Multimedia Tools and Applications (2023) 82:22759–22776
https://doi.org/10.1007/s11042-023-14680-y

End-to-end emotional speech recognition using acoustic model adaptation based on knowledge distillation

Hong-In Yun [1] · Jeong-Sik Park [2]

Received: 16 March 2022 / Revised: 28 July 2022 / Accepted: 3 February 2023 /
Published online: 13 February 2023

Abstract

The end-to-end approach provides better performance in speech recognition compared to the traditional hidden Markov model-deep neural network (HMM-DNN)-based approach, but still shows poor performance in abnormal speech, especially emotional speech. The optimal solution is to build an acoustic model suitable for emotional speech recognition using only emotional speech data for each emotion, but it is impossible because it is difficult to collect sufficient amount of emotional speech data for each emotion. In this study, we propose a method to improve the emotional speech recognition performance by using the knowledge distillation technique that was originally introduced to decrease computational intensity of deep learning-based approaches by reducing the number of model parameters. In addition to its use as model compression, we employ this technique for model adaptation to emotional speech. The proposed method builds a basic model (referred to as a teacher model) with a number of model parameters using an amount of normal speech data, and then constructs a target model (referred to as a student model) with fewer model parameters using a small amount of emotional speech data (i.e., adaptation data). Since the student model is built with emotional speech data, it is expected to reflect the emotional characteristics of each emotion well. In the emotional speech recognition experiment, the student model maintained recognition performance regardless of the number of model parameters, whereas the teacher model degraded performance significantly as the number of parameters decreased, showing performance degradation of about 10% in word error rate. This result demonstrates that the student model serves as an acoustic model suitable for emotional speech recognition even though it does not require much emotional speech data.

✉ Jeong-Sik Park
parkjs@hufs.ac.kr

Hong-In Yun
gnlenfn@gmail.com

[1] Department of English Linguistics, Hankuk University of Foreign Studies, Seoul, Republic of Korea

[2] Department of English Linguistics & Language Technology, Hankuk University of Foreign Studies, Seoul, Republic of Korea

Keywords Emotional speech recognition · Deep neural network · Model adaptation · Model compression · Knowledge distillation

1 Introduction

Automatic speech recognition (ASR) performs two main functions: extracting the acoustic features from audio signals and classifying them into appropriate text using acoustic models. Many types of acoustic models have been introduced according to pattern recognition approaches including dynamic time warping (DTW), support vector machine (SVM), hidden Markov model (HMM) and deep neural network (DNN) [1, 14, 23, 36]. Since DTW, SVM and HMM are simple but classic techniques, they are currently only used in limited speech recognition domains [25, 33].

The HMM-based approach estimates the phonetic characteristics of speech signals by a statistical measure with a Gaussian distribution, whereas the DNN-based approach aims to estimate model parameters of multiple layers using a large amount of speech data. The hybrid approach combining HMM and DNN has achieved much lower word error rate (WER) than the HMM-based approach [8].

One state-of-the-art ASR approach is end-to-end (E2E) speech recognition, which outperforms the conventional HMM–DNN hybrid approach. The E2E approach merges the separate acoustic and language models constructed by the conventional approach. Rather than concentrating on the specific tasks in each ASR procedure, E2E integrates the procedures into one system. Therefore, it requires a tremendous amount of computation for training the DNN.

By virtue of the E2E ASR systems, the state-of-the-art performance of speech recognition accuracy exceeds 95%. However, although ASR has considerably benefited from the technical breakthroughs in deep learning approaches, its performance is excellent only on normal data, and degrades on abnormal speech, especially emotional speech.

Like typical speech recognition models, emotional speech recognition was previously based on HMM. However, the word recognition rate (WRR) of HMM-based emotional speech recognition was only 30% [4]. In [28], the WRR was improved to 60%, but this performance was still below that of standard speech recognition in that era [13]. Emotional speech recognition usually weakens on standard speech models because emotional states cause meaningful variations in the speech parameters such as pitch frequency [10].

A way to achieve a reliable performance is to collect a considerable amount of emotional speech data and construct speech recognition module for each emotion [28]. This method needs to build several emotion-dependent acoustic models using the corresponding emotional speech data and requires an emotion-detection system at the front end to classify and assign the emotions to proper models, as shown in Fig. 1. However, the construction of emotion-dependent acoustic models is a difficult undertaking because it is not easy to obtain naturally verbalized emotional utterances from multiple speakers.

Another solution is extracting the acoustic features appropriate for emotional speech recognition. This requires knowledge of emotional expression through speech to extract useful features that properly represent emotions. However, identifying the characteristics of emotion from the speech is very ambiguous [16, 29]. Most studies analyze the speech characteristics such as prosody, pitch, and formant, but emotional expressions (and sometimes emotional states) are accompanied by facial expressions or other non-verbal languages. Classifying emotions themselves is highly challenging. Many of the emotion detection studies report

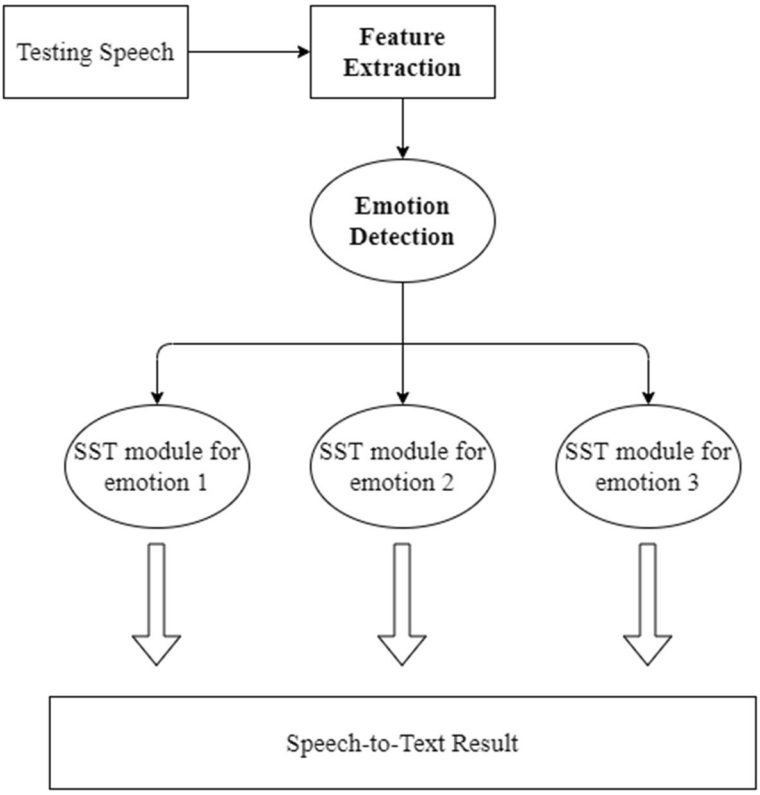

Fig. 1 Emotion-dependent model-based approach

better accuracy than 80% for two classes of emotions, but rapid degradation to 50% for over five classes of emotions [16, 30].

