SIP (2021), vol. 10, e6, page 1 of 11 © The Author(s), 2021 published by Cambridge University Press in association with Asia Pacific Signal and Information Processing Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.
doi:10.1017/ATSIP.2021.7

ORIGINAL PAPER

Speech emotion recognition based on listener-dependent emotion perception models

ATSUSHI ANDO,[1,2] TAKESHI MORI,[1] SATOSHI KOBASHIKAWA[1] AND TOMOKI TODA[2]

This paper presents a novel speech emotion recognition scheme that leverages the individuality of emotion perception. Most conventional methods simply poll multiple listeners and directly model the majority decision as the perceived emotion. However, emotion perception varies with the listener, which forces the conventional methods with their single models to create complex mixtures of emotion perception criteria. In order to mitigate this problem, we propose a majority-voted emotion recognition framework that constructs listener-dependent (LD) emotion recognition models. The LD model can estimate not only listener-wise perceived emotion, but also majority decision by averaging the outputs of the multiple LD models. Three LD models, fine-tuning, auxiliary input, and sub-layer weighting, are introduced, all of which are inspired by successful domain-adaptation frameworks in various speech processing tasks. Experiments on two emotional speech datasets demonstrate that the proposed approach outperforms the conventional emotion recognition frameworks in not only majority-voted but also listener-wise perceived emotion recognition.

Keywords: Speech emotion recognition, perceived emotion, adaptation

Received 3 November 2020; Revised 22 March 2021

I. INTRODUCTION

Human speech is the most basic and widely used form of daily communication. Speech conveys not only linguistic information but also other factors such as speaker and emotion, all of which are essential for human interaction. Thus, speech emotion recognition (SER) is an important technology for natural human–computer interaction. There are a lot of SER applications such as voice-of-customer analysis in contact center calls [1, 2], driver state monitoring [3], and human-like responses in spoken dialog systems [4].

SER can be categorized into two tasks: dimensional and categorical emotion recognition. Dimensional emotion recognition is the task of estimating the values of several emotion attributes present in speech [5]. Three primitive emotion attributes, i.e. valence, arousal, and dominance are commonly used [6]. Categorical emotion recognition is the task of identifying the speaker's emotion from among a discrete set of emotion categories [7]. The ground truth is defined as the majority of perceived emotion class as determined by multiple listeners. Comparing these two tasks, categorical emotion recognition is more suitable for most applications because it is easy to interpret. This paper aims to improve categorical emotion recognition accuracy.

A large number of SER methods have been proposed. One of the basic approaches is based on utterance-level heuristic features including the statistics of frame-level acoustic features such as fundamental frequency, power, and Mel-frequency cepstral coefficients (MFCC) as determined by a simple classifier [8, 9]. Although they can recognize several typical emotions, their performance is still far from satisfactory because emotional cues exhibit great diversity, which demands the use of hand-crafted features with simple criteria. In contrast to this approach, several recent studies have achieved remarkable improvements through the use of deep neural network (DNN)-based classifiers [10–19]. The main advantage of DNN-based classifiers is that they can learn complex cues of emotions automatically by combining several kinds of layers. Recurrent neural network (RNN)-layers have been used to capture the contextual characteristics of utterances [12, 13]. Attention mechanism has also been employed to focus on the local characteristics of utterances [13, 14]. Furthermore, DNN-based models can utilize low-level features, e.g. log power-spectrogram or raw waveform, which have rich but excessively complex information that simple classifiers are unable to handle [7, 15].

However, SER is still a challenging task despite these advances. One of the difficulties lies in handling two types of individuality: speaker and listener dependencies. The way in which emotions are presented strongly depends on the speaker. It is reported that prosodic characteristics such as pitch and laryngealization differ among speakers [20]. This

[1] NTT Corporation, Yokosuka, Kanagawa 239-0847, Japan
[2] Nagoya University, Nagoya, Aichi 464-8601, Japan

Corresponding author:
A. Ando
Email: atsushi.ando.hd@hco.ntt.co.jp

is similar in emotion perceptions, and depends on age [21], gender [22], and cultures [23] of listeners. Given these issues, speaker dependency has often been considered for SER [24, 25]. However, the dependency of listeners has received little attention in SER tasks even though it influences the determination of the majority-voted emotions.

This paper presents a new SER framework based on listener-dependent (LD) models. The proposed framework aims to consider the individuality of emotional perceptions. In the training step of the proposed method, LD models are constructed so as to learn criteria for capturing the emotion recognition attributes of individual listeners. This allows the LD models to estimate the posterior probabilities of perceived emotions of specific listeners. Majority-voted emotions can be estimated by averaging these posterior probabilities as given by LD models. Inspired by domain adaptation frameworks in speech processing, three LD models are introduced: fine-tuning, auxiliary input, and sub-layer weighting. The fine-tuning method constructs as many LD models as listeners, while the remaining models cover all listeners by a single model. We also propose adaptation frameworks that allow the LD models to handle unseen listeners in the training data. Experiments on two emotional speech corpora show the individuality of listener perception and the effectiveness of the proposed approach. The main contributions of this paper are as follows:

(1) A scheme to recognize majority-voted emotions by leveraging the individuality of emotion perception is presented. To the best of our knowledge, this is the first work to take listener characteristics into consideration for SER.
(2) The performance of listener-oriented emotion perception is evaluated in addition to that of majority-voted emotion recognition. The proposed LD models show better performance than the conventional method in both metrics, which indicates that the proposed scheme is suitable for estimating not only majority-voted emotions, but also personalized emotion perception.

This paper is organized as follows. Section II introduces studies related to this work. Conventional emotion recognition is shown in Section III. The proposed framework based on LD models is shown in Section IV. Evaluation experiments are reported in Section V and the conclusion is given in Section VI.

II. RELATED WORK

A large number of emotion recognition methods have been investigated. The traditional approaches are based on utterance-level heuristic features. The statistics of frame-level acoustic features such as pitch, power, and MFCC are often used [8, 9]. However, it is difficult to create truly effective features because emotional cues exhibit great diversity. Thus, recent studies employ DNNs to learn emotion-related features automatically. The initial studies integrate frame-level emotion classifiers based on DNNs [10, 11]. Speaker emotion of individual frames are estimated by the classifiers, then integrated to evaluate utterance-level decisions by other DNN-based models such as extreme learning machines or RNNs. In recent years, end-to-end mechanisms that directly estimate utterance-level emotion from input feature sequences have been developed to allow the use of local and contextual characteristic [12, 13]. The end-to-end model consists of multiple DNN layers such as convolutional neural networks (CNNs), long short-term memory-RNNs, attention mechanisms, and fully-connected (FC) layers [7, 14, 15]. Recent works also utilize low-level inputs such as raw waveform and log power spectra in order to leverage the rich information of low-level features for estimation [16–19]. One of the advantages of the DNN-based framework is that the model can automatically learn complex emotional cues from training data. Our proposals employ this framework to construct LD models.

