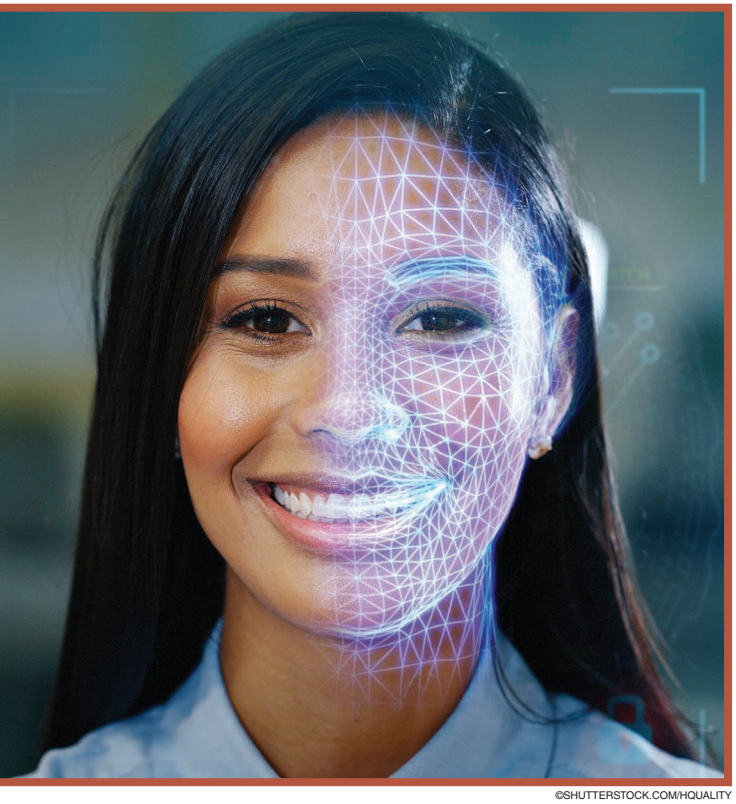
Chi-Chun Lee, Kusha Sridhar, Jeng-Lin Li, Wei-Cheng Lin,
Bo-Hao Su, and Carlos Busso

Deep Representation Learning for Affective Speech Signal Analysis and Processing

Preventing unwanted signal disparities



©SHUTTERSTOCK.COM/HQUALITY

Digital Object Identifier 10.1109/MSP.2021.3105939
Date of current version: 27 October 2021

Speech emotion recognition (SER) is an important research area, with direct impacts in applications of our daily lives, spanning education, health care, security and defense, entertainment, and human–computer interaction. The advances in many other speech signal modeling tasks, such as automatic speech recognition, text-to-speech synthesis, and speaker identification, have led to the current proliferation of speech-based technology. Incorporating SER solutions into existing and future systems can take these voice-based solutions to the next level. Speech is a highly nonstationary signal, with dynamically evolving spatial-temporal patterns. It often requires a sophisticated representation modeling framework to develop algorithms capable of handling real-life complexities.

Most of the variability in a speech signal comes from the interplay between lexical, paralinguistic, idiosyncratic, and many other pieces of contextual information, which are simultaneously conveyed in the speech signal. In particular, emotion directly affects the speech-production process, modulating the acoustic signal with expressive characteristics in a subtly complex manner. Many of the traditional signal processing methods are designed based on psychoacoustical knowledge to characterize these fine-grained patterns of the affect-related acoustic modulation [1]. Researchers in the field of SER have empirically derived standard feature sets accompanied with open toolboxes to obtain an off-the-shelf emotion recognizer [2], [3]. However, there is inevitable information loss due to the knowledge-driven—as compared to data-driven—process in computing these handcrafted features, which can underestimate the complexity in modeling the affective speech signal in real life.

With continuous advancements in deep representation algorithms, the modeling challenges around deploying SER solutions in our daily lives are being addressed across three core dimensions using deep networks for affective speech modeling:

- *Robustness*: How do we learn speech representations that can be robust against settings of signal acquisition and the nature of emotion manifestation to achieve robustness?