The last approach for emotional speech recognition is normalization of the features [20]. Emotional speech contains more information than typical automatic speech recognition. Normalizing these additional features can enhance the recognition accuracy. Cepstral parameters are essential feature in speech recognition systems, but vary in complex ways when emotion enters speech. For this reason, the cepstral characteristics are often difficult to normalize.

This study proposes an efficient emotional speech recognition approach based on an adaptation technique. Although several adaption techniques have been successfully applied to various pattern recognition problems, including image classification and speech recognition, they need to be carefully handled in an emotional speech recognition task that has domain-oriented ambiguity. In this study, we propose a knowledge distillation-based model adaptation approach for emotional speech recognition.

The knowledge distillation technique was originally introduced to decrease computational intensity of deep learning-based approaches by reducing the number of model parameters [15]. This technique builds a teacher model with a number of model parameters, and transfers the knowledge of the teacher model to the student model that has fewer model parameters. Thus, the student model performs similarly to the teacher model even with a relatively shallow model structure.

In this study, we attempt to build a student model that serves as an acoustic model suitable for emotional speech recognition. A teacher model is constructed from a large amount of

normal speech data that can be easily collected, and then a student model is built with a small amount of emotional speech data according to the knowledge distillation procedure. The student model is expected to have the characteristics of acoustic model adapted to emotional speech, because the model includes both the excellent speech recognition performance of the teacher model trained with a large amount of normal speech data and the emotional knowledge required for emotional speech recognition.

The main contributions of this study are summarized as follows:

(1) An efficient approach for emotional speech recognition is proposed.
(2) A new model adaptation technique based on knowledge distillation is proposed.
(3) The proposed approach efficiently performs both model compression and model adaptation.
(4) The proposed adaptation technique can be applied to various speech recognition tasks handling abnormal speech data that is difficult to obtain in large amounts, such as emotional speech and accented speech.

2 Material and methods

Emotional speech recognition differs from usual speech recognition in one important respect: its input data contain the emotional information. This additional information degrades the performance of standard speech recognition models. This study attempts to enhance the ASR model for emotional speech recognition by adapting the acoustic model to emotional speech using the knowledge distillation-based model adaptation.

2.1 Model adaptation for emotional speech recognition

In general, model adaptation is associated with domain adaptation, which is an algorithm that transfers information from a model trained in one or more "source domains" to a different but related "target domain" for the purpose of constructing a model pertinent to the target domain [9, 11, 22]. When trained on the source domain, a model can effectively deduce from a target domain that is insufficient or non-existent. For example, domain adaptation has been applied to diagnostic algorithms for artificial intelligence. A domain-adapted network trained on the labeled data of previous diseases is applied to new unlabeled data associated with a new disease like COVID-19.

Domain adaptation has recently gained much attention as a breakthrough technique that arrests the performance degradation caused by differences between the learning data and real-world data and the insufficient quality of benchmarking. Many domain adaptation approaches maximize the domain confusion by minimizing the difference between the distributions of the features extracted from the source and target domains [11, 22].

The domain adversarial neural network (DANN) is a representative adaptation method that generates target-domain data through an adversarial method and retrains the model to improve its performance [9]. However, this method has some shortcomings to be applied for emotional speech recognition. First, it requires two training sessions. The DANN generates target-domain data and is retrained for adaptation. Moreover, modifying the hyperparameters is difficult while training the networks. Because the performances of the generator and discriminator in DANN directly depend on the generated data quality, modifying the hyperparameters is costly.

Finally, as the generated data depend on the input target dataset, a DANN-based method is not easily generalizable.

This study proposes a domain adaptation approach based on knowledge distillation, which offers several advantages over the adversarial methods in emotional speech recognition tasks. First, it has a straightforward training process that reduces the cost. Furthermore, since models are compressed by knowledge distillation, this technique can be efficiently applied for emotional speech recognition that requires very complex model architecture to handle ambiguous emotional characteristics.

2.2 Acoustic model adaptation based on knowledge distillation

2.2.1 Knowledge distillation

In many cases, an ensemble model combining two or more networks provides a good performance [26]. Although the ensemble strategy majorly enhances the performance of a model, a whole ensemble model is cumbersome and demands excessive computational power when accessed by many users, especially if each model employs a large neural network. [15] proposed knowledge distillation to overcome the limitations of an ensemble-heavy model. The performance of knowledge distillation almost matches that of the big teacher model on a distilled student model but requires fewer parameters than the teacher model. Particularly, knowledge distillation offers model compression.

$$q_i = \frac{exp(z_i)/T}{\sum_i exp(z_i)/T} \tag{1}$$

Neural networks typically compute the class probability in a softmax output layer that converts the logit z_i computed for each class into a probability q_i, as described in (1). In this computation, the z_i is compared with other logits. The softmax function outputs a one-hot binary vector indicating the class assignments of z_i. This typical labeling system is called a hard target shown in Fig. 2.

In (1), the temperature parameter T creates a soft label (for a standard softmax function, T = 1). As pointed out in [15], a hard target accepts only the highest probability and dumps the others. The probability of abandonment might also play a role in transfer learning.