Several studies have investigated speaker dependency on emotional expression. They indicate that the speaker differences significantly affect emotion representation. For example, each speaker exhibits different laryngealization and pitch characteristics [20]. It has been suggested that speaker variability is a more serious factor than linguistic content [26]. Therefore, a lot of speaker adaptation methods for SER have been developed. Some attempt feature-level normalization; a speaker-dependent utterance feature is transformed into its speaker-independent equivalent [24]. Another approach is model-level adaptation. A speaker-independent emotion recognition model can, with a small amount of adaptation data, be adjusted to yield a speaker-dependent model [25]. Recent studies employ multi-task learning to construct gender-dependent models without inputting speaker attributes [18, 27]. Personal profiles have also been utilized to estimate speaker-dependent emotion recognition [28]. In this paper, we do not employ speaker adaptation in order to investigate the influence of just listener dependency; it will be possible, however, to combine the proposed LD model with existing speaker adaptation methods.

It has also been reported that emotion perception varies with the listener. Younger listeners tend to perceive emotions more precisely than their elders [21]. It is reported that female listeners are more sensitive to emotion than males [22]. The perception also depends on culture [23]. Even though listener variability affects the majority decision as to emotion perception, there is little work that considers the listener in SER. One related work is soft-label / multi-label emotion recognition; it models the distribution of emotion perception of listeners [17, 29, 30], but it cannot distinguish individuals. In music emotion recognition, several studies have tackled listener-wise perception [31, 32]. However, to the best of our knowledge, this is the first work to utilize listener variability for SER.

It is considered that constructing listener-oriented emotion recognition models is strongly related to the frameworks created for domain or speaker adaptation. As mentioned with regard to speaker adaptation in emotion recognition, adaptation has two approaches: feature-based and model-based adaptation. In recent speech processing methods such as those for speech recognition and speech synthesis, model-based adaptation is dominant because it is very powerful in handling complex changes in domains or listeners. One of the common adaptation approaches is updating the parameters of a pre-trained model by using a target domain dataset [33]. Another approach is developing a recognition model that includes the domain-dependent part. Technologies along these lines such as switching domain-dependent layers [34], projections with auxiliary input [35, 36], or summation of multiple projection outputs with speaker-dependent weights [37] have been proposed. Inspired by these successful frameworks, our proposal yields LD emotion recognition models.

III. EMOTION RECOGNITION BY MAJORITY-VOTED MODEL

This section describes the conventional emotion recognition approach based on DNN model [14, 18]. In this paper, we call this model the *majority-voted model* because it directly models majority-voted emotion of multiple listener perceptions.

Let $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_T]$ be the acoustic features of an input utterance and T be their total length. $C = \{1, \cdots, K\}$ is the set of target emotion indices, e.g. 1 means neutral and 2 is happy. K is the total number of target emotions. The task of SER is formulated as estimating the majority-voted emotion of utterance $c \in C$ from $\boldsymbol{X}$,

$$\hat{c} = \arg\max_{c} P(c|\boldsymbol{X}), \tag{1}$$

where $\hat{c}$ is the estimated majority-voted emotion. $P(c|\boldsymbol{X})$ is the posterior probability indicated by the input utterance. The ground truth of majority-voted emotion c is defined as the dominant choice of multiple listener's perception results,

$$c \equiv \arg\max_{k} \sum_{l \in L} f(c^{(l)} = k), \tag{2}$$

where $c^{(l)} \in C$ is the perceived emotion of human listener l[1]. $f(\cdot)$ is a binary function of emotion presence / absence, $f(c^{(l)} = k) = 1$ if l perceived the k-th target emotion from the utterance, otherwise 0. L is a set of the listeners annotated emotion perceptions given the input utterance, where $L \subset \mathbb{L}$ and $\mathbb{L}$ is a set of listeners in the training data. Note that the set of the listeners, L, can vary for each utterance in SER task.

The posterior probabilities of the majority-voted emotions $\boldsymbol{y} = [P(c = 1|\boldsymbol{X}), \cdots, P(c = K|\boldsymbol{X})]^\top$ are evaluated by the estimation model composed of an encoder and decoder. An example of the estimation model is shown in Fig. 1. The encoder projects an arbitrary length of acoustic features $\boldsymbol{X}$ into a fixed-length hidden representation in order to extract context-sensitive emotional cues. It consists of CNN, Bidirectional LSTM-RNNs (BLSTM), and self-attention layers such as a structured self-attention network [38]. The decoder estimates $\boldsymbol{y}$ from the hidden representation. It is composed of several FC layers.

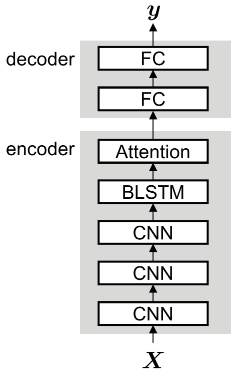


Fig. 1. An example of the conventional emotion recogntion model based on direct modeling of majority-voted emotion.

Fig. 1 - B/W online

The parameters of the estimation model are optimized by stochastic gradient descent with cross entropy loss,

$$\mathcal{L} = -\sum_{c} q(c) \log P(c|\boldsymbol{X}), \tag{3}$$

where $q(\cdot)$ is the reference distribution. $q(c = k)$ is 1 if the majority-voted emotion is the k-th target emotion, otherwise 0.

IV. EMOTION RECOGNITION BY LD MODELS

This section proposes a majority-voted emotion recognition framework based on LD models. The key idea of our proposal is to consider the individuality of emotion perception. Every majority-voted emotion is determined from different sets of listeners in the SER task. However, the characteristics of emotional perceptions vary with the listener. Direct modeling of the majority-voted emotion will result in conflating multiple different emotion perception criteria, which may degrade estimation performance. To solve this problem, the proposed method constructs LD models to learn listener-specific emotion perception criteria.

This framework determines the posterior probability of majority-voted emotion by averaging the posterior probabilities of the LD perceived emotions,

$$P(c|\boldsymbol{X}) = \frac{1}{N_L} \sum_{l \in L} P(c^{(l)}|\boldsymbol{X}, l), \tag{4}$$

where N_L is the total number of listeners L. In vector representation,

$$\boldsymbol{y} = \frac{1}{N_L} \sum_{l \in L} \boldsymbol{y}^{(l)}, \tag{5}$$

where $\boldsymbol{y}^{(l)} = [P(c^{(l)} = 1|\boldsymbol{X}, l), \cdots, P(c^{(l)} = K|\boldsymbol{X}, l)]^\top$ is the LD posterior probability vector evaluated by the LD model.

[1] In this paper, notation $^{(l)}$ means a LD variable.