- *Generalization*: How do we learn speech representations that can handle source-target-domain mismatch in cross-context application scenarios to achieve generalization?
- *Usability*: How do we learn speech representations that can be practical during deployment, handling privacy and ethical concerns to achieve usability?

These modeling challenges are critical in leading SER technology into an integrative component of our daily lives. Our aim with this article is to provide the readers easy-to-follow materials that demonstrate the use of deep representation learning approaches in addressing these affective speech signal processing and analysis tasks. Our focus is mostly on the modeling part (i.e., back end), rather than on the feature-extraction part (i.e., front end), emphasizing deep learning strategies that are appropriate solutions for these challenges. For a detailed review of representation learning at the front end of SER, readers are referred to the work of Alisamir and Ringeval [4].

Directly using source speech representations on the target domain would result in significant degradation in performance, often due to inadequate generalization capability.

Three modeling challenges

Robustness, generalization, and usability are the three major affective signal modeling goals in deploying SER in the real world. Figure 1 provides an overview, highlighting key issues that can be addressed with deep representation learning methods for each of the three modeling goals. These three challenges are all interconnected; for example, a complex architecture might be more robust against noise in a real-world situation with better generalization to different conditions, but its computation power, memory requirement, and even privacy concerns may prevent it from deploying this system in actual applications. For the sake of presentation, however, we discuss each of them in separate sections.

In terms of robustness, the current available databases and models hardly cover the possible emotional speech variability space. To develop an SER system, we rely on the affective speech samples collected in predetermined contexts or scenarios, where the ground truths are often derived from manually annotated labels collected with perceptual evaluations. A key source of variability is the differences in the situated data acquisitions. The differences in microphones and in their placement create wide variations of affective speech data. Environmental noises also modulate emotion perception, reducing speech intelligibility [5]. Another source of variability is the individual's personal traits in expressing and perceiving emotions. The subjective differences in emotion perception affect the labels used to train the models as raters may perceive different emotions from the same speech [2]. Furthermore, the boundary between emotional categories are blurred for emotional expressions observed in daily interactions, introducing variations in the labels [6]. These uncontrollable factors are embedded within the collected speech signals and create additional technical difficulty in handling the expanded speech variability space to achieve robust SER.

In terms of generalization, an important modeling goal in SER is to learn the representations that can handle cross-domain mismatches, which result from data with different languages, labeling protocols, speaker traits, and interaction settings [7], [8]. The available representation models trained with data from a source domain should be able to maintain their emotion discriminatory power in a target domain, even when both domains may have inevitable mismatches. This goal is often difficult due to insufficiently labeled samples in the target domain, vast distributional emotional differences between the target and source domain, and inconsistent emotional descriptors. Directly using source speech representations on the target domain would result in significant degradation in performance, often due to inadequate generalization capability. Technically,