Typically, the softmax layer outputs an integer value. However, if T is greater than 1, the probability distribution is softened over the classes. After training on the target data set with assuming a soft target distribution of each case in the target data set, the softmax function of the high-temperature teacher model transfers information to the student model. The student model is trained on the same high-temperature softmax model, but in the post-training test stage, the softmax function is reverted to a standard softmax with T = 1. Figure 3 and Eqs. (2) to (5) show the calculation process of knowledge distillation loss.

$$C_{hard}(x,y) = -\sum_{i=1}^{K} q_i \log P_i(x) \tag{2}$$

$$C_{soft}(x,q) = -\sum_{i=1}^{K} q_i \log P_i(x) \tag{3}$$

$$Loss = (1-\alpha) C_{hard}(x,y) + \alpha C_{soft}(x,q) \tag{4}$$

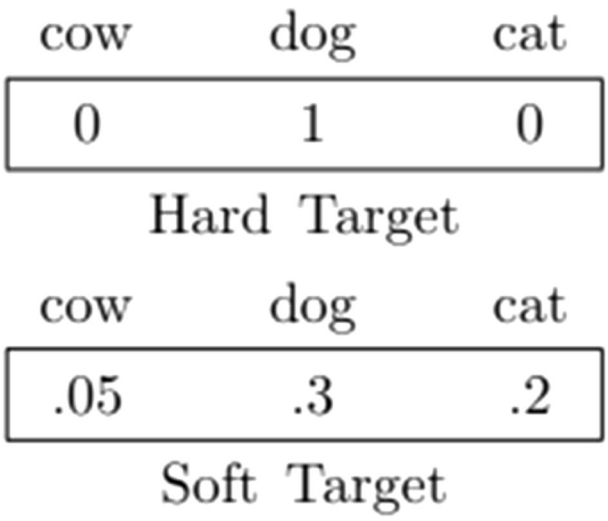

Fig. 2 Hard target and soft target

$$q_i = \frac{exp(z_i(x)/T)}{\sum_{j=1}^{K} exp(z_i(x)/T)} \tag{5}$$

The total loss in the network is the sum of the distillation and student losses [15]. The distillation loss is contributed by the soft label of the teacher model and the soft prediction of the student model, and the student loss is contributed by the hard prediction of the student model and the hard label of target domain data. The soft label from the teacher model and two types of predictions from the student model are obtained according to softmax with different temperatures.

The above equations are mathematical descriptions of knowledge distillation. In (2) and (3), $p_i(x)$ is the output probability of the i-th class of the student model, and q_i is a soft target of the input feature x. Therefore, $C_{hard}(x, y)$ is a one-hot vector output by the softmax layer, and $C_{soft}(x, y)$ is a softly distributed one-hot vector of softmax probabilities. In (4), α is the weight of the hard or soft label in the cross-entropy loss.

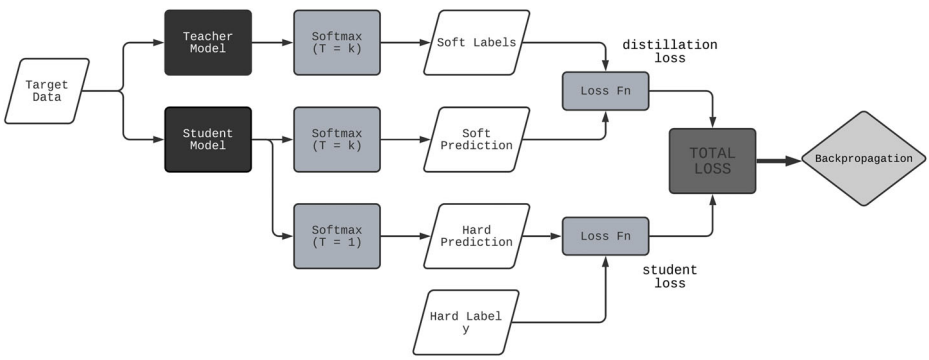

Fig. 3 Flowchart of knowledge distillation

As described above, knowledge distillation transfers the excellent performance knowledge of teacher models to shallow student models. After training on the extensive teacher knowledge, the student model delivers its best performance. Thus, knowledge distillation has been widely applied for model compression in DNN-based speech recognition tasks [7].

2.2.2 Knowledge distillation for end-to-end emotional speech recognition

Recent studies proposed knowledge distillation-based model compression approaches for E2E speech recognition [18, 32, 37]. As introduced in Section 1, E2E is a state-of-the-art ASR approach, which outperforms the conventional HMM–DNN hybrid approach. It is implemented as two types: connectionist temporal classification (CTC) and listen, attend and spell (LAS) [6, 12].

Among the two types, CTC is advantageous for E2E emotional speech recognition because it removes the need for post-processing; instead, the CTC decoder transforms the neural network output into the final text. Additionally, it provides the correct alignment between the training and transcript data, which is very important in emotional speech recognition, as it ensures the correctness of the acoustic model. For this reason, the CTC-based ASR approach is used as the baseline for emotional speech recognition in this study.

Knowledge distillation approaches typically transfer either the probability values of the classes in the teacher model, or the hidden layers in the middle of the teacher model. This study adopts the former approach. The equation that transfers the probability values of the frame units to the CTC model for training is given by (6) where x_t is the t-th frame from an input sequence of total length T [27, 31].

$$L_{CTC-KD_{frame}} = -\sum_{x \in Z} \sum_{t=1}^{T} \sum_{k \in K} P_{teacher}(k \mid x_t) \ln P_{student}(k \mid x_t) \tag{6}$$

2.2.3 Model adaptation process via knowledge distillation for emotional speech recognition

Knowledge distillation transfers the knowledge of the teacher model to the student model through a particular loss function, as described in Section 2.2.1. From a certain perspective, the loss Eq. (4) can be interpreted as model adaptation rather than as model compression. The first term of (4) is the standard cross-entropy loss function. The other term can be understood as a regularization term that restricts the student model from imitating the teacher model. From this perspective, we can apply the knowledge distillation method to model adaptation.

In order to handle emotional speech recognition, this study proposes an efficient model adaptation framework in which student models corresponding to emotions are respectively constructed and model compression of knowledge distillation is considered along with model adaptation. Figure 4 illustrates the conventional knowledge distillation concept and the proposed framework combining model compression and model adaptation.

The general concept of knowledge distillation focuses on model compression where the student model has fewer parameters than the teacher model, and the teacher and student models have the same input data, as shown in Fig. 4a. Considering in terms of model adaptation, the student model has the same model structure as the teacher model, and the teacher model is built with the source domain data and the student model is built with the target domain data, as shown in Fig. 4b.

In the proposed framework, the teacher models have a number of parameters trained with a huge amount of source domain data (normal speech data), whereas the student model is

considered as a simple network having fewer parameters trained with a small number of target domain data (emotional speech data). We expect that this combined framework realizes two usages of knowledge distillation at the same time.

When the compression and adaptation effects are applied simultaneously, we obtain a model with fewer parameters that adapts to the target domain (emotional speech). To this end, we alter the structure of the student model and set the input data of the student model as the target-domain data for model adaptation, thus reducing the number of parameters of the student model (obtaining a compressed model).

The adaptation process is focused on minimizing the knowledge distillation loss described in (4). The first term in this equation ensures that the adapted model distinguishes in domain data after training, while the second term allows the model to generalize by training on class-similarity information from the source model.