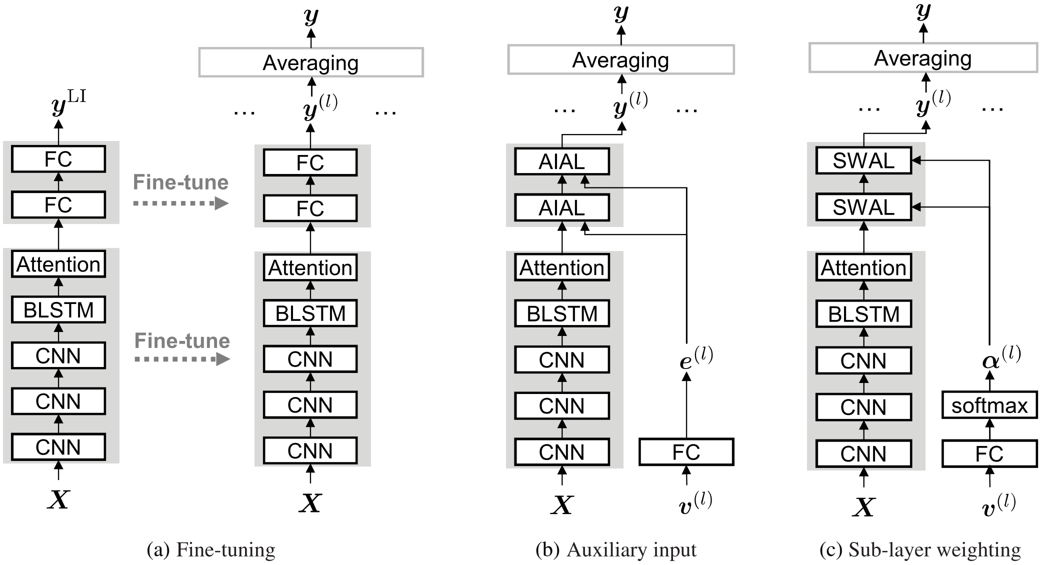


Fig. 2. Overview of the proposed majority-voted emotion recognition based on listener-dependent (LD) models. (a) Fine-tuning. (b) Auxiliary input. (c) Sub-layer weighting.

In this paper, three LD models are introduced: fine-tuning, auxiliary input, and sub-layer weighting. All of them are inspired by adaptation techniques in speech processing. The proposed frameworks based on LD models are overviewed in Fig. 2.

A) Model overview

1) Fine-tuning based LD model

A listener-independent (LI) model is retrained with specific listener training data to create a LD model. This is inspired by fine-tuning based domain adaptation in speech recognition [33].

Two-step training is employed. First, the LI model is trained with all utterances and their listeners in the training data. Listener-wise annotations are used for the reference distributions without distinguishing among listeners. The trained LI model outputs LI posterior probabilities y^{LI}. Second, the LI model is retrained with a particular listener's labels and utterances. This yields as many isolated LD models as there are listeners in the training data.

The optimization methods of LI / LD models are cross-entropy loss with LD perceived emotion, see as equation (3),

$$\mathcal{L} = -\sum_{c^{(l)}} q(c^{(l)}) \log P(c^{(l)}|\boldsymbol{X}, l). \tag{6}$$

2) Auxiliary input based LD model

The second model adapts particular layers of the estimation model through the auxiliary use of listener information. This is inspired by speaker adaptation in speech recognition [35] and speech synthesis [36].

One-hot vector of listener l, $\boldsymbol{v}^{(l)}$, is used to enhance acoustic features. $\boldsymbol{v}^{(l)}$ is projected into listener embedding vector $\boldsymbol{e}^{(l)}$ by an embedding layer,

$$\boldsymbol{e}^{(l)} = \sigma(\boldsymbol{W}_e \boldsymbol{v}^{(l)} + \boldsymbol{b}_e), \tag{7}$$

where $\boldsymbol{W}_e, \boldsymbol{b}_e$ are the parameters of the embedding layer. σ is an activation function such as hyperbolic tangent. Then $\boldsymbol{e}^{(l)}$ is used as the auxiliary input of the adaptation layers, named auxiliary input-based adaptation layers (AIALs), in the decoder so as to adjust the decoder to the chosen listener,

$$\boldsymbol{h}_{a,o} = \boldsymbol{W}_a \left[\boldsymbol{h}_{a,i}^{\top}, \boldsymbol{e}^{(l)\top}\right]^{\top} + \boldsymbol{b}_a, \tag{8}$$

where $\boldsymbol{h}_{a,i}, \boldsymbol{h}_{a,o}$ are the input and the output of the AIAL, respectively, and $\boldsymbol{W}_a, \boldsymbol{b}_a$ are the parameters. Note that the embedding layer, the encoder and decoder are optimized jointly.

The advantage of the auxiliary input based approach is that it offers greater stability than fine-tuning based models. There are two reasons for this. First, it has fewer parameters than fine-tuning based models. The fine-tuning models have to store as many encoders and decoders as there are listeners. However, auxiliary input based models share the encoder and decoder among all listeners, which suppresses the number of parameters. Second, the auxiliary input model can utilize the similarity of listeners. The fine-tuning models learn for just particular listeners. On the other hand, similar listeners will be mapped into similar latent vectors by the projection function, which reinforces the encoder's ability to learn LD emotion perception.

Note that only the decoder of the LD model is adapted to the selected listener. We consider that every listener perceives the same emotional cues from acoustic features, e.g. pitch raise / fall and fast-talking, and decision making from the emotional cues depends on listeners.

3) Sub-layer weighting based LD model

The sub-layer weighting approach combines multiple projection functions to adapt to the listener. This is inspired by context adaptive DNN proposed for source separation [37].

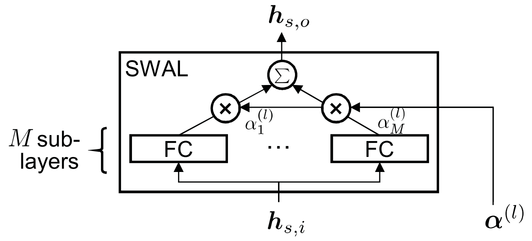


Fig. 4 - B/W online

Fig. 3. Structure of the Sub-layer Weighting-based Adaptation Layer (SWAL).

Sub-layer weighting-based adaptation layers (SWALs) are used to adapt the decoder to the selected listener. SWAL consists of multiple FC sub-layers,

$$\boldsymbol{h}_{s,o} = \sum_{m=1}^{M} \alpha_m^{(l)} \left(\boldsymbol{W}_{s,m}\boldsymbol{h}_{s,i} + \boldsymbol{b}_{s,m}\right), \qquad (9)$$

where $\boldsymbol{h}_{s,i}, \boldsymbol{h}_{s,o}$ are the input and output of the SWAL. $\boldsymbol{W}_{s,m}, \boldsymbol{b}_{s,m}$ is the parameters of the m-th sub-layer and M is the total number of sub-layers. $\alpha_m^{(l)}$ is the adaptation weight associated with the selected listener,

$$\boldsymbol{\alpha}^{(l)} = \mathsf{SOFTMAX}(\boldsymbol{W}_e\boldsymbol{v}^{(l)} + \boldsymbol{b}_e), \qquad (10)$$

where $\boldsymbol{\alpha}^{(l)} = [\alpha_1^{(l)}, \cdots, \alpha_M^{(l)}]^\top$ is an adaptation weight vector determined by listener representation $\boldsymbol{v}^{(l)}$ and $\mathsf{SOFTMAX}(\cdot)$ is the softmax function. The structure of the SWAL is shown in Fig. 3. The model parameters including the embedding layer, equation (10), are jointly optimized as the auxiliary input approach.

The main advantage of sub-layer weighting is that it is more expressive than auxiliary input based models. LD estimation is conducted by means of combining the perception rule of embedded listeners. However, it will require more training data than the auxiliary input-based approach because it has more parameters.

B) Adaptation to a new listener

The LD models can be directly applied to listener-closed situations, i.e. evaluation listeners are present in the training data. Although common SER tasks are listener-closed, SER in practice is listener-opened so evaluation listeners are not included in the training set. Our solution is to propose adaptation methods that allow the LD models to handle open listeners using a small amount of adaptation data.

