the smoothing coefficient $\alpha = 1$.

Annotation	hard-target	soft-targets	modified soft-targets
{*neu, neu, neu*}	$[1, 0, 0, 0]$	$[1, 0, 0, 0]$	$[\frac{4}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}]$
{*neu, neu, sad*}	$[1, 0, 0, 0]$	$[\frac{2}{3}, 0, \frac{1}{3}, 0]$	$[\frac{3}{7}, \frac{1}{7}, \frac{2}{7}, \frac{1}{7}]$
{*neu, fru, fru*}	*no use*	$[1, 0, 0, 0]$	$[\frac{2}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}]$

It is desirable to distinguish *ambiguous* emotional utterances from those of *clear*. Hence we propose a new soft-target variant named modified soft-targets to handle both types of emotional utterances properly. The modified soft-targets are the additive smoothed form of the conventional soft-targets;

$$q(c_k) = \frac{\alpha + \sum_n h_k^{(n)}}{\alpha K + \sum_{k'} \sum_n h_{k'}^{(n)}}, \quad (5)$$

where α is the smoothing coefficient and the modified soft-targets equal the conventional ones if $\alpha = 0$. This gives more flattened reference distributions in *ambiguous* emotional utterances.

The modified soft-targets can be regarded as Maximum a Posteriori (MAP) estimation by annotated labels with uniform prior distribution, while the conventional is Maximum Likelihood (ML) estimation. In general, the performance of MAP estimation is more robust given small sample size than ML estimation. Hence the proposed method is suitable for representing emotional soft-targets.

4. EXPERIMENTS

4.1. Setup

Speaker-independent speech emotion classification experiments were conducted to evaluate the proposed method against the IEMOCAP database. The number of target emotions was four; *neutral*, *happy*, *sad*, and *angry*. 8 speakers (4 males and females) were selected for the training set and the remaining 2 speakers (1 male and female) were used as the test set. The training set was divided into *clear* and *ambiguous* sets. *clear* set includes *clear* emotional utterances of the target emotions, and *ambiguous* set has those of *ambiguous*. The numbers of the utterances in each set are shown in Table 2.

47 dimensional acoustic features were extracted as frame-level utterance features; 12 dimensional Mel-Frequency Cepstral Coefficients (MFCCs), loudness, fundamental frequency (F_0), voice probability, zero cross rate, Harmonics-to-Noise Ratio (HNR), the first order derivatives of them,

Table 2. Number of the utterances in the dataset. The first four of the major emotion classes were the target emotions, while the rests, *frustration, excitement, surprised, fear, disgust, other, xxx* (no major) were not.

	Total	Major emotion class										
		neu	*hap*	*sad*	*ang*	(*fru*)	(*exc*)	(*sur*)	(*fea*)	(*dis*)	(*oth*)	(*xxx*)
Train - *clear*	3548	1324	460	890	874	-	-	-	-	-	-	-
Train - *ambig.*	3693	0	0	0	0	1049	568	21	11	0	2	2042
Test	942	384	135	194	229	-	-	-	-	-	-	-

and the second order derivatives of MFCCs and loudness. The frame length and frame shift were 20 ms and 10 ms, respectively. All features were extracted by openSMILE [14]. These were normalized by mean and standard deviation of all of the utterance features in the training set. Finally, the features of every 4 frame sequence were used as the input feature sequence of the model.

BLSTMs with attention mechanism were used in both the baseline and the proposed method. The model was composed of fully-connected layers with Rectified Linear Unit (ReLU) activation function, BLSTMs with attention layer and fully-connected layers with softmax function. The number of hidden units was 256 in the first fully-connected layer, 128 in BLSTMs and 256 in the last fully-connected layer. Dropout rate was 50% in all layers. Two methods were evaluated as the baseline; hard-target training and the conventional soft-target training [9]. The proposed method was the modified soft-target training with smoothing coefficient $\alpha = 0.75$. Both baselines used the *clear* set for training, while the proposed used either or both *clear* and *ambiguous* set. Early-stopping criteria were applied in all conditions to minimize the loss of the validation set. In this paper we used testset as the validation set in the same way as [9] because of the limitation of the dataset.