Figure 5 shows the process of model adaptation based on knowledge distillation for emotional speech recognition. The teacher model is firstly trained with source domain data. It is used to make soft label for the knowledge distillation process. Then the student model is trained with emotional speech data while minimizing the knowledge distillation loss. The loss is calculated from labels and predictions, as described in Section 2.2.1. The soft label and the

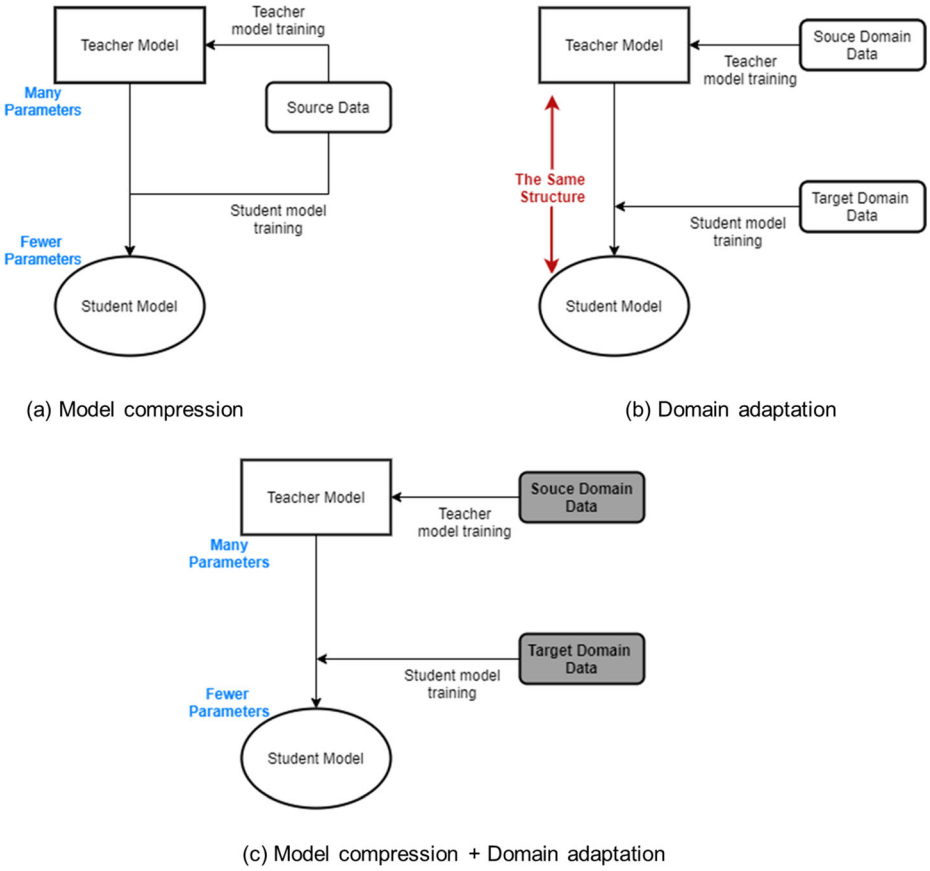

Fig. 4 Knowledge distillation concept of model compression and domain adaptation, and a combined framework in the proposed method

two predictions are the outputs of the teacher and student models, respectively. Hard label is obtained from the target domain data (i.e. emotional speech data).

The model adaptation task attempts to provide an adapted model matching the target domain from a source domain model. Various model architectures have been introduced in terms of DNN architecture, and each architecture provides different performance according to recognition target data and tasks. In this study, we employ the RNN architecture that is known to be an efficient model for sequentially varying data to construct teacher and student models for emotional speech. The models are composed of convolutional layers, recurrent layers, and fully connected layers. In particular, all training processes are conducted via the E2E speech recognition framework.

In the recognition process, a given emotional speech data is entered into each student model as an input data, and then recognized with a result of a model that provides highest output probability. The model is expectedly a target student model pertinent to the emotion of the given speech data.

2.3 Advantages of model adaptation based on knowledge distillation in emotional speech recognition

As described in Section 2.1, many domain adaptation methods have been applied to various research fields. The DANN that is a well-known domain adaptation method has achieved

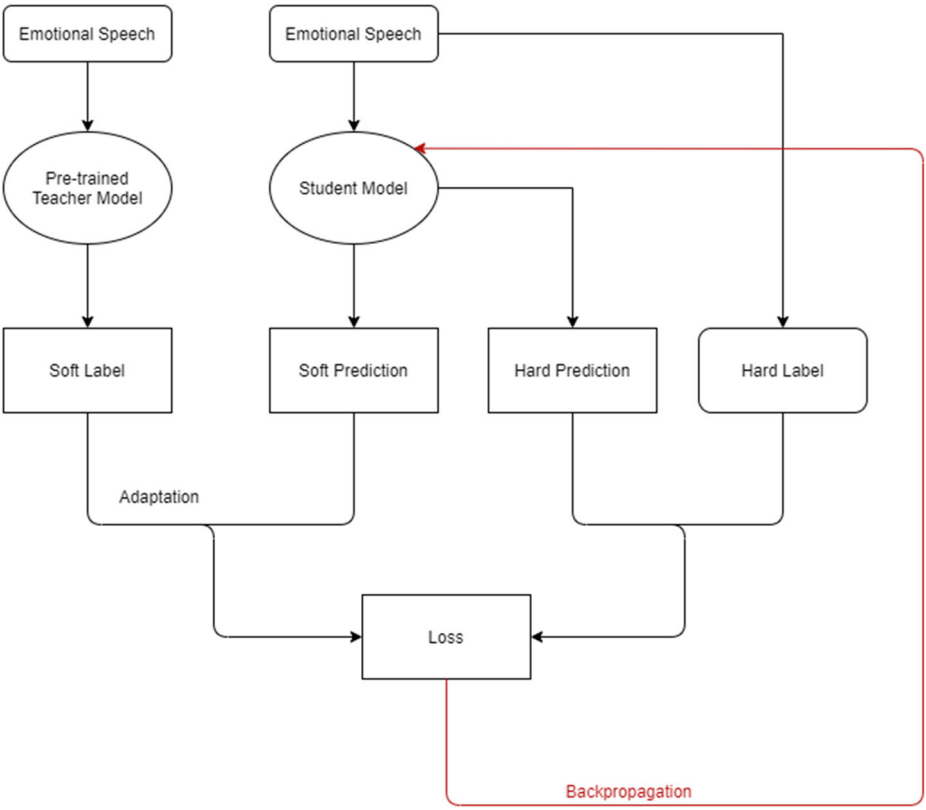

Fig. 5 Model adaptation process based on knowledge distillation for emotional speech recognition

superior performance in many fields, but it has some drawbacks. The method is a two-step process of generating new data and then training a model on the generated data. To accomplish these tasks, it requires two models: a generator that creates the target-domain data and a discriminator that distinguishes whether the generated data are target data or not.