The adaptation for the fine-tuning based LD model can be achieved by the retraining method that is the same as the second step of the flat-start training of the model. The LI model constructed by the training set is fine-tuned using the adaptation utterances of the particular listener.

The auxiliary input and sub-layer weighting based LD models adapt to a new listener to estimate the most similar listener code in the training listeners. Let $\hat{\boldsymbol{v}}^{(l)}$ and $\boldsymbol{u}^{(l)}$ be the estimated listener code and its indicator whose sizes are the same as $\boldsymbol{v}^{(l)}$. The initial value of $\boldsymbol{u}^{(l)}$ is a zero vector. $\boldsymbol{u}^{(l)}$ is updated by backpropagating the loss of the adaptation data while freezing all the model parameters, as shown in Fig. 4. Note that the proposed adaptation does not update $\hat{\boldsymbol{v}}^{(l)}$ directly so as to restrict that the sum of $\hat{\boldsymbol{v}}^{(l)}$ to be 1 and all the dimensions to be non-negative, which is the same constraint as $\boldsymbol{v}^{(l)}$ in the training step. After the estimated listener vector $\hat{\boldsymbol{v}}^{(l)}$ is obtained from the adaptation data, it is fed to the LD models as the listener code, and the posterior probabilities of the perceived emotion of the new listener are derived. This approach is similar to those proposed in speech recognition [39].

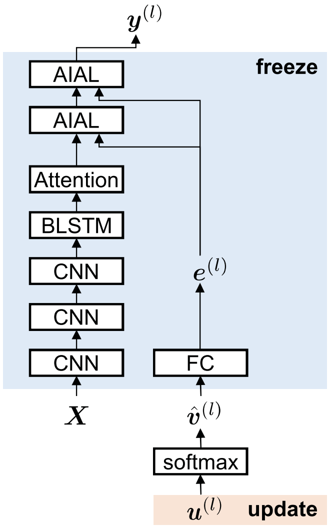


Fig. 4. Adaptation for the auxiliary input-based LD model.

V. EXPERIMENTS

We evaluated the proposed LD models in two scenarios. The first was a flat-start evaluation. The estimation models were trained from scratch and evaluated by listeners present in the training dataset, i.e. a listener-closed condition. The second was an adaptation evaluation. It was a listener-open condition; the utterances and listeners separated from the training data were used for the adaptation and evaluation data to investigate estimation performance for unseen listeners.

A) Datasets

Two large SER datasets, Interactive Emotional Dyadic Motion Capture (IEMOCAP) [40] and MSP-Podcast [41], were used in evaluating the proposal. IEMOCAP and MSP-Podcast contain acted and natural emotional speech, respectively. We selected four target emotions, neutral (*Neu*), happy (*Hap*), sad (*Sad*), and angry (*Ang*). All non-target emotion classes in the datasets were set as other (*Oth*) class.

IEMOCAP contains audiovisual data of 10 skilled actors (five males and five females) in five dyadic sessions. The database consists of a total of 12 h of English utterances generated by improvised or scripted scenarios specifically

Table 1. Number of utterances in IEMOCAP

		Neu	*Hap*	*Sad*	*Ang*	*Oth*	Total
Majority		1099	947	608	289	0	2943
Listener	1	412	1166	589	284	456	2907
	2	951	876	586	269	99	2781
	3	1225	717	324	155	150	2571
	Rest	226	119	113	56	56	570

Table 2. Number of utterances in MSP-Podcast

		Neu	*Hap*	*Sad*	*Ang*	*Oth*	Total
Majority		22,681	12,302	2351	2893	0	40,227
Listener	1	5475	380	27	59	45	5986
	2	1130	1026	120	69	800	3145
	3	421	1072	191	128	440	2252
	...						...
	154	78	37	4	2	27	148
	Rest	74,524	57,200	12,459	14891	44,470	203,544

written to represent the emotional expressions. As in several conventional studies [11, 14, 16, 27, 29], we used only audio tracks of the improvised set since scripted data may contain undesired contextual information. There are six listeners in the corpus and every utterance was annotated by three of them. The annotated categorical emotion labels are 10: neutral, happy, sad, angry, disgusted, excited, fearful, frustrated, surprised, and other. We combine happy and excited into *Hap* class in accordance with conventional studies [18, 19]. Although listeners were allowed to give multiple emotion labels to each utterance, to evaluate listener-wise emotion perception performance we unified them so that all listeners labeled one emotion per utterance. The unification rule was to select the majority-voted emotion if it is included in the multiple annotations, otherwise the first annotation is the unique perceived emotion. The listeners who gave fewer than 500 annotations were clustered as the "rest listeners" because they provided too little information to support learning LD emotion perception characteristics. Finally, the utterances whose majority-voted emotion is one of the target emotions were used to form the evaluation dataset. The numbers of utterances are shown in Table 1. The estimation performances were compared by leave-one-speaker-out cross-validation; one speaker was used for testing, another for validation, and the other eight speakers were used for training.

MSP-Podcast contains English speech segments from podcast recordings. Collected from online audio shows, they cover a wide range of topics like entertainment, politics, sports, and so on. We used Release 1.7 which contains approximately 100 h of speaking turns. Annotations were conducted by crowdsourcing. There are 11,010 listeners and each utterance was annotated by at least three listeners (6.7 listeners per utterance on average). This dataset has two types of emotion annotations, primary and secondary emotions; we used only the primary emotions as listener-wise perceived emotions. The variety of annotated primary emotions consisted of neutral, happy, sad, angry, disgust, contempt, fear, surprise, and other. We used the utterances whose majority-voted emotions were one of the target emotions. A predetermined speaker-open subset was used in the flat-start evaluation; 8215 segments from 60 speakers for testing, 4418 segments from 44 speakers for validation, and the remaining 25,332 segments from more than 1000 speakers for training. The listeners who gave fewer than 100 annotations in the training set were clustered as "rest listeners", same as IEMOCAP. The total numbers of emotional utterances are shown in Table 2.

