THE GENERALIZATION EFFECT FOR MULTILINGUAL SPEECH EMOTION RECOGNITION ACROSS HETEROGENEOUS LANGUAGES

Shi-wook Lee

National Institute of Advanced Industrial Science and Technology, Japan

ABSTRACT

Regularization approaches, such as multi-task learning and dropout, prevent overfitting and improve generalization ability. Speech emotion recognition suffers from insufficiently transcribed databases, where labels are subjectively annotated. Because emotions are a more universally recognized language, the paralinguistic feature space of emotional speech can be better generalized, even across substantially heterogeneous languages. We investigate the effect of regularization and normalization frameworks on two emotional speech databases, the IEMOCAP for English and the JTES for Japanese. We obtain absolute gains of unweighted average recall over ten runs (1.48% for the IEMOCAP and 1.03% for the JTES) and achieve a maximum of 59.49% on the IEMOCAP. From comparative experiments, we confirm that dropout and multi-task learning strategies are effective for multilingual speech emotion recognition, and common normalization over two languages leads to further improvement under all conditions, which suggests that better generalization is available even when two highly heterogeneous languages are merged.

Index Terms— speech emotion recognition, data normalization, generalization, multi-task learning, multilingual

1. INTRODUCTION

Speech emotion recognition (SER) is essential for empathic human-machine communication. Since the dictation performance of automatic speech recognition (ASR) has reached the stage of practical use, there has been growing interest in how to make machines more human-like and how to enhance smart communication between human and machine. However, in contrast to ASR, SER still suffers from low resources. Emotional speech databases are extremely limited, as emotions in speech are expressed and perceived subjectively. Whether emotions serve as a universal language is still debated in the psychology and cognitive neuroscience fields, depending on which indicators researchers used to measure emotions - verbal or nonverbal expressions, facial, vocal, or gestural signals, and expressions or perceptions [1]. Setting aside this debate, it is worth considering how emotions expressed in different languages can be integrated into SER tasks. Unlike linguistic information, a paralinguistic feature space can be implemented robustly across languages. Through a series of INTERSPEECH Paralinguistic Challenges, paralinguistic information is enriched and widely adopted to improve SER performance [2, 3, 4]. Furthermore, recent deep-learning-based approaches have yielded remarkable improvements in SER performance [5, 6, 7, 8, 9, 10]. In [5], researchers adopted deep neural network (DNN) for SER using a generalized discriminant analysis; the results on nine databases showed significant improvement over support vector machines (SVM). In [6], the authors proposed using a DNN to estimate emotion states for each speech segment in an utterance, constructing an utterance level feature from segment-level estimations and then employing an extreme learning machine to SER. Their experimental results indicated that a DNN-based approach substantially improved the performance of SER. In [7], the authors compared recurrent neural networks (RNN)-based SER system with a DNN-based SER system. In [8], researchers also reported RNN-based transfer learning from dimensional to categorical emotion attributes of a single emotional speech database. In [9], the authors compared several DNN-based and RNN-based systems over SVM-based SER. In [10], researchers proposed attentive convolutional neural network (CNN) for SER, which combines the benefits of CNN and attention mechanisms. This previous research clarifies that a large amount of data is needed to improve SER performance, especially in the case of deep-learning-based approaches. Consequently, we investigate the generalization effect for multilingual SER, using a recently released, large-scale Japanese database, the Japanese Twitter-based Emotional Speech (JTES) [11], and a widely used English database, the Interactive Emotional Dyadic Motion Capture (IEMOCAP) speech database [12].

2. RELATED WORK

In this work, we investigate multilingual SER across two highly heterogeneous languages, English and Japanese. Emotions are influenced by culture and society, and therefore emotional speech databases have been developed and constructed heterogeneously [5, 13, 14]. Thus, multilingual SER is a challenging but implementable task. In [15], researchers applied data normalization via histogram equalization to remove cross-speaker and cross-language variability for SER. In [16], the authors presented a comprehensive overview us-

ing eight languages from four language families, showing that cross-language SER is feasible but has notably lower accuracy than mono-lingual SER. In [13], researchers first applied automatic language identification for model selection and then performed SER on a language-dependent model. Most recently, [17] explored cross-lingual and multilingual speech emotion recognition in English and French; they found that multilingual SER was feasible without adaptation to the language and presented promising results for cross-lingual training, followed by fine-tuning of the target language.