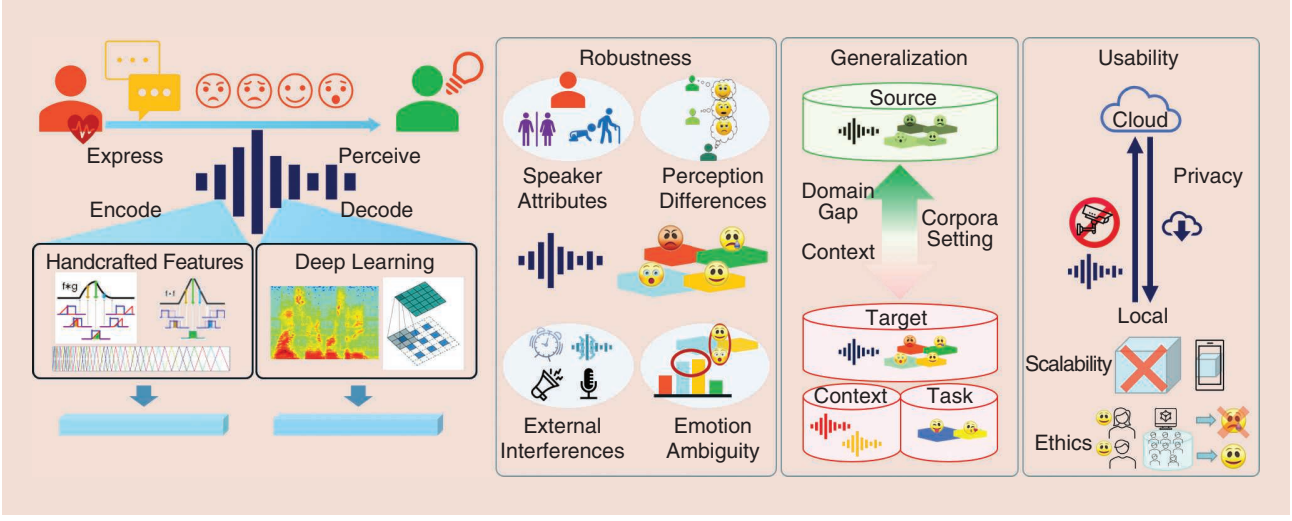


FIGURE 1. An overview of a deep representation learning scheme for SER, and the three real-world modeling challenges: robustness, generalization, and usability.

there is a tradeoff between model performance and generalizability. Without a careful design, insufficient training and fine-tuning can lead to negative transfer and domain shift problems [3], [9]. We can facilitate improving SER toward generalization by tackling the source-target mismatch on both speech features and emotion labels using deep learning approaches.

In terms of usability, practical issues in deploying SER need to be considered, such as model compactness, sensitive attribute masking, and fair representation. Real-world SER solutions often follow a cloud-local architecture that handles large-scale training and inferences. Current SER models tend to utilize deep architectures to achieve both robustness and generalization. However, edge devices allow only limited computation, storage, and memory access [10]. This constraint creates technical issues in learning SER to attain high efficiency. Privacy leakage during cloud-local transmission can also threaten the trustability of the SER model. Relying exclusively on edge computing would limit the model capacity. A better handling of the tradeoff between achieving high performance and addressing privacy issues is key for deploying a trustable SER system [11]. Another practical issue that should be carefully avoided is the exclusion of a target user demographic while building the representation. The collected affective speech databases can introduce ethics concerns, such as the unconscious inclusion of stereotypical bias and unfair representations. The protocol in data collection and speech feature would likely come under scrutiny as the SER model is rapidly being deployed and utilized as a decision-making aid in our daily lives.

Most of the studies addressing speech representation learning focus on the advancement of deriving unified front-end representations that can be applied to downstream speech tasks [12], [13]. In contrast, we discuss using deep representation learning to overcome these three key modeling challenges to enable SER in real-world applications: robustness, generalization, and usability. We first highlight the use of deep representation learning for each of the three associated affective speech modeling tasks. Then, we identify the current effective approaches in addressing these issues.

Affective speech modeling: Robustness

Robustness is an important consideration when modeling speech for emotion recognition. SER systems should be robust against signal- and emotional-based variabilities. Figure 2 shows a general overview of robustness challenges in the SER field. Signal-based variations are typically related to differences in the way in which we express emotions (e.g., interspeaker, gender, and phonetic variability) and the presence of external interferences (e.g., noise, channel variability, reverberation and far-field speech). Emotional-based variations are associated with the natural perceptual ambiguity among subjects in interpreting emotions, which affects the labels. Recent SER formulations have addressed signal- and emotional-based variations with model formulations, providing important insights for researchers working in this area. In the following sections, we focus mostly on discussing these formulations based on deep representation approaches.