To achieve outstanding performance, a robust discriminator is essential. Building a robust discriminator for adversarial domain adaptation requires dataset-dependent parameter optimization. Such a corpus-dependent optimized model is expensive and difficult to generalize. Therefore, it is unsuitable when models must be adapted to many domains or when a domain is continuously changing. Building many-case or constantly evolving domains is prohibitively expensive. In emotional speech recognition, a model is often built for each emotion. Applying DANN to emotional speech recognition will thus incur a tremendous computational cost.

The proposed adaptation approach based on knowledge distillation offers two advantages over the adversarial method. First, the knowledge distillation-based method requires only one training process. Knowledge distillation employs two models: a teacher model and a student model. Only the student model is adapted and the training of the teacher model is unrelated to the domain adaptation process. The teacher model can be trained on a pre-trained model to conserve time and resources. In DANN, building a model for each emotion requires a lot of training processes, but knowledge distillation greatly reduces the cost of building respective emotion models.

The second advantage of knowledge distillation is relaxation of the limitations on training the target domain. The adversarial method requires dataset-specific hyperparameter optimization when training the discrimination model. Accordingly, the generated data might be biased toward the given target-domain data, and additional target-domain data might be rejected. Adding more target-domain data is much easier in knowledge distillation than in the adversarial method. Moreover, the adaptation dataset is smaller than the training set. The time and cost of training the student model for adaptation are much reduced in knowledge distillation.

2.4 Experimental environments

To verify the efficiency of the proposed approach, we performed several experiments on emotional speech recognition. Some information about experimental setup is addressed in this section.

2.4.1 Data set

For the evaluation, we used LibriSpeech and interactive dyadic motion capture (IEMOCAP), which are representative speech data in speech recognition and emotion recognition domains [5, 19, 24]. First, we built the teacher model as the baseline model on LibriSpeech. The IEMOCAP data were then used for the target data in the adaptation process and the training data for the student model.

LibriSpeech is a corpus of English speech suitable for training and evaluating speech recognition systems [24]. It is derived from audiobooks that are part of the LibriVox project and contains 1000 h of speech. The corpus is divided into three subsets with approximate sizes of 100, 360, and 500 h. The speakers in the corpus were ranked by the WERs of the wall street journal (WSJ) model's transcripts and were divided into two roughly equal parts: the lower-WER speakers ("clean" group) and the higher-WER speakers (the "other" group). From the "clean" pool, 20 male and 20 female speakers were drawn randomly and assigned to the

development set. This process was repeated to form the test set. For each speaker in the development and test sets, we extracted ~8 min of speech (approximate total time = 5 h and 20 min in each group; see Table 1).

The "other" pool was similarly split into test and development sets and a single training set of ~500 h. However, the "other" pool was extracted from a subset with more challenging data than the "clean" pool. The WER computed by the WSJ models ranks the speakers in order of increasing difficulty of comprehensibility, and the speakers in the test and development sets were randomly chosen from the third quartile of this sorted list. Table 1 summarizes the subsets in the corpus. In this study, the baseline was trained on all 1000 h of speech.

The IEMOCAP corpus was designed for multi-modal emotion recognition [5]. It contains the data of 10 adult actors recorded during dyadic sessions. The actors were asked to read three selected scripts with explicit emotional content. The actors were also asked to improvise dialogues in hypothetical scenarios that elicit specific emotions (happiness, anger, sadness, frustration, and the neutral state). The speech data are augmented with the motion capture data of the subjects' faces (not used in this study). The database contains ~12 h of data in total. As the adaptation data, we used the data of three emotions (i.e., anger, happiness, and sadness) as representative emotion types. A total of 5 h and three emotions of speech data were finally used as the adaptation data. Table 2 shows the total lengths of the speeches associated with each emotion and the total number of audio files.

2.4.2 Experimental setup

The input feature was a 128-dimensional Mel-Spectrogram, and the answer label of CTC comprised 28 labels, including alphabets, blank space, and an apostrophe. The teacher and student models were based on Deep Speech2, and training was conducted with PyTorch. The knowledge distillation was also encoded in Python.

The teacher model (based on Deep Speech2 as mentioned above) was composed of three 2D convolutional layers, five bi-directional gated recurrent unit (Bi-GRU) layers, and one fully connected layer. Figure 6 shows the structure of Deep Speech2.

The teacher model was trained on a Titan Xp GPU for 30 epochs with a batch size of 10. Optimization was performed using an Adam optimizer with an initial learning rate of 5e-4 that gradually decreased every 5 epochs. The teacher model was configured as described in Table 3. The initial learning rate was 0.01, with a weight decay of 0.03. In the knowledge distillation, the temperature was set to 20. Training was performed for 20 epochs with a batch size of 10.

Table 1 Data subsets in LibriSpeech [24]

Subset	Hours	Per-spk Minutes	Female Spkrs	Male Spkrs	Total spkrs
Dev-clean	5.4	8	20	20	40
Test-clean	5.4	8	20	20	40
Dev-other	5.3	10	16	17	33
Test-other	5.1	10	17	16	33
Train-clean-100	100.6	25	125	126	251
Train-clean-360	363.6	25	439	482	921
Train-other-500	496.7	30	564	602	1166

Table 2 Size of adaptation set for each emotion

	Train (min)	Test (min)	Total Files
Anger	73.103	9.86	1103
Happiness	102.104	23.79	1636
Sadness	80.117	19.135	1084

Model adaptation with knowledge distillation requires the same structure of the teacher and student models. Both models were constructed with three convolutional layers and five Bi-GRU layers. The pre-trained teacher model (identical to the baseline model) was used to train the student model on additional adaptation data. In this experiment, we compared the WERs and CERs of the baseline and student models. This experiment also aimed to confirm whether model adaptation through knowledge distillation works properly for emotional speech recognition.

We additionally investigated the performance of emotional speech recognition when model adaptation and model compression were conducted simultaneously. The procedure was identical to the previous procedure, but the number of layers in the student model was reduced to lower the number of parameters. We also compared the compression rates of the convolutional and recurrent layers to determine which layer was most related to compression. For this purpose, each layer of the student model was decreased to 2. The total number of parameters was reduced by one-third.