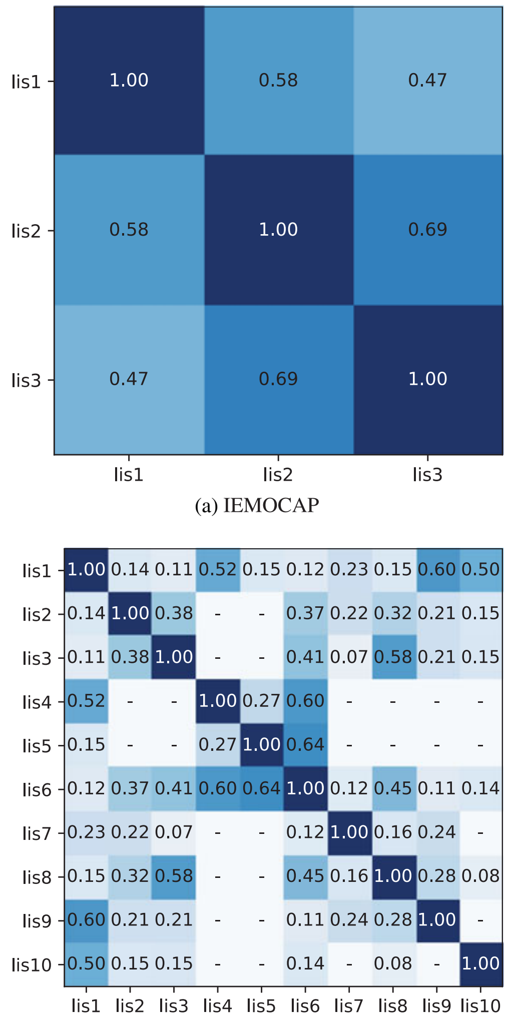
To clarify the impact of listener dependency on emotion perception, we first investigated the similarity of listener annotations. Fleiss' and Cohen's kappa coefficients were employed as the similarity measures of the overall and the individual pairs of listeners, respectively. The coefficients were calculated through 5-class matching (4 target emotions + *Oth*) from only the utterances in the evaluation dataset. Cohen's kappa coefficients of the listener pairs in which both listeners annotated less than the same 20 utterances were not evaluated ('—' in results). Fleiss' kappa were 0.57 in IEMOCAP and 0.35 in MSP-Podcast. There are two reasons for the lower consistency rate of MSP-Podcast. First, MSP-Podcast speech segments are completely natural, unlike IEMOCAP utterances which contained acted speech; this increased the ambiguous emotional speech in MSP-Podcast. Second, MSP-Podcast listeners will have larger diversity than those of IEMOCAP. All the listeners in IEMOCAP are students in the same university [40]. Cohen's kappa coefficients of IEMOCAP and MSP-Podcast listeners are shown in Fig. 5. It is shown that listener 2 showed relatively high similarity with listeners 1 and 3, while listeners 1 and 3 showed low similarity in IEMOCAP. The MSP-Podcast result showed the same property. Listener 1 showed high similarity with listeners 4, 9, 10, but low similarity with the remaining listeners. Listener 6 was similar to listeners 4 and 5. These indicate that emotion perception depends on listeners, and that there are several clusters of emotion perception criteria.

B) Flat-start evaluation

1) Setups

Log power spectrograms were used as acoustic features. The conditions used in extracting spectrograms followed those of conventional studies [16, 42]. Frame length and frame shift length were 40 and 10 ms, respectively. The window type was Hamming window. DFT length was 1600 (10 Hz grid resolution) and we used 0–4 kHz frequency range, which yielded 400-dimensional log power spectrograms. All the spectrograms were z-normalized using the mean and variance of the training dataset.

The baseline was the majority-voted emotion recognition model described in Section III. An ensemble of multiple majority-voted models with different initial parameters was also employed to compare with the proposed method that unifies several outputs of LD models. The number of



(a) IEMOCAP

(b) MSP-Podcast

Fig. 5. Cohen's kappa coefficients of listener annotations. (a) IEMOCAP. (b) MSP-Podcast.

ensembles was the average number of listeners per utterance, i.e. 3 and 7 in IEMOCAP and MSP-Podcast, respectively. The structure of the baseline is shown in Table 3. Each CNN layer was followed by batch normalization [43], rectified linear activation function, and 2×2 max pooling layers. Early stopping was performed using development set loss as the trigger. The optimization method was Adam [44] with a learning ratio of 0.0001. In the training step, inverse values of the class frequencies were used as class weights to mitigate the class imbalance problem [45]. Minibatch size was 8 in IEMOCAP and 16 in MSP-Podcast evaluations. Data augmentation was performed by means of speed perturbation with speed factors of 0.9, 0.95, 1.05, and 1.1 [7]. SpecAugment [46] was also applied with two time and frequency masking. The ensemble of multiple majority-voted models with different initial parameters was also compared because the proposed method unifies the multiple outputs of LD models. The number of ensembled models was the average number of listeners per utterance, i.e. 3 and 7 in IEMOCAP and MSP-Podcast, respectively.

Table 3. Network architectures of emotion recognition model

	Layer-type	Parameters
Encoder	CNN	16 ch, [12×16] kernel, [2×2] stride
	CNN	24 ch, [4×6] kernel, [1×1] stride
	CNN	32 ch, [3×4] kernel, [1×1] stride
	BLSTM	1 layer, 128 dim.
	Attention	structured self-attention [38], 4 head
Decoder	FC / AIAL / SWAL	1 layer, 64 dim.
	FC / AIAL / SWAL	1 layer, 4 dim.

Table 4. Number of model parameters

			IEMOCAP	MSP-Podcast
Baseline	Majority		1.04M	
	Majority (ens.)		3.11M	7.26M
Proposed	Fine-tuning	LI	1.04M	
		LD	4.15M	160.68M
	Auxiliary		1.04M	
	Weighting		1.09M	

The proposals were LD models by fine-tuning, auxiliary input, and sub-layer weighting. LI model, the base model of the fine-tuning based LD model, was also compared to investigate the difference before and after fine-tuning. These model structures were the same as the baseline except for FC layers in the decoder, which were replaced with AIALs or SWALs. The numbers of listener embedding vector dimensions and sub-layers were 2, 3, 4, 8, 16 and we selected the best parameters for each dataset. The learning ratio was 0.0001 and 0.00005 in flat-start and fine-tuning, respectively. The class weights were calculated by each listener in LD model training. The other training conditions and data augmentation setup were those of the baseline. All the baseline and the proposed methods were implemented by PyTorch [47]. Comparisons of the model parameters are shown in Table 4. The numbers of dimensions of listener embedding vector dimensions and sub-layers shown in the table were 4.

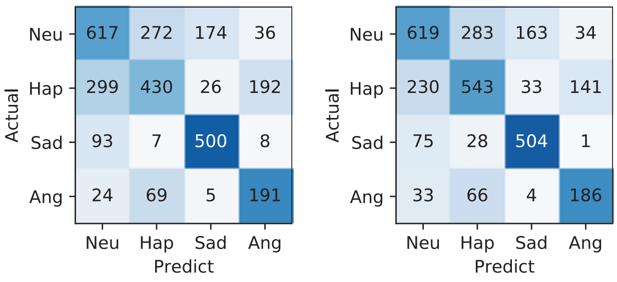
Two evaluation metrics common in emotion recognition studies were employed: weighted accuracy (WA) and unweighted accuracy (UA). WA is the classification accuracy of all utterances and UA is the macro average of individual emotion class accuracies. We evaluated not only the performances of majority-voted emotion estimation but also those of listener-wise emotion recognition to investigate the capability of the proposed LD models.

2) Results

The results of majority-voted emotion estimation are shown in Table 5. The notation Majority (ens.) means the ensemble result of the majority-voted models. Comparing the two datasets, MSP-Podcast yielded lower overall accuracy than IEMOCAP. It is considered that MSP-Podcast contains natural speech with a large number of speakers, which makes it more difficult to recognize emotion than IEMOCAP, which holds acted utterances from limited speakers. The LD models showed significantly better WAs ($p < 0.05$ in paired t-test) as almost the same or better UAs than the

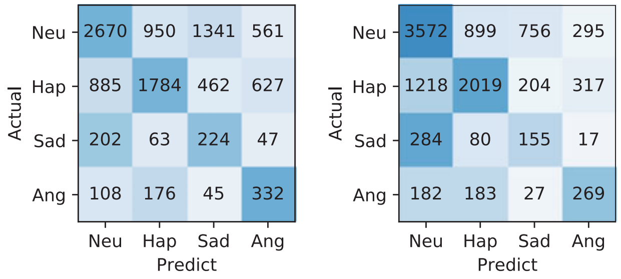
Table 5. Estimation accuracies of the majority-voted emotions. Bold means the highest accuracy.