Signal-based variations

Signal-based variations of affective speech can be represented with (1). The input audio signal $x[n]$ is transformed by the

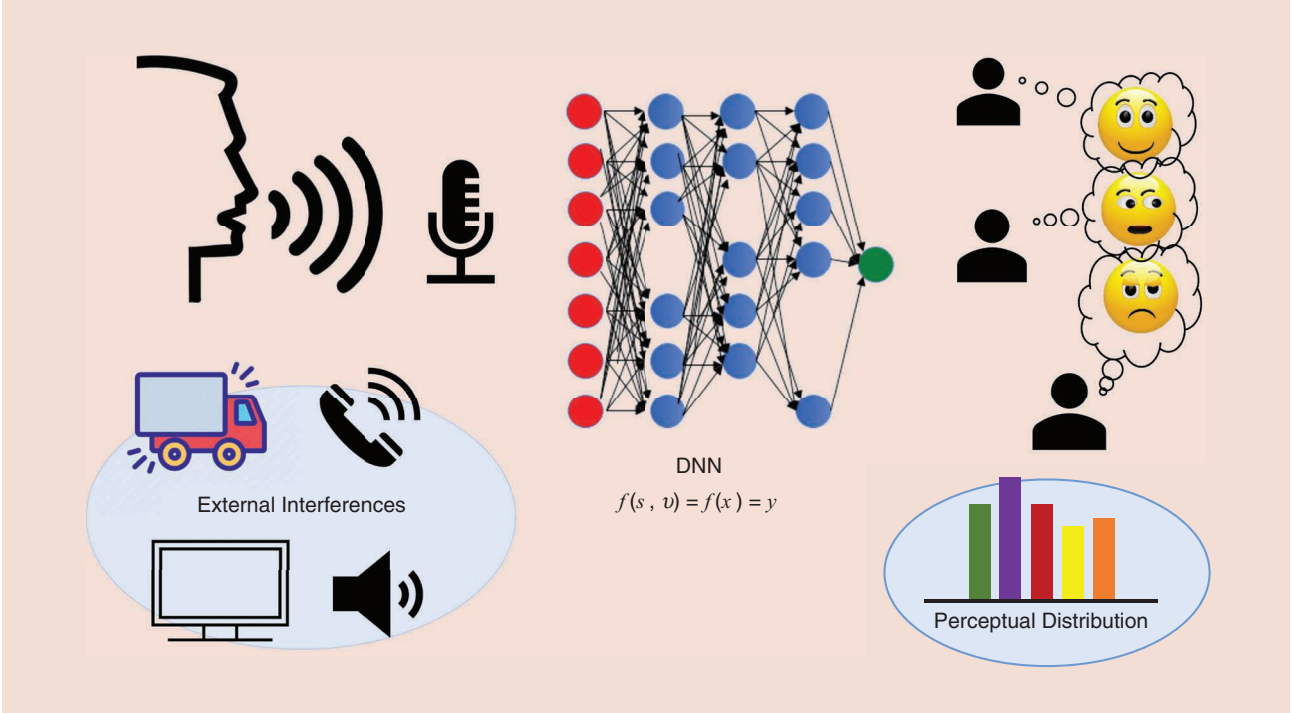


FIGURE 2. An overview of the robustness challenges for SER, which are generally divided into signal- and emotional-based variations. Modeling these variations by deep learning approaches are recent popular research trends in the SER field. DNN: deep neural network.

function $f(\cdot)$ into a vector, from where we can predict its emotional label y. This target output can represent either emotional categories (e.g., happy, neutral, or anger) or emotional attributes (e.g., arousal, valence, or dominance). In this article, we consider cases when function $f(\cdot)$ is implemented with a deep learning model.

$$f(x[n]) = y \tag{1}$$

$$x[n] = s[n] + v[n]. \tag{2}$$

When solving the issues of signal-based variations, we can further decompose the input audio signal $x[n]$ as the sum of a clean raw speech $s[n]$ and external interferences $v[n]$ [(2)]. The term $v[n]$ is a general one that includes different interference during data acquisition, such as the environment, background noise (e.g., stationary and nonstationary noise), and recording conditions (e.g., reverberations, far-field speech, and microphone settings). Likewise, the speech production of an individual varies with speaker traits such as vocal tract length and vocal fold size. Researchers have proposed solving these two components by utilizing deep learning approaches [5], [15].