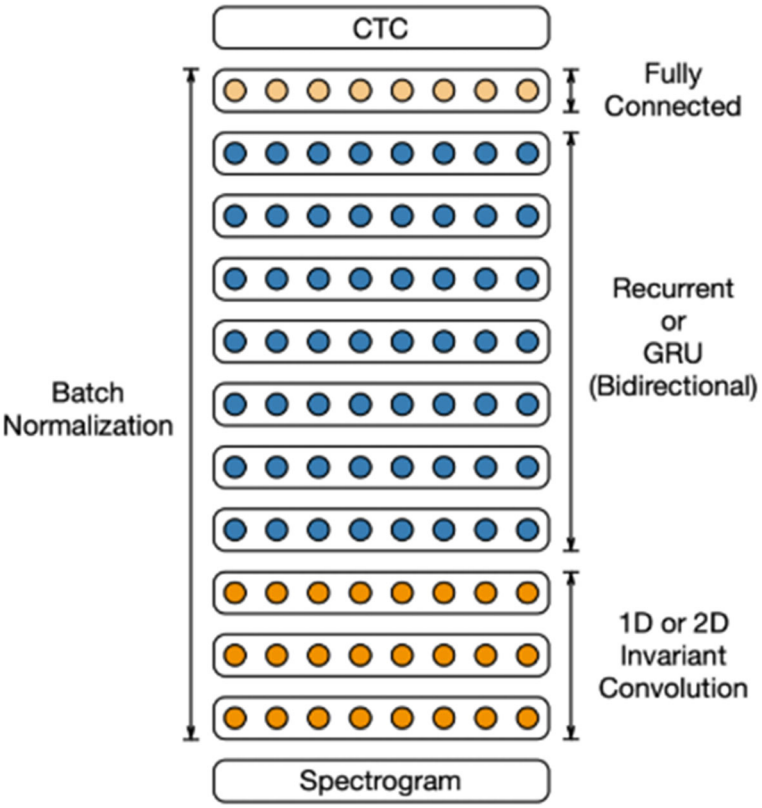

Fig. 6 Architecture of Deep Speech2 [3]

We repeated the experiment for each of the five sessions of the IEMOCAP corpus. Each session consists of a recording of a conversation between two speakers (male and female). The overall length of the test data depends on the emotion type, as shown in Table 4.

3 Results

Table 5 shows the performance of emotional speech recognition experiments. The results in the left two columns indicate that the student model (1) provided slightly lower performance than the teacher model (1). But when the size of the baseline model is reduced, the student model (2) gave similar performance to the student model (1), whereas the teacher model (2) yielded significant performance degradation compared to the teacher model (1).

The best CER performance of the baseline (teacher model (1)) was 24.6% on anger speech data, but the baseline achieved 39.4% CER on happiness speech data. Meanwhile, the best and worst WER performances of the baseline were 27.8% (anger) and 42.7% (happiness), respectively. When listening to the audio files, we did not recognize speech well because of lots of laughter sounds in the happiness data, which in part explains the low recognition result of the happiness category. Equally poor performance on sadness is due to the lower voice amplitude while expressing sadness.

Meanwhile, the student model achieved its best performance on anger (CER = 25.9%; WER = 29.1%) and its worst performance on happiness (CER = 41.4%; WER = 44.3%). However, a notable performance was observed in the student model (2). When the model size was reduced to about 40%, the performance of the student model was maintained, while the performance of the teacher model was significantly degraded. The student model (2) achieved a 25%, 9%, and 17% performance improvement in anger, happiness, and sadness, respectively, compared to the teacher model (2). In particular, the teacher model required 24 hours of training time, whereas the student model was trained in ~30 min. These results demonstrate that the student model provides distinct advantages over the baseline model in terms of training time and performance, and model adaptation with model compression via knowledge distillation works well in emotional speech recognition.

Next, we conducted additional experiments to examine the efficiency of model adaptation with model compression via knowledge distillation in more detail. For this work, we deleted each layer one by one for reducing parameters to cause model compression effects. Table 6 shows the results. In this table, the numbers before and after the solidus refer to the numbers of layers in the convolutional and Bi-GRU layers, respectively. Figure 7 represents average performance of emotional speech recognition results for the three emotion types and the number of parameters according to five model structures.

The Bi-GRU layer exerted little influence on the model compression. Reducing the number of Bi-GRU layers did not alter the file size or the number of parameters in the model. In contrast, the convolutional layer directly affected the number of parameters. Reducing the

Table 3 Configurations of the teacher model

Layer name	Kernel	Output channels	Dropout	Repeat
Conv1	(3,3)	32	0.1	3
Bi-GRU		512	0.1	5
FC		512	0	1

Table 4 Total length of test data

	Anger	Happiness	Sadness
Length (min)	9.86	23.79	19.13
Speakers	2	2	2

number of convolutional layers reduced the model size. The number of parameters was not reduced after decreasing the number of GRU layers but decreased rapidly after deleting a convolutional layer. Despite their small size, these models outperformed the reduced baseline. In fact, the performances were similar to those of the adapted model observed in Table 5. It showed difference less than 2%p in all models. The compressed student model maintained the performance of the original student model with minimal parameters.

4 Discussion

Table 5 is the performance of the first experiment conducted to evaluate whether model adaptation works properly in the proposed approach. In this experiment, we configured two experimental environments by making the teacher model and the student model of the same size. In the first environment with a large network size (about 23.7 million parameters), the performance of the student model was slightly lower than that of the teacher model. On the other hand, in the second environment with a small size (about 9.5 million parameters), the performance of the teacher model decreased significantly, while the performance of the student model was maintained, indicating much better performance than that of the teacher model.

In the proposed approach, a relatively small amount of emotional speech data was added when constructing the student model. Therefore, the results of this experiment indicate that the characteristics of emotional speech were well reflected in the student model even though the amount of emotional speech data was not large, so that the student model worked properly as an adaptive model.

Table 6 is the performance of the second experiment conducted to check whether model adaptation still works when model compression is performed by knowledge distillation. In this experiment, we fixed the teacher model with a large network size (about 23.7 million parameters) and investigated the performance while changing the network structure of the student model. In the experimental results, all student models with five types of networks showed similar performance for both CER and WER. In particular, even in the simplest structure (two convolutional layers and two Bi-GRU layers), the performance degradation was not significant.