Two common evaluation measures were used in the experiments: weighted accuracy (WA) and unweighted accuracy (UA). WA is the overall accuracy and UA is average recall over every emotional category. In both cases, the majority annotated class labels are regarded as the true emotions in the test set. We made five trials of model training and evaluation, and averaged WA and UA were taken as the final result.

4.2. Results

Performance comparisons of the baseline and the proposed method are shown in Table 3. The performance of the proposed method with only *ambiguous* set was close to that of the baseline via hard-target with *clear* set, even though their training set had no *clear* emotional utterances. This indicated that *ambiguous* emotional utterances certainly contain common cues of the target emotions. Next, comparing the proposed modified soft-targets to the two baselines with *clear* set, shows that WA scores were equivalent while those UA were better. Finally, the modified soft-targets with both *clear* and *ambiguous* datasets dominated the baseline in terms of relative error reduction; 9.7% in WA and 21.6% in UA. These results indicate that both *clear* and *ambiguous* emotional utterances are useful for model training.

Table 3. Accuracy comparison between the baseline and the proposed method.

Method	Label	Train set *clear*	Train set *ambig.*	Acc. [%] WA	Acc. [%] UA
Baseline	hard	✓		58.6	53.7
	soft	✓		58.1	54.9
Proposed	modified soft	✓		58.5	57.4
			✓	53.6	54.0
		✓	✓	**62.6**	**63.7**

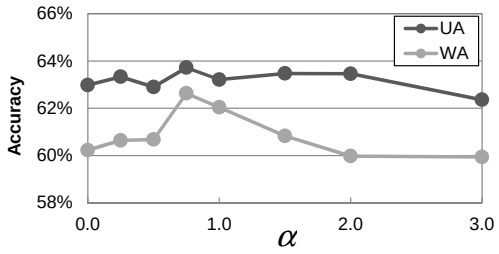

Fig. 3. Accuracies of the modified soft-targets with both *clear* and *ambiguous* training set.

In addition, we compared the conventional soft-targets with the modified soft-targets. The result of $\alpha = 0$ in Fig. 3 represents the conventional soft-target performance while $\alpha > 0$ represents the modified. In the figure, the soft-targets from $\alpha = 0.75$ to 1.5 showed better performance than the conventional method, $\alpha = 0$. Thus the modified soft-targets are more suitable for emotion classification.

5. CONCLUSIONS

In this paper, we proposed a novel speech emotion classification method that can leverage *ambiguous* emotional utterances for training. The proposed method focused on the soft-target training to utilize *ambiguous* utterances. In addition, this paper modified the conventional soft-targets in order to effectively handle both *clear* and *ambiguous* emotional utterances. Experiments showed that the proposal yields performance improvements in terms of both weighted and unweighted accuracies.

6. REFERENCES

[1] P. Gupta and N. Rajput, "Two-stream emotion recognition for call center monitoring," in *Proc. of INTERSPEECH*, 2007, pp. 2241–2244.

[2] S. Fujie, Y. Ejiri, Y. Matsusaka, H. Kikuchi, and T. Kobayashi, "Recognition of para-linguistic information and its application to spoken dialogue system," in *Proc. of ASRU*, 2003, pp. 231–236.

[3] M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[4] R. Banse and K. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Psychology*, vol. 70, no. 3, pp. 614–636, 1996.

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

[6] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Proc. of INTERSPEECH*, 2015.

[7] C. W. Huang and S. S. Narayanan, "Attention assisted discovery of sub-utterance structure in speech emotion recognition," in *Proc. of INTERSPEECH*, 2016, pp. 1387–1391.

[8] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. of ICASSP*, 2017, pp. 2227–2231.

[9] H. M. Fayek, M. Lech, and L. Cavedon, "Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels," in *Proc. of IJCNN*, 2016, pp. 566–570.

[10] S. Steidl, M. Levit, A. Batliner, E. Noth, and H. Niemann, "Of all things the measure is man: Automatic classification of emotions and inter-labeler consistency," in *Proc. of ICASSP*, 2005, pp. 317–320.

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

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

[13] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[14] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proc. of ACM Multimedia*, 2010, pp. 1459–1462.