Noisy speech enhancement

Ambient noises severely affect both the time- and frequency-domain structures of a signal, distorting the acoustic representation for an SER system. Various approaches have been proposed to resolve this issue, including data augmentation, feature compensation, and the extraction of robust acoustic features. A complementary technique to these methods is to apply a speech enhancement (SE) algorithm prior to the implementation of an SER model to attenuate the influence of the external interference $v[n]$. This approach is recommended because it is a straightforward and effective preprocessing step. For example, Triantafyllopoulos et al. [5] introduced a spectrogram-based SE network as a preprocessing module for an SER model. The SE network was independently trained from the SER by optimizing the mean square error (MSE) between the magnitude spectrum of the enhanced signal $|\hat{X}(t, w)|$ and the magnitude spectrum of the clean signal $|X(t, w)|$ [(3)]:

$$\mathcal{J} = \frac{1}{TW}\sum_{t}\sum_{w}(|\hat{X}(t, w)| - |X(t, w)|)^2. \tag{3}$$

The original noisy input audio $x[n]$ is first enhanced by the trained SE system to output $\hat{x}[n]$, which is the inverse of the short-time Fourier transform. The enhanced vector $\hat{x}[n]$ is then fed into the SER model. As the $v[n]$ term in the enhanced signal is attenuated, the modeling power of the SER model used to capture emotionally relevant features improves under different noisy environments, increasing its robustness against external interference. The evaluation on the Emo-DB corpus showed that the approach improved the unweighted average recall (UAR) from 14.73 to 20.75% on the −5-dB signal-to-noise ratio (SNR) condition, and from 18.05 to 26.85% on the 0-dB SNR condition [5]. Another elegant alternative is to jointly learn the SE and SER tasks. The enhancement model is embedded in the intermediate layers of an end-to-end SER framework [14] (see "The Robustness of SER Against Environment Noise"). This strategy allows the model to remove background noise or music while preserving important speech emotional cues. Note that although integrating SE techniques into SER tasks can effectively improve the system robustness, it also increases the complexity of the entire framework.

Robustness against speaker variability

We express emotion differently due to idiosyncratic or physical variations, which are reflected in the speech signal. The gender and phonetic variabilities are also important speaker-dependent traits included in $s[n]$. The approaches in SER aim to remove or adapt the idiosyncratic speaker information contained in $s[n]$ to improve the robustness of an SER model for unseen speakers. The conventional methods include speaker normalization, domain adaptation, and data selection. Recent advances in deep neural learning offer powerful adversarial schemes to achieve the goal of removing speaker characteristics. For example, Li et al. [15] proposed an adversarial training network for SER to disentangle the speaker and emotional characteristics. In addition to the cross-entropy loss ($\mathcal{L}_{\text{Emo}}$) for the primary SER task, they incorporated an entropy loss (H_{Spk}) in the cost function, as shown in (4). This term in the cost function encourages the model to increase the uncertainty or randomness of another independently trained speaker classifier output, forming a multitask setting to train the model:

$$\mathcal{L} = \lambda\mathcal{L}_{\text{Emo}} - (1 - \lambda)\mathcal{H}_{\text{Spk}}. \tag{4}$$

The maximization of the entropy term during the training is achieved by implementing a gradient-reversal layer. Because the SER model is trained to remove speaker traits, the approach reduces the sensitivity toward acoustic variability due to physical differences across individual speakers. This technique improved the performance of an SER system built without considering speaker variations, increasing the UAR from 57.45 to 59.91%, using the recordings of the IEMOCAP corpus. A limitation of performing adversarial learning is typically the optimization of a minimax problem, which might lead to unstable convergence issues.

Robustness against sentences of different lengths

Another practical problem that can increase variability is the duration of a sentence. The varying durations of each utterance are common in SER databases. The SER model should maintain robust recognition performance regardless of the duration of the signal. However, SER models are typically not capable of handling long sequences [16], which may even include dynamic emotional changes within a sentence (i.e., nonuniform emotion expressions over time). Given the fixed structure in the network, sentences are often zero padded or truncated to reach a target duration. These approaches affect the robustness

of the SER model, leading to severe performance degradation. For instance, Lin and Busso [16] demonstrated that the absolute concordance correlation coefficient (CCC) values across emotional attributes for long sentences (i.e., 8–11 s) were between 3 and 14% lower than the corresponding performances for short sentences (i.e., fewer than 5 s). A simple but effective method to address this problem is to split a sentence of arbitrary length T_i into a fixed number of segments or chunks with the same duration [16]. This goal is achieved by dynamically adjusting the step size Δc_i between chunks for different duration inputs. Equation (5) gives the step size for that sentence, where w_c is the fixed length of the chunks (e.g., $w_c = 1$ s) and C is the fixed number of desired chunks (e.g., $C = 10$, assuming the maximum duration of the sentences is 10 s):