			IEMOCAP		MSP-Podcast	
			WA	UA	WA	UA
Baseline	Majority		59.1	62.5	47.8	47.0
	Majority (ens.)		61.0	64.7	47.8	47.0
Proposed	Fine-tuning	LI	61.2	63.7	54.9	**48.9**
		LD	**62.9**	**65.2**	56.6	**48.9**
	Auxiliary		**62.9**	**65.2**	57.4	47.0
	Weighting		62.3	64.0	**58.7**	46.3



(a) Majority-voted model (b) LD model

Fig. 6. Confusion matrices for IEMOCAP. (a) Majority-voted model. (b) LD model



(a) Majority-voted model (b) LD model

Fig. 7. Confusion matrices for MSP-Podcast. (a) Majority-voted model. (b) LD model.

Table 6. Macro-average of estimation accuracies of the listener-dependent perceived emotions. Bold means the highest accuracy.

			IEMOCAP		MSP-Podcast	
			WA	UA	WA	UA
Baseline	Majority		57.0	62.0	44.9	43.2
	Majority (ens.)		58.9	64.1	44.0	43.3
Proposed	Fine-tuning	LI	59.7	63.6	50.0	44.3
		LD	**61.6**	63.3	51.8	44.8
	Auxiliary		61.5	**64.7**	52.1	**45.1**
	Weighting		60.9	63.8	**53.1**	44.9

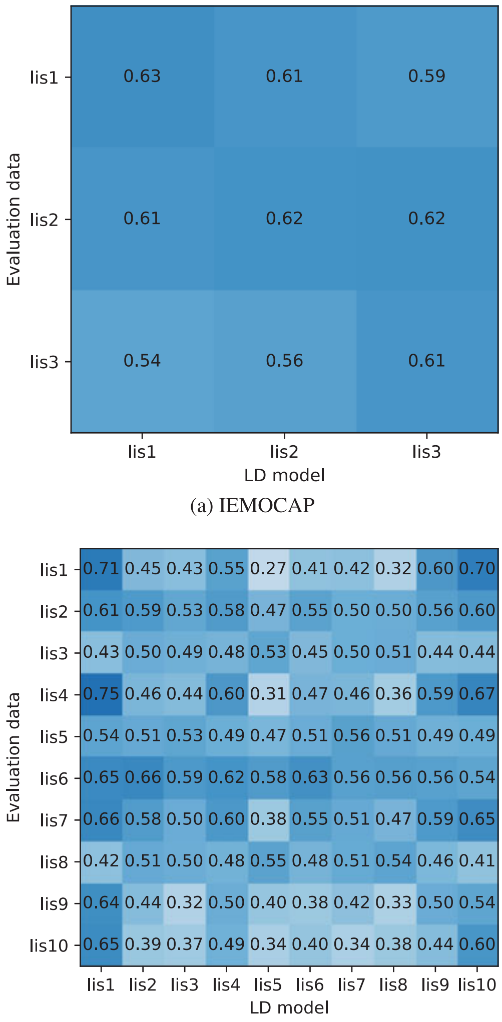
baselines on both datasets. For example, fine-tuning based LD models achieved 3.8 and 2.7 % improvements from the single majority-voted model in WA and UA for IEMOCAP, 8.8 and 1.9% for MSP-Podcast. These results indicate that majority-voted emotion recognition based on LD models is more effective than the conventional majority-voted emotion modeling framework. Figs. 6 and 7 show the confusion matrices of the baseline and the auxiliary input-based LD model. Comparing numbers of the corrected samples for each emotion, *Hap* was improved on both IEMOCAP and MSP-Podcast, while *Sad* and *Ang* were degraded on MSP-Podcast. One possible reason for the degradation is data imbalance. These two emotions were hardly observed by some listeners, e.g. listener 154 annotated only two utterances with *Ang* emotion as shown in Table 2, which leads to overfitting in the LD model. Comparing the LD models, there were no significant differences ($p \geq 0.05$), while fine-tuning and auxiliary input were slightly better for IEMOCAP, while sub-layer weighting yielded the best WA and fine-tuning attained the best UA for MSP-Podcast. Taking the number of parameters (see Table 4) into consideration, the auxiliary input based model is suitable for all conditions, while sub-layer weighting may become better for large datasets. Note that even the LI model significantly outperformed the model ensemble baseline in MSP-Podcast ($p < 0.05$). One possible reason is that training by listener-specific labels allows the model to learn inter-emotion similarities. For example, a set of listener-wise labels *neu*, *neu*, *hap* indicates that the speech may contain both *neu* and *hap* cues. On the other hand, its majority-voted label just indicates the speech has *neu* characteristics.

Macro-averages of listener-wise emotion recognition performances are shown in Table 6. In this evaluation, WA / UAs of all the listeners except for "rest listeners" were averaged to compare overall performance. Table 6 represents that all LD models showed better performance than the baseline. The improvements were significant in MSP-Podcast ($p < 0.05$ in paired t-test), while not in IEMOCAP. It is considered that IEMOCAP has only three listeners, which is too few samples for a paired t-test. Note that there are no significances among the three proposed LD models. The matrices of listener-wise WA with LD models are also shown in Fig. 8. All the LD models were constructed by fine-tuning. Comparing the matrices to Fig. 5, the evaluations of high similarity listener pairs tend to show relatively high WAs. For example, LD models of listeners 1, 4, 9, 10 showed higher WAs than the remaining LD models for listener 1 evaluation data. These results indicate that LD models can accurately learn LD emotion perception characteristics. Note that there are several listeners in which the listener-mismatched LD model showed better WAs than the listener-matched model. One possible reason is the difference in the amount of training data in listeners. For example, listener 1 has several times of training data compared with other listeners, which yields a better emotion perception model in spite of listener-mismatched conditions.

C) Adaptation evaluation

1) Setups

We resegmented MSP-Podcast evaluation subsets to create an utterance and listener open dataset. First, the utterances contained only "rest listeners" were selected from the original training, validation, and testing dataset as the open data candidates. Second, the listeners who annotated more than



(a) IEMOCAP

(b) MSP-Podcast

Fig. 8. WAs of listener-wise emotion recognitions with LD models. (a) IEMOCAP. (b) MSP-Podcast.

two utterances with each target emotion and 30 utterances in total of the candidates were selected as the open listeners. Finally, the utterances that had one or more open listeners in the candidates were regarded as the open dataset, while the remaining candidates were returned to the original training, validation, and test sets. We selected 24 listeners with 1080 utterances for the open dataset. The average number of utterances per listener was 42.8. Note that we did not use IEMOCAP in the adaptation evaluation because no open utterances were available.