Table 5 The performance of emotional speech recognition

Models	Teacher Model (1) 91 Mb		Student Model (1) 91 Mb		Teacher Model (2) 37 Mb		Student Model (2) 37 Mb	
File size # of parameters	23,705,373		23,705,373		9,506,269		9,506,269	
	CER	WER	CER	WER	CER	WER	CER	WER
Anger speech	24.6	27.8	25.9	29.1	34.6	40.5	26.0	29.9
Happiness speech	39.4	42.7	41.1	44.3	44.6	49.8	41.4	44.1
Sadness speech	39.1	39.2	42.3	41.4	50.2	51.8	42.4	41.4

Table 6 Result of adaptation and compression

Models	3/5		3/3		3/2	
File Size	91 Mb		55 Mb		55 Mb	
Parameters	23,705,373		14,251,805		14,233,053	
	CER	WER	CER	WER	CER	WER
Anger	25.9	29.1	26.4	30.0	26.1	30.3
Happiness	41.1	44.3	42.5	44.4	41.2	44.2
Sadness	42.3	41.4	41.4	43.0	41.7	41.2
Models	2/3		2/2			
File Size	37 Mb		37 Mb			
Parameters	9,525,021		9,506,269			
	CER	WER	CER	WER		
Anger	26.9	29.8	27.9	31.1		
Happiness	42.2	44.7	42.5	45.4		
Sadness	42.3	41.5	41.4	42.3		

This experimental result demonstrates that model compression and model adaptation work well at the same time through knowledge distillation in the emotional speech recognition task. If the student model was built for the purpose of model compression according to the general knowledge distillation procedure, and emotional speech data was not applied when building the student model, the performance of the student model may have deteriorated significantly. However, the student model built by the proposed method showed good performance, which demonstrates that the student model serves as an acoustic model suitable for emotional speech recognition even though it does not require much emotional speech data.

Regarding comparative verification with other research works, there are not many studies on emotional speech recognition compared to speech recognition studies on standard speech. Before 2010, there were studies based on HMM [35], a classical method, and recently, several studies based on DNN have been published [17, 34]. One reason is that there are many speech corpora for standard speech recognition, but few corpora for emotional speech recognition. In [34], experiments were conducted using the IEMOCAP corpus used in this study, but in

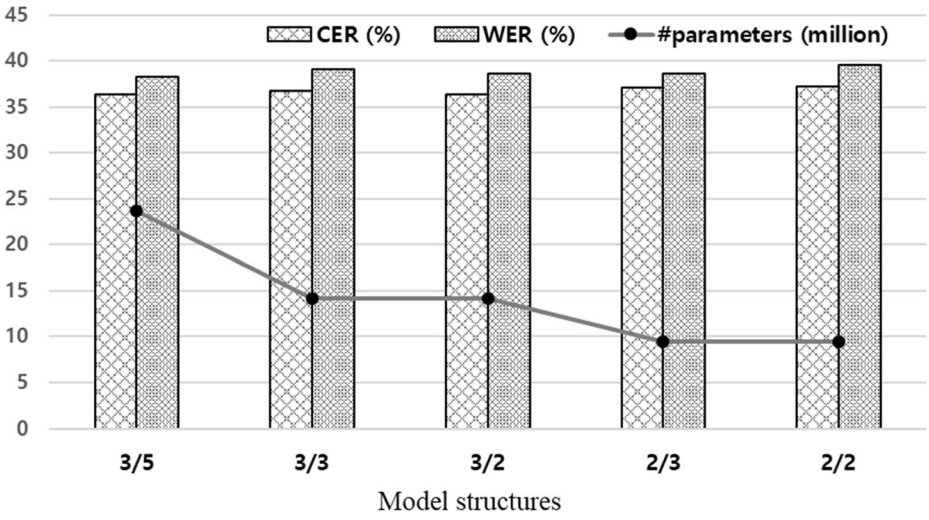

Fig. 7 Average performance and the number of parameters for each model structure

abnormal speech recognition tasks such as emotional speech recognition, learning and evaluating with the same data has limitations in terms of accuracy and reliability.

The main contribution of this study is to build a simple structure model (student model) that reflects the characteristics of emotional speech with a small amount of emotional speech data (IEMOCAP) from a complex structure model (teacher model) built with a large amount of standard speech data (LibriSpeech) via an model adaptation technique based on knowledge distillation. There are few studies related to model adaptation in DNN-based emotional speech recognition. In [17], an acoustic model built by simply including emotional speech data rather than a model using a specific model adaptation technique, is described as an adapted model. Therefore, we think that it is meaningless to intuitively compare the performance of this study with the results of the conventional research works. Meanwhile, in Table 5, teacher models (1) and (2) with a complex neural network structure represent acoustic models built by the existing DNN-based emotional speech recognition approach. For this reason, the superior performance of the student model (2) compared to the teacher model (2) indicates the efficiency of the proposed approach.

Based on the experiments, we confirmed that the proposed method works efficiently on the emotional speech recognition task. Nevertheless, the emotional speech recognition performance was still lower than the standard speech recognition performance with a WER of 5–8% in Deep Speech2 proposed by Baidu [3]. This explains why it is difficult to recognize abnormal speech such as emotional speech compared to standard speech. In this study, we confirmed the possibility that the proposed method could play a role in building a deep learning model reflecting the characteristics of abnormal data that are difficult to collect, such as emotional speech.

Knowledge distillation is similar to transfer learning in that it creates another model from one model, but the two techniques are distinctly different [2]. Recently, negative transfer is treated as an important issue in transfer learning [21], but the problem of negative transfer in knowledge distillation has not yet been identified. If knowledge is transferred negatively from the teacher model to the student model in knowledge distillation, there is a possibility of generating an incorrect student model, which may cause performance degradation. Thus, if this is elucidated, it will be possible to create a more robust and correct student model.

5 Conclusions

In order to overcome the limitations of emotional speech recognition, where it is difficult to obtain a large amount of data required to build a complete model, this study proposed an efficient emotional speech recognition approach based on knowledge distillation.

Knowledge distillation was originally proposed for model compression. We employed this technique to build an acoustic model suitable for emotional speech recognition, expecting that the student model would have the characteristics of an acoustic model adapted to emotional speech with a small amount of emotional speech data. For this task, we constructed the teacher model having a number of model parameters with an amount of speech data of normal voice, and then constructed the student models for respective emotions having fewer model parameters with a small amount of emotional speech data.

The experimental result showed that the performance of the teacher model significantly decreased according to the compression, whereas the performance of the student model was maintained. This result demonstrates that the proposed approach can be effectively applied for

emotional speech recognition in terms of model compression and model adaptation. In addition, the student model constructed in the proposed approach can be used as a deep learning model reflecting the characteristics of abnormal data that are difficult to collect.