The Robustness of SER Against Environment Noise

Figure S1 provides an example of a network implementation of an end-to-end SER system that is robust against noise interference. We follow the technique presented by Trigeorgis et al. [14]. Generally, the network consists of the following two parts:

1) *Part 1*: The input of an end-to-end framework is a time-domain audio waveform. The raw signal might contain noisy background, including music or other interferences that impact the performance of an SER system. Therefore, the first part of the network is two 1D convolutional neural network (CNN) layers that perform short- and long-term temporal convolution, respectively, on the raw signal. The short-term convolution aims to extract fine-scale spectral information from the high-sampling rate signal followed by a temporal-wise max-pooling operation. The long-term convolution intends to extract more higher-level, abstract features from the speech signal by intentionally increasing the kernel size for the second 1D CNN. Finally, the information is aggregated with a channel-wise max-pooling operation.
2) *Part 2*: After the CNN-based network, a recurrent-based network [i.e., bidirectional long short-term memory (BLSTM)] is concatenated to serve as the emotional discriminator. The optimization object is shared across the CNN and the BLSTM networks, jointly updating learnable weights to form the end-to-end training framework. The CNN network is considered a feature extractor, which has noise-reduction ability to remove undesired noises, while preserving important emotional cues for the BLSTM discriminator. The choice of the object function depends on the recognition task, where the cross-entropy loss is often used for emotional category tasks (i.e., the classification problem), and the concordance correlation coefficient is often used for emotional attribute prediction tasks (i.e., the regression problem).