The baseline method was the majority-voted emotion recognition model without adaptation. It was trained by the resegmented training and validation set. The proposed was the auxiliary input-based LD model with adaptation. The LD model was trained by the resegmented training and validation set first, then adapted to the specific listener in the open set with adaptation data. five-fold cross-validation was used in the LD model adaptation; 80 % of the open dataset was used for adaptation and the rest 20 % was for the evaluation. To evaluate the performance of the listener code estimation alone, we also ran a comparison with the auxiliary input-based LD model in the oracle condition in which the one-hot listener code that showed the highest geometric mean of WA and UA was selected for each open listeners. We used the same LD model in adaptation and oracle conditions. For the adaptation, the minibatch size was the same as the amount of listener-wise utterances in the adaptation set. The learning rate was 0.05. Earlystopping was not used and the adaptation was stopped at 30 epochs.

Table 7. Macro-average of WAs and UAs in listener-open dataset

			MSP-Podcast	
			WA	UA
Baseline	Majority		41.4	42.0
Proposed	Auxiliary	Adapted	48.4	44.2
		Oracle	58.6	52.7

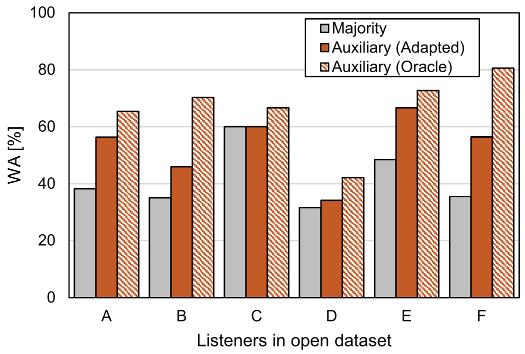


Fig. 9. WA for each listeners in open dataset.

Evaluation metrics were macro-averages of the listener-wise WAs and UAs. Note that we did not evaluate the performance of the majority-voted emotion recognition because the majority-voted emotions were not *open*; the listeners of the utterances in the open dataset were almost “rest listeners” who included in the training subset and the majority-voted emotions were mostly determined by them.

2) Results

The macro-average of listener-wise WAs and UAs are shown in Table 7. Relative to the baseline, the auxiliary input-based LD model with adaptation achieved significantly better WA ($p < 0.05$ in paired t-test) with the same level of UA ($p > 0.05$). Furthermore, the oracle of the auxiliary model showed very high WA and UA ($p < 0.05$ compared with the auxiliary model with adaptation). These indicate that the auxiliary model is capable of LD emotion recognition and the proposed adaptation is effective for unseen listeners, while there is room for improvement to estimate better listener code from a limited adaptation set. The same trend is present in the examples of the listener-wise WAs and UAs shown in Fig. 9. The LD model with adaptation showed the same or better performances than the majority model for all listeners, and the auxiliary model in the oracle setup greatly outperformed the adapted model for some listeners such as listener B.

Tables 6 and 7 show that the auxiliary model with oracle evaluation in the open set attained higher accuracies than those with listener-closed training in the test set. One possibility is that there are some listeners who gave noisy annotations, which degrades estimation performance even in listener-closed conditions. It has been reported that there are several noisy annotators in crowdsourced data like MSP-Podcast [48].

VI. CONCLUSION

This paper proposed an emotion recognition framework based on LD emotion perception models. The conventional approach ignores the individuality of emotional perception. The key idea of the proposal lies in constructing LD models that account for individuality. Three LD models were introduced: fine-tuning, auxiliary input, and sublayer weighting. The last two models can adapt to a wide range of listeners with limited model parameters. Experiments on two large emotion speech corpora revealed that emotion perception depends on listeners and that the proposed framework outperformed the conventional method by means of leveraging listener dependencies in majority-voted emotion recognition. Furthermore, the proposed LD models attained higher accuracies in listener-wise emotion recognition, which indicates that the LD models were successful in learning the individuality of emotion perception.

Future work includes investigating the effectiveness of the proposed approach in other languages and cultures, improving the adaptation framework to unseen listeners, and combining the LD models with the speaker adaptation frameworks.

REFERENCES

[1] Devillers, L.; Vaudable, C.; Chastagnol, C.: Real-life emotion-related states detection in call centers: a cross-corpora study, in *Proc. of INTERSPEECH*, 2010, 2350–2353.

[2] Ando, A.; Masumura, R.; Kamiyama, H.; Kobashikawa, S.; Aono, Y.; Toda, T.: Customer satisfaction estimation in contact center calls based on a hierarchical multi-task model. *IEEE/ACM Trans. Audio Speech Lang. Process.*, **28** (2020), 715–728.

[3] Tawari, A.; Trivedi, M.: Speech based emotion classification framework for driver assistance system, in *Proc. of IEEE IVS*, 2010, 174–178.

[4] Acosta, J.C.: Using emotion to gain rapport in a spoken dialog system, in *Proc. of NAACL HLT Student Research Workshop and Doctoral Consorium*, 2009, 49–54.

[5] Kowtha, V. *et al.*: Detecting emotion primitives from speech and their use in discerning categorical emotions, in *Proc. of ICASSP*, 2020, 7164–7168.

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

[7] Sarma, M.; Ghahremani, P.; Povey, D.; Goel, N.K.; Sarma, K.K.; Dehak, N.: Emotion identification from raw speech signals using DNNs, in *Proc. of Interspeech*, 2018, 3097–3101.

[8] Schuller, B.; Batliner, A.; Steidl, S.; Seppi, D.: Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. *Speech Commun.*, **53** (2011), 1062–1087.

[9] Luengo, I.; Navas, E.; Hernàez, I.; Sànchez, J.: Automatic emotion recognition using prosodic parameters, in *Proc. of Interspeech*, 2005, 493–496.

[10] Han, K.; Yu, D.; Tashev, I.: Speech emotion recognition using deep neural network and extreme learning machine, in *Proc. of Interspeech*, 2014, 223–227.

[11] Lee, J.; Tashev, I.: High-level feature representation using recurrent neural network for speech emotion recognition, in *Proc. of Interspeech*, 2015.

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

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

[14] Li, P.; Song, Y.; McLoughlin, I.; Guo, W.; Dai, L.: An attention pooling based representation learning method for speech emotion recognition, in *Proc. of Interspeech*, 2018, 3087–3091.

[15] Tzirakis, P.; Zhang, J.; Schuller, B.: End-to-end speech emotion recognition using deep neural networks, in *Proc. of ICASSP*, 2018, 5089–5093.

[16] Satt, A.; Rozenberg, S.; Hoory, R.: Efficient emotion recognition from speech using deep learning on spectrograms, in *Proc. of Interspeech*, 2017, 1089–1093.

[17] Ando, A.; Masumura, R.; Kamiyama, H.; Kobashikawa, S.; Aono, Y.: Speech emotion recognition based on multi-label emotion existence model, in *Proc. of Interspeech*, 2019, 2818–2822.

[18] Li, Y.; Zhao, T.; Kawahara, T.: Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning, in *Proc. of Interspeech*, 2019, 2803–2807.

[19] Neumann, M.; Vu, N.T.: 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, 1263–1267.

[20] Batliner, A.; Huber, R.: Speaker characteristics and emotion classification. *Springer*, **434** (2007), 138–151.

[21] Ben-David, B. M.; Gal-Rosenblum, S.; van Lieshout, P.H.H.M.; Shakuf, V.: Age-related differences in the perception of emotion in spoken language: the relative roles of prosody and semantics. *J. Speech Lang. Hearing Res.*, **62**, (2019), 1188–1202.