In further study, we will investigate the model adaptation performance, increasing the number of adaptation data. In particular, if a sufficient amount of emotional speech data is provided, we can also investigate the capability of the proposed approach by comparing the performance with standard models constructed only with the target emotional speech data. In addition, we will investigate the performance improvement by a hybrid approach with other adaptation techniques used in different domains, such as DANN and transformer.

Acknowledgements This research was supported by Hankuk University of Foreign Studies Research Fund and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2020R1A2C1013162).

Declarations

Conflicts of interests/Competing interests There's no conflict of interest between the authors.

References

1. Aida-zade K, Xocayev A, Rustamov S (2016) Speech recognition using support vector machines. In: 2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT), pp 1–4
2. Alkhulaifi A, Alsahli F, Ahmad I (2021) Knowledge distillation in deep learning and its applications. PeerJ Comput Sci 7:e474
3. Amodei D, Ananthanarayanan S, Anubhai R et al (2016) Deep speech 2: End-to-end speech recognition in English and Mandarin. In: International Conference on Machine Learning, pp 173–182
4. Athanaselis T, Bakamidis S, Dologlou I, Cowie R, Douglas-Cowie E, Cox C (2005) ASR for emotional speech: clarifying the issues and enhancing performance. Neural Netw 18:437–444
5. Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan SS (2008) IEMOCAP: interactive emotional dyadic motion capture database. Lang Resour Eval 42(4):335–359
6. Chan W, Jaitly N, Le Q, Vinyals O (2016) Listen, attend and spell: a neural network for large vocabulary conversational speech recognition. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing, pp 4960–4964
7. Chebotar Y, Waters A (2016) Distilling knowledge from ensembles of neural networks for speech recognition. In: Interspeech, pp 3439–3443
8. Dahl GE, Yu D, Deng L, Acero A (2011) Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans Audio Speech Lang Process 20:30–42
9. Ganin Y, Ustinova E, Ajakan H et al (2016) Domain-adversarial training of neural networks. J Mach Learn Res 17(1):2096–2030
10. Gharavian D, Sheikhan M, Janipour M (2010) Pitch in emotional speech and emotional speech recognition using pitch frequency. Majlesi J Electr Eng 4(1):19
11. Goodfellow I, Pouget-Abadie J, Mirza M et al (2014) Generative adversarial nets. Adv Neural Inf Process Syst 27:2672–2680
12. Graves A, Fernández S, Gomez F, Schmidhuber J (2006) Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: 23rd International Conference on Machine Learning, pp 369–376
13. Graves A, Jaitly N, Mohamed A (2013) Hybrid speech recognition with deep bidirectional LSTM. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pp 273–278
14. Hinton G, Deng L, Yu D, Dahl G (2012) Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Sig Proc Mag 29(6):82–97
15. Hinton G, Vinyals O, Dean J (2015) Distilling the knowledge in a neural network. ArXiv preprint arXiv; 1503.02531

16. Kim JB, Park JS (2016) Multistage data selection-based unsupervised speaker adaptation for personalized speech emotion recognition. Eng Appl Artif Intell 52:126–134
17. Kosaka T, Aizawa Y, Kato M et al (2018) Acoustic model adaptation for emotional speech recognition using twitter-based emotional speech corpus. In: Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC), pp 1747–1751
18. Kurata G, Audhkhasi K (2018) Improved knowledge distillation from bi-directional to uni-directional LSTM CTC for end-to-end speech recognition. In: SLT, pp 411–417
19. Li Y, Zhao T, Kawahara T (2019) Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning. In: Interspeech, pp 2803–2807
20. Liu J, Zheng TF, Wu W (2006) Pitch mean based frequency warping. In: International Symposium on Chinese Spoken Language Processing, pp 87–94
21. Minoofam SAH, Bastanfard A, Keyvanpour MR (2021) TRCLA: a transfer learning approach to reduce negative transfer for cellular learning automata. IEEE Trans Neural Netw Learn Syst
22. Na HJ, Park JS (2021) Accented speech recognition based on end-to-end domain adversarial training of neural networks. Appl Sci 11:1–13
23. Najkar N, Razzazi F, Sameti H (2010) A novel approach to HMM-based speech recognition systems using particle swarm optimization. Math Comput Model 52(11–12):1910–1920
24. Panayotov V, Chen G, Povey D, Khudanpur S (2015) Librispeech: an ASR corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, pp 5206–5210
25. Park JS, Na HJ (2020) Front-end of vehicle-embedded speech recognition for voice-driven multi-UAVs control. Appl Sci 10(19):6876
26. Sagi O, Rokach L (2018) Ensemble learning: a survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8(4):e1249
27. Senior A, Sak H, Quitry FC et al (2015) Acoustic modelling with CD-CTC-SMBR LSTM RNNs. In: ASRU, pp 604–609
28. Sheikhan M, Gharavian D, Ashoftedel F (2012) Using DTW neural-based MFCC warping to improve emotional speech recognition. Neural Comput Appl 21:1765–1773
29. Singh YB, Goel S (2021) An efficient algorithm for recognition of emotions from speaker and language independent speech using deep learning. Multimed Tools Appl 80(9):14001–14018
30. Siriwardhana S, Reis A, Weerasekera R, Nanayakkara S (2020) Jointly fine-tuning "BERT-like" self-supervised models to improve multimodal speech emotion recognition. ArXiv preprint arXiv: 2008.06682
31. Takashima R, Li S, Kawai H (2018) An investigation of a knowledge distillation method for CTC acoustic models. In: ICASSP, pp 5809–5813
32. Takashima R, Li S, Kawai H (2019) Investigation of sequence-level knowledge distillation methods for CTC acoustic models. In: ICASSP, pp 6156–6160
33. Thiruvengatanadhan R (2018) Speech recognition using SVM. Int Res J Eng Technol 5(9):918–921
34. Trinh L, Dao T, Le T, Castelli E (2022) Emotional speech recognition using deep neural networks. Sensors 22(4):1414
35. Ververidis D, Kotropoulos C (2006) Emotional speech recognition: resources, features, and methods. Speech Commun 48(9):1162–1181
36. Xihao S, Miyanaga Y (2013) Dynamic time warping for speech recognition with training part to reduce the computation. In: International symposium on signals, circuits and systems (ISSCS), pp 1–4. https://doi.org/10.1109/ISSCS.2013.6651195
37. Yoon JW, Lee H, Kim HY, Cho WI, Kim NS (2021) TutorNet: towards flexible knowledge distillation for end-to-end speech recognition. IEEE/ACM Trans Audio Speech Lang Process 29:1626–1638