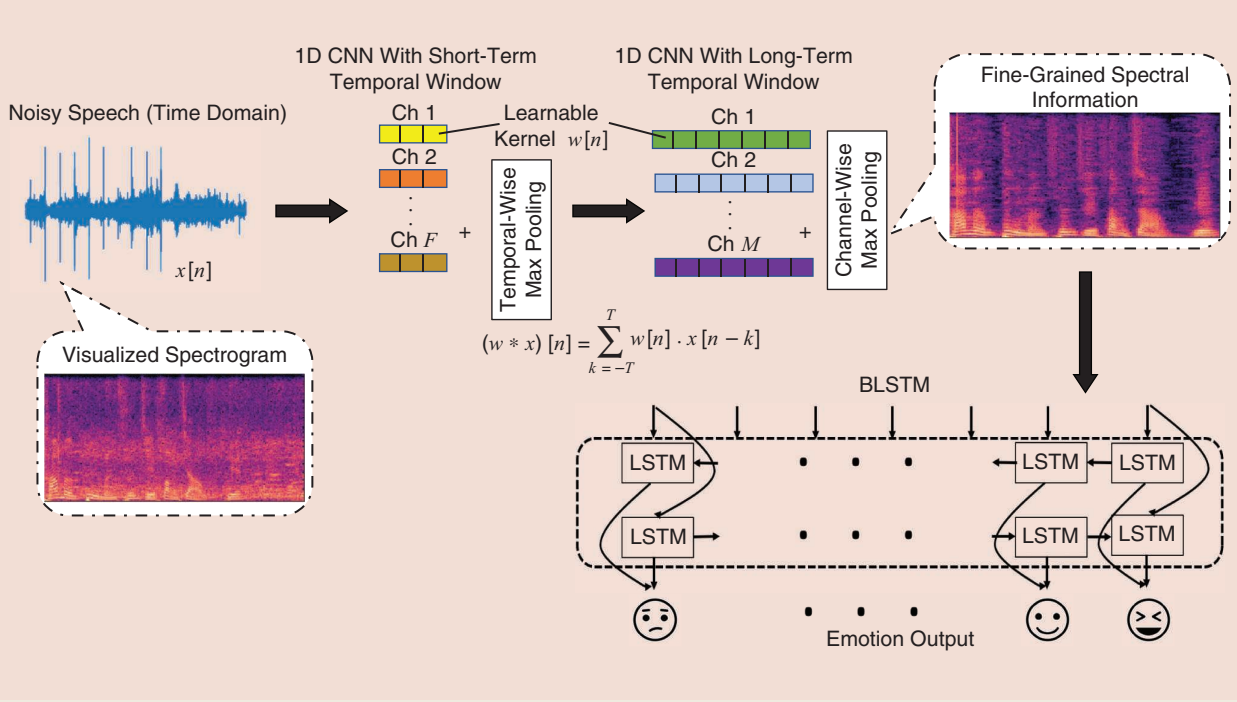


FIGURE S1. An example of the SER architecture proposed to learn deep emotional speech representation to deal with signal-based variations in an end-to-end framework [14]. BLSTM: bidirectional long short-term memory; Ch: channel.

$$\Delta c_i = \frac{T_i - w_c}{C - 1}. \tag{5}$$

As the method creates chunks with a fixed duration, it is straightforward to extract acoustic representations using methods such as long short-term memory (LSTM) or convolutional neural networks (CNNs) with fixed structures. As the number of chunks is also fixed, it is easy to aggregate the temporal information; for example, we can combine chunk-level information with attention models. The attention weights can effectively capture the emotionally salient regions within a speech regardless of its duration, resulting in robust performance toward sentences of different length. Furthermore, the model is highly parallelizable, leading to a computationally effective SER implementation. The experimental results systematically showed significant improvements in the CCC, with an absolute gain between 2 and 8% by using the adjusted step-size approach for different models (e.g., LSTM and CNN) across different emotional attributes (i.e., arousal, dominance, and valence).

Emotional-based variations

A second source of variability is the subjective nature of humans' emotional expression and perception. The ground-truth labels of emotion are often derived from human perceptual evaluations. As depicted in Figure 2, individuals may perceive different emotional content despite listening to the exact same audio clips. The perceptual variability leads to low interevaluator agreements in the labels, resulting in noisy, emotional labels that directly influence the robustness of machine learning-based SER systems. Studies have proposed strategies to deal with this problem using either label- or model-based solutions.

The most straightforward way to incorporate the differences in emotional perception in SER tasks is to train the model with soft labels.

Joint hard-soft emotional label learning

The most straightforward way to incorporate the differences in emotional perception in SER tasks is to train the model with soft labels. For example, if a speech recording is perceived as happy by two raters, angry by one rater, and sad by one rater, the model is trained with the vector [0.5, 0.25, 0.25, 0], where the dimensions represent the emotions of happiness, anger, sadness, and neutral, respectively. This approach explicitly includes label variation/noise (i.e., perception variabilities) in the optimization process while training the SER system. Instead of training an SER model based solely on the consensus ground truth (e.g., the majority or average of annotations), Chou and Lee [2] introduced a framework that jointly learns both consensus assignment (hard label) and emotion distribution (soft label). Equation (6) shows the soft label $q(c_k)$, which indicates the class probability for the kth emotional class. The variable K refers to the total number of emotion classes, and $\mathbb{1}_k^{(n)}$ is an indicator that is one when the nth annotator selects the kth class and zero otherwise:

$$q(c_k) = \frac{\alpha + \sum_n \mathbb{1}_k^{(n)}}{\alpha K + \sum_{k'} \sum_n \mathbb{1}_{k'}^{(n)}}. \tag{6}$$

The novelty in the soft-label definition is the use of a smoothing coefficient α, which controls the sharpness of the output label distribution. Therefore, the model not only encourages learning a single emotional target but also captures the label uncertainty reflected by the spread of the soft distribution. As a result, the trained SER model typically achieves higher recognition performance because it explicitly considers the natural variations of the emotional labels. The other methods used to leverage the difference in perception across evaluators are 1) finding trends across labels, 2) applying oversampling techniques that take into account the distribution of the labels provided to the samples, and 3) modeling uncertainty directly from the labels.