[22] Zhao, Y.; Ando, A.; Takaki, S.; Yamagishi, J.; Kobashikawa, S.: Does the lombard effect improve emotional communication in noise? - analysis of emotional speech acted in noise -, in *Proc. of Interspeech*, 2019, 3292–3296.

[23] Dang, J. *et al.*:Comparison of emotion perception among different cultures. *Acoust. Sci. Technol.*, **31** (6) (2010), 394–402.

[24] Sethu, V.; Epps, J.; Ambikairajah, E.: Speaker variability in speech based emotion models - analysis and normalisation, in *Proc. of ICASSP*, 2013, 7522–7526.

[25] Kim, J.; Park, J.-S.; Oh, Y.-H.: Speaker-characterized emotion recognition using online and iterative speaker adaptation. *Cogn. Comput.*, **4** (2012), 398–408.

[26] Sethu, V.; Ambikairajah, E.; Epps, J.: Phonetic and speaker variations in automatic emotion classification, in *Proc. of Interspeech*, 2008, 617–620..

[27] Nediyanchath, A.; Paramasivam, P.; Yenigalla, P.: Multi-head attention for speech emotion recognition with auxiliary learning, in *Proc. of ICASSP*, 2020, 7179–7183.

[28] Li, J.-L.; Lee, C.-C.: Attentive to individual: a multimodal emotion recognition network with personalized attention profile, in *Proc. of Interspeech*, 2019, 211–215.

[29] Ando, A.; Kobashikawa, S.; Kamiyama, H.; Masumura, R.; Ijima, Y.; Aono, Y.: Soft-target training with ambiguous emotional utterances for DNN-based speech emotion classification, in *Proc. of ICASSP*, 2018, 4964–4968.

[30] Fayek, H.M.; Lech, M.; Cavedon, L.: Modeling subjectiveness in emotion recognition with deep neural networks: ensembles vs soft labels, in *Proc. of IJCNN*, 2016, 566–570.

[31] Chen, Y.; Wang, J.; Yang, Y.; Chen, H. H.: Component tying for mixture model adaptation in personalization of music emotion recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.*, **25** (7) (2017), 1409–1420.

[32] Wang, J.; Yang, Y.; Wang, H.; Jeng, S.: Personalized music emotion recognition via model adaptation, in *Proc. of APSIPA*, 2012, 1–7.

[33] Yu, D.; Yao, K.; Su, H.; Li, G.; Seide, F.: KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition, in *Proc. of ICASSP*, 2013, 7893–7897.

[34] Ochiai, T.; Matsuda, S.; Lu, X.; Hori, C.; Katagiri, S.: Speaker adaptive training using deep neural networks, in *Proc. of ICASSP*, 2014, 6349–6353.

[35] Garimella, S.; Mandal, A.; Strom, N.; Hoffmeister, B.; Matsoukas, S.; Parthasarathi, S.H.K.: Robust i-vector based adaptation of DNN acoustic model for speech recognition, in *Proc. of Interspeech*, 2015, 2877–2881.

[36] Hojo, N.; Ijima, Y.; Mizuno, H.: DNN-based speech synthesis using speaker codes. *IEICE Trans. Inf. Syst.*, **E101.D** (2) (2018), 462–472.

[37] Delcroix, M.; Zmolikova, K.; Kinoshita, K.; Ogawa, A.; Nakatani, T.: Single channel target speaker extraction and recognition with speaker beam, in *Proc. of ICASSP*, 2018, 5554–5558.

[38] Lin, Z; Feng, M.; dos Santos, C. N.; Yu, M.; Xiang, B.; Zhou, B.; Bengio, Y. *et al.*: A structured self-attentive sentence embedding, in *Proc. of ICLR*, 2017.

[39] Abdel-Hamid, O.; Jiang, H.: Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code, in *Proc. of ICASSP*, 2013, 7942–7946.

[40] Busso, C. *et al.*: IEMOCAP: interactive emotional dyadic motion capture database. *J. Lang. Res. Eval.*, **42** (4) (2008), 335–359.

[41] Lotfian, R.; Busso, C.: Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Trans. Affect. Comput.*, **10** (4) (2019), 471–483.

[42] Ma, X.; Wu, Z.; Jia, J.; Xu, M.; Meng, H.; Cai, L.: Emotion recognition from variable-length speech segments using deep learning on spectrograms, in *Proc. of Interspeech*, (2018), 3683–3687.

[43] Ioffe, S.; Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift, in *Proc. of ICLR*, (2015), 448–456.

[44] Kingma, D.; Ba, J.: Adam: a method for stochastic optimization, in *Proc. of ICLR*, 2015.

[45] Elkan, C.: The foundations of cost-sensitive learning, in *Proc. of IJCAI*, (2001), 973–978.

[46] Park, D. S. *et al.*: SpecAugment: a simple data augmentation method for automatic speech recognition, in *Proc. of Interspeech*, 2019, 2613–2617.

[47] Paszke, A. *et al.*: Automatic differentiation in PyTorch, in *Advances in NIPS*, 2017.

[48] Lotfian, R.; Busso, C.: Curriculum learning for speech emotion recognition from crowdsourced labels. *IEEE/ACM Trans. Audio Speech Lang. Process.*, **27** (4) (2019), 815–826.

Atsushi Ando received the B.E. and M.E. degrees from Nagoya University, Nagoya, Japan, in 2011 and 2013, respectively. Since joining Nippon Telegraph and Telephone Corporation (NTT) in 2013, he has been engaged in research on speech recognition and non-/para-linguistic information processing. He received the Awaya Kiyoshi Science Promotion Award from the Acoustic Society of Japan (ASJ) in 2019. He is a member of ASJ, the Institute of Electronics, Information and Communication Engineers (IEICE), and the International Speech Communication Association (ISCA).

Takeshi Mori received the B.E. and the M.E. degrees from Tokyo Institute of Technology in 1994 and 1996 and the D.E. degree from University of Tsukuba in 2007, respectively. Since joining NTT in 1996, he has been engaged in research on speech and audio processing algorithms. He is a member of the IEEE, ASJ, and IEICE.

Satoshi Kobashikawa received the B.E., M.E., and Ph.D. degrees from the University of Tokyo, Tokyo, Japan, in 2000, 2002, and 2013, respectively. Since joining NTT in 2002, he has been engaged in research on speech recognition and spoken language processing. He received the Kiyasu Special Industrial Achievement Award in 2011 from IPSJ, the 58th Maejima Hisoka Award from the Tsushinbunka Assocsiation in 2012, and the 54th Sato Paper Award from ASJ in 2013. He is a member of ASJ, IPSJ, IEICE, and ISCA.

Tomoki Toda received the B.E. degree from Nagoya University, Japan, in 1999 and the M.E. and D.E. degrees from the Nara Institute of Science and Technology (NAIST), Japan, in 2001 and 2003, respectively. He was a Research Fellow of the Japan Society for the Promotion of Science, from 2003 to 2005. He was then an Assistant Professor, from 2005 to 2011, and an Associate Professor, from 2011 to 2015, at NAIST. Since 2015, he has been a Professor with the Information Technology Center, Nagoya University. He received more than 15 article/achievement awards, including the IEEE SPS 2009 Young Author Best Paper Award and the 2013 EURASIP-ISCA Best Paper Award (Speech Communication Journal). His research interest includes statistical approaches to sound media information processing.