Another difficulty of SER is that classifiers are easily overfitted because there is so little training speech, much less than that for current ASR. Multi-task learning (MTL) is introduced to improve generalization ability [18] and has also been recently adopted in SER [19, 20, 21, 22]. In [19, 20], the multi-task is composed of arousal, valence, and dominance, three dimensional components of emotion. In [21], researchers adapted shared hidden layers (SHLs) to the task of SER, setting the multi-task with nine heterogeneous, emotion-related tasks in three different languages (English, German, and Danish). However, as their multi-task is composed only of emotion-based components, SHLs can be generalized over various languages but may be specified on emotion. In [22], the authors built an attention-based, weighted pooling framework with MTL for emotion recognition, where the multi-task is composed of emotion, speaker, and gender but is trained only in one database. In this work, we set three different tasks in MTL - language, gender, and emotion classifications - using two highly heterogeneous languages, English and Japanese.

3. GENERALIZATION METHODS

3.1. Multi-task learning

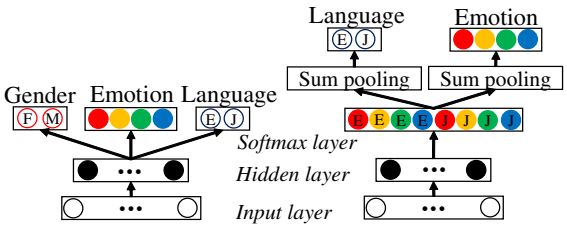
Multi-task learning (MTL) can be interpreted as implicit regularization, as it improves generalization ability [18]. In MTL, the output layer of multiple tasks can be composed of several task-specified output layers, as depicted in Fig. 1. As shown in Fig. 1-(a), each softmax layer is assigned to a specific task among gender, emotion, and language classifications, where the output layers are identical to the softmax layers. Each paralinguistic feature $\mathbf{x}_i$ has three labels, which are three one-hot vectors indicating language, gender, and emotion separately. The output $\mathbf{y}_i = \phi(\mathbf{x}_i)$ of the MTL on a paralinguistic feature vector $\mathbf{x}_i$ is divided into sub-vectors for each task $\{1, \cdots, c, \cdots, C\}$:

$$\mathbf{y}_i = \left[\mathbf{y}_i^{(1)}; \cdots; \mathbf{y}_i^{(c)}; \cdots; \mathbf{y}_i^{(C)}\right] \quad (1)$$

Because each paralinguistic feature $\mathbf{x}_i$ has multiple task-specified labels $\mathbf{z}_i^{(c)}$, the loss function is a weighted sum of the multiple task-specified losses, calculated as:

$$\mathcal{L}^{MTL} = \sum_i \sum_c \alpha_c(\mathbf{y}_i^{(c)}, \mathbf{z}_i^{(c)}) \quad (2)$$

where α_c is a weight for each task, usually determined by an empirical optimization method. In this work, we weight all losses equally, to allow for substantial comparison. As shown in Fig. 1-(b), the network has one softmax layer, where each node has two types of information: emotion and language. Each paralinguistic feature $\mathbf{x}_i$ has only one label, called a one-hot vector, which indicates both language and emotion. Thus, the loss function is calculated by one-hot categorical entropy from the softmax layer in training. For recognition, first posterior probabilities are calculated for each node of the softmax layer and the subsequent task-specified sum pooling layers merge the posterior probabilities for each task-specified classification.



(a) Three softmax layers (b) One softmax layer

Fig. 1. Two multi-task networks. (a) The output layer is composed of three task-specified softmax layers; (b) each node of one softmax layer contains two tasks, language and emotion.

3.2. Dropout

Dropout is a simple and effective way to prevent DNN from overfitting, where dropout randomly removes nodes of hidden layers with the predefined rate. Dropout is widely adopted in many classification tasks, such as image classification, ASR, and natural language processing. The problem of overfitting is much more serious in SER than in ASR due to the former's limited sample of emotional speech and the variation across speakers and annotators. Thus, generalization ability easily deteriorates in SER. In robust SER, dropout effectively eliminates the variability of speakers' subjectively expressed emotions.

3.3. Feature normalization

Feature normalization is a scaling method that is widely used to standardize a range of features. Variation in paralinguistic features of emotional speech is influenced not only by society and culture, but also by speakers' subjective emotional expressions. In addition, compared with spectral features, such as filterbank coefficients and mel-frequency cepstral coefficients (MFCC) in ASR, features used in SER such as F0 and energy vary more widely. Furthermore, functionals for SER are more varied than low-level descriptors (LLD) [2, 3, 4]. In [16], researchers reported better SER performance for the same language databases than that for cross-language databases. They concluded that cross-language and even cross-language-family acoustic emotion recognition is feasible, but they recommend relying on a suitable language resource for each desired target language. In [15], the authors improved SER performance with multilingual databases and data normalization. In [23], the authors applied speaker nor-

malization to reduce SER variance due to speaker variation, retaining only variance attributable to emotion variation. In this work, we use two kinds of normalization methods: *individual normalization*, where feature normalizations are conducted on each individual database, and *common normalization*, where features are commonly normalized by means and variances calculated from the two databases. For both methods, feature normalization makes the values of each feature in the databases have a mean of zero and variance of one.

4. EXPERIMENTS

4.1. Datasets

We use two speech emotion databases, the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database for English [12] and the Japanese Twitter-based Emotional Speech (JTES) database for Japanese [11]. The IEMOCAP database consists of approximately 12 hours and 33 minutes of speech from five females and five males. For our classification experiments, we only use 4,490 turns of four emotional labels. These utterances have a majority agreement amongst the annotators (at least two-thirds of annotators) for the emotion label. The IEMOCAP is a widely adopted benchmark test for SER [9, 10, 24, 25, 26]. The JTES database consists of about 23 hours and 31 minutes, where 50 spoken sentences of each emotion are acted out emotionally by 50 females and 50 males. In this work, the emotion annotation schemes of two databases in English and Japanese have four common emotion categories: {*anger, happiness/joy, neutral, sadness*}. Happiness in the IMEOCAP database is annotated as joy in the JTES database.

4.2. Experimental setup

We use the openSMILE toolkit 2.3.0 to extract a set of 1,582 features [4]. The feature set was introduced in the INTERSPEECH 2010 Paralinguistic Challenge [2]. We conduct our experiments using five-fold cross-validation, using one IEMOCAP session as a test set and the other four sessions as a training set. This method ensures that the models are trained and tested on speaker-independent sets. For a comparison experiment, the JTES database equally divides the 100 speakers into five sessions. As an evaluation measure, we use unweighted average recall (UAR) [3]. We run the experiments ten times with different initial parameters for twenty DNN architectures; the results are compared by averaging ten UAR values. The twenty DNN architectures are composed of varied hidden layers, 1024, 512, 256, and 128 nodes, and from one to five layers.

Table 1. Number of emotion utterances per category in the IEMOCAP and the JTES databases.

Database	IEMOCAP	JTES
Language	English	Japanese
Anger	1,103	5,000
Happiness	595	5,000
Neutral	1,708	5,000
Sadness	1,084	5,000
Total turns	4,490 (of 10,039)	20,000
Length	5h 36m (of 12h 33m)	23h 31m
# of speakers	10 (f:5, m:5)	100 (f:50, m:50)
Emotion	scripted/improvised	acted
Speech	fixed/free	fixed

5. RESULTS

First, we evaluate the monolingual and cross-lingual performance, wherein feature normalization and training are conducted on each individual database. We then evaluate the effectiveness of adopting dropout and common normalization. As shown in the second row of Table 2, the baseline performance of monolingual SER is 57.54% average UAR for the IEMOCAP and 80.17% for the JTES. Because the emotions are acted, the performance of the JTES is better than that of the IEMOCAP. For the cross-lingual experiments, the performances are substantially degraded without adaption or additional fine-tune training, as shown in the first row of Table 2. When we adopt dropout with a rate of 0.5, the performances of the IEMOCAP are improved in both normalization methods, but those of the JTES are degraded. For the IEMOCAP, the UARs increase from 57.54% to 58.79% with individual normalization and from 58.43% to 59.02% with common normalization, while the UARs decrease from 80.17% to 79.77% and from 81.21% to 81.16% for the JTES, respectively. The results show that dropout for SER is effective, but that additional costs are needed to optimize hyper-parameters such as dropout rate or epoch numbers. By adopting common normalization instead of individual normalization, in both the IEMOCAP and the JTES and with and without dropout, the UARs are consistently improved, from 57.54% to 58.43% and from 58.79% to 59.02% for the IEMOCAP, and for the JTES increasing from 80.17% to 81.21% and from 79.77% to 81.16%. Finally, we achieve 59.49% maximum UAR for the IEMOCAP (which outperforms values in previous research [24, 26]) and 81.44% maximum UAR for the JTES.

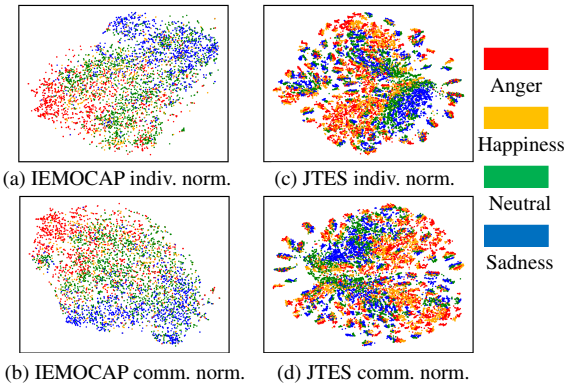
Table 2. Performance of cross-lingual and monolingual SER. The average UARs (%) with standard error are estimated over 10 runs and are the best among 20 DNN architectures. The maximum among 200 UARs is indicated in parentheses.

	IEMOCAP	JTES
Cross-lingual	42.68±0.11 (43.57)	39.84±0.08 (41.85)
Monolingual Individual normalization	57.54±0.14 (58.42)	80.17±0.04 (80.35)
+ dropout (rate=0.5)	58.79±0.12 (59.43)	79.77±0.06 (80.18)
Monolingual Common normalization	58.43±0.14 (59.11)	**81.21**±0.05 **(81.44)**
+ dropout (rate=0.5)	**59.02**±0.07 **(59.49)**	81.16±0.08 (81.61)

Table 3. Performance of multilingual SER. The average UARs (%) with standard error are estimated over 10 runs and are the best among 20 DNN architectures. The maximum among 200 UARs is indicated in parentheses.

	IEMOCAP	JTES
Two task-specified softmax layers (Emotion and language)	57.13±0.09 (57.82)	79.75±0.04 (79.96)
Two task-specified softmax layers (Emotion and gender)	57.03±0.16 (57.68)	79.50±0.05 (79.76)
Three task-specified softmax layers (Emotion, language and gender)	55.99±0.19 (57.05)	78.19±0.05 (78.50)
One softmax layer × 4 nodes (Four emotions on both language)	57.06±0.16 (57.79)	79.64±0.05 (79.95)
One softmax layer × 8 nodes (Four emotions per each language)	**58.02**±0.17 **(58.88)**	**80.47**±0.05 **(80.70)**
One softmax layer × 16 nodes (Four emotions per each language and gender)	57.42±0.10 (58.15)	78.78±0.06 (79.19)

Next, we evaluate the multilingual SER performance, based on adopting common normalization. Table 3's top three rows show results from the performance of the multi-task DNNs, which are composed of multiple task-specified softmax layers. The third row, with three task-specified softmax layers, is depicted in Fig. 1-(a); the first and second rows present the variants of Fig. 1-(a), where the softmax layers are specified to two tasks. Due to the equally assigned weights of the loss function in Eq. (2), the performances are degraded from the monolingual SER performance in the second row of Table 2. These degraded performances show that empirical optimization of the loss function is needed when the SHL is used for generalization. The bottom three rows in Table 3 present the performance of the multi-task DNNs with one softmax layer, where the network in the fifth row is depicted in Fig. 1-(b) and the fourth and sixth rows present the variants of Fig. 1-(b). Surprisingly, multilingual SER as



(a) IEMOCAP indiv. norm. (c) JTES indiv. norm.

(b) IEMOCAP comm. norm. (d) JTES comm. norm.

Fig. 2. Visualization of four emotions by t-SNE, where a paralinguistic feature, $\mathbf{x}_i$ is mapped by $f : \mathbb{R}^{1582} \rightarrow \mathbb{R}^2$.

Table 4. Recalls (%) for each emotion, improved by changing individual to common normalization.

	IEMOCAP		JTES	
Norm.	Indiv.	Comm.	Indiv.	Comm.
Anger	78.24	79.06	83.10	83.44
Happiness	*32.44*	*28.57*	73.80	75.32
Neutral	54.27	58.20	81.32	82.80
Sadness	68.73	72.14	83.16	84.18
UAR	58.42	**59.49**	80.35	**81.44**

presented in the fifth row of Table 3 performs better than the monolingual SER presented in the second row of Table 2, for both the IEMOCAP and the JTES. These improved performances of monolingual SER in the fourth and fifth rows of Table 2 and of multilingual SER in the fifth row of Table 3 show that common normalization enhances generalization of the network. Here, we find that emotional speech of various language databases can be used to improve generalization ability by common normalization.

Finally, the t-distributed stochastic neighbor embedding (t-SNE) [27] graphs of the normalized feature are shown in Fig. 2, where the two left-most distributions depict 4,490 turns of the IEMOCAP and the two right-most distributions depict 20,000 turns of the JTES. The two right-hand distributions of the JTES reveal that the feature space for male and female is divided into two and that the distributions have isolated speaker space, because the JTES is performed by 100 speakers. The two top-most distributions in Fig. 2 are individually normalized in each database, and the two bottom-most distributions are commonly normalized for all data in the two databases. In the JTES distributions, it is hard to find a difference between normalization methods. However, in the distributions of the IEMOCAP, for the ability to discriminate among three of the emotions - anger, neutral, and sadness - is improved. Happiness, indicated by the color orange, is widely superimposed on the other three emotions such that it is difficult to discriminate among them. The improved generalization ability is shown in Table 4, which presents detailed recall values for each emotion class. All recalls are improved by common normalization, while only one recall value (that is of happiness for the IEMOCAP) sees its performance degraded.

6. CONCLUSIONS

In this study, we investigated the generalization effect of multilingual SER on two highly heterogeneous languages, English and Japanese. We confirmed that the performance of monolingual SER was improved by common normalization performed in two databases. Furthermore, we achieved slightly better performance in the multilingual SER than in monolingual SER, using MTL (with one softmax layer and eight nodes) and common normalization. We found that, although emotions vary across societies, cultures, and languages, a multilingual SER system is feasible.

7. REFERENCES

[1] D. A. Sauter, F. Eisner, P. Ekman, and S. K. Scott, "Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations," in *Proc. of the National Academy of Sciences*, 2010, pp. 2408–2412.

[2] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Muller, and S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. of Interspeech*, 2010, pp. 2794–2797.

[3] B. Schuller and A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*, Wiley, 2013.

[4] F. Eyben, *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*, Springer Theses. Springer International Publishing, 2015.

[5] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *Proc. of ICASSP*, 2011, pp. 5688–5691.

[6] 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.

[7] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Proc. of Interspeech*, 2015, pp. 1537–1540.

[8] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation learning for speech emotion recognition," in *Proc. of Interspeech*,2016, pp. 3603–3607.

[9] 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.

[10] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," in *Proc. of Interspeech*, 2017, pp. 1263–1267.

[11] E. Takeishi, T. Nose, Y. Chiba, and A. Ito, "Construction and analysis of phonetically and prosodically balanced emotional speech database," in *Proc. of 2016 Conference of The O-COCOSDA*, 2016, pp. 16–21.

[12] 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," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[13] H. Sagha, P. Matejka, M. Gavryukova, F. Povolny, E. Marchi, and B. Schuller, "Enhancing multilingual recognition of emotion in speech by language identification," in *Proc. of Interspeech*,2016, pp. 2949–2953.

[14] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[15] B.-C. Chiou and C.-P. Chen, "Speech emotion recognition with cross-lingual databases," in *Proc. of Interspeech*, 2014, pp. 558–561.

[16] S.M. Feraru, D. Schuller, and B. Schuller, "Cross-language acoustic emotion recognition: An overview and some tendencies," in *Proc. of International Conference on Affective Computing and Intelligent Interaction*, 2015, pp. 125–131.

[17] M. Neumann and N. T. Vu, "Cross-lingual and multilingual speech emotion recognition on english and french," in *Proc. of ICASSP*, 2018, pp. 5769–5773.

[18] R. Caruana, "Multitask learning," *Journal of Machine Learning Research*, vol. 28, pp. 41–75, 1997.

[19] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2D continuous space," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 3–14, 2017.

[20] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Proc. of Interspeech*, 2017, pp. 1103–1107.

[21] Y. Zhang, Y. Liu, F. Weninger, and B. Schuller, "Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations," in *Proc. of ICASSP*, 2017, pp. 4990–4994.

[22] F. Tao and G. Liu, "Advanced LSTM: A study about better time dependency modeling in emotion recognition," in *Proc. of ICASSP*, 2018, pp. 2906–2910.

[23] P.-W. Hsiao and C.-P. Chen, "Effective attention mechanism in dynamic models for speech emotion recognition," in *Proc. of ICASSP*, 2018, pp. 2526–2530.

[24] S. Sahu, R. Gupta, G. Sivaraman, W. AbdAlmageed, and C. Espy-Wilson, "Adversarial auto-encoders for speech based emotion recognition," in *Proc. of Interspeech*, 2017, pp. 1243–1247.

[25] S. Sahu, R. Gupta, G. Sivaraman, and C. Espy-Wilson, "Smoothing model predictions using adversarial training procedures for speech based emotion recognition," in *Proc. of ICASSP*, 2018, pp. 4934–4938.

[26] 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.

[27] L.V.D. Maaten and G.E. Hinton, "Visualizing high-dimensional data using t-SNE," *Journal of Machine Learning Research*, pp. 2579–2605, Nov. 2008.