Model-uncertainty prediction

An alternative approach used to deal with emotional variability is capturing uncertainty in the model prediction as a part of the learning process. In addition to recognizing emotions, the SER system also estimates the confidence in its predictions. This information can be valuable for practical applications, especially in human-in-the-loop scenarios where ambiguous cases can be further reviewed. Sridhar and Busso [17] presented a model-level tactic to deal with label uncertainty. They designed an SER model with a reject option to recognize categorical emotions, enabling the model to abstain from providing a classification result when the confidence of the prediction falls below a certain threshold. The performance of a reject option is measured by reporting the accuracy of the system as a function of the test coverage, which is defined as the percentage of the test set over which the system provides a prediction. As the model rejects more samples, the SER performance is expected to increase as the most ambiguous samples are removed from the test set. The confidence in the results was estimated using two alternative criteria. The first criterion was used to minimize the empirical risk of the selective classifier while maintaining the test coverage as high as possible. The second criterion was used to compute the difference between the two highest classes predicted by the deep neural network (DNN) model. If the difference was above a certain threshold, the model was confident to make a prediction; otherwise, it rejected the sample. This approach effectively improved the recognition accuracy without significantly compromising the test coverage.

Another technique used to capture model uncertainty is the Monte Carlo (MC) dropout. MC dropout provides a way to calculate the intractable posterior distribution of the predictions by approximating variational inference using deterministic neural networks with dropout regularization. Dropout needs to be applied during both the training and testing stages. During training, dropout effectively samples a smaller network at every iteration in a tractable and feasible manner. This

approach is also computationally effective. Sridhar and Busso [6] used this technique to quantify uncertainty in the predictions of emotional attributes. Their study found that sentences with extreme values for valence, arousal, and dominance were predicted with less uncertainty, whereas more ambiguity was observed among neutral samples. Although this method is beneficial for modeling prediction uncertainties in SER, it requires multiple inference steps to calculate meaningful uncertainties. Also, a DNN with MC dropout is still dependent on the dropout and activation hyperparameters, which can modify the calculated uncertainty.

They designed an SER model with a reject option to recognize categorical emotions, enabling the model to abstain from providing a classification result when the confidence of the prediction falls below a certain threshold.

Affective signal modeling: Generalization

Generalization is the ability of a system to respond to novel situations not observed in the training data. Generalizing SER systems to adapt to new conditions is an important problem, with the increasing presence of speech-based systems across multiple domains, such as health care, security and defense, and education. It is difficult to learn the acoustic representations that capture general trends across multiple unseen scenarios. The factors that affect the generalization of SER systems are the scarcity of emotionally rich and balanced databases, accurate and consistent ground-truth labels, and differences in the interaction settings. Studies in SER have demonstrated that it is possible to circumvent several factors hindering generalization by employing machine learning approaches such as dropout [7], early stopping, data augmentation [18], [19], l_1 and l_2 weight regularization, and the use of a speaker-independent hold-out set to test the models. Figure 3 provides a broad perspective of the different ways to achieve generalization. This following sections describe promising tactics to improve the generalization of SER models, focusing on regularization-based, cross-domain adaptation, self-supervised learning, and proxy label approaches.

Regularization-based methods

DNNs often have millions of parameters that are extremely hard to optimize, especially when the train and test conditions are mismatched. This mismatch can be due to factors such as acquisition conditions, acoustic and emotion distributions, and even merely the types of human interactions. Reducing codependencies between the layers of a DNN can help with learning more generic trends across input instances, leading to better performance on unseen instances and improving generalization. In the following, we describe some of the recent and promising regularization techniques that have been used to achieve this goal.

Regularization with layer reconstruction

An approach used to regularize a DNN is with the reconstruction losses of intermediate layers, which are added to the main supervised problem. The most common formulations for this task are autoencoders and noisy autoencoders. The benefits of these auxiliary tasks is that they do not require emotional labels, so the unlabeled data from the target domain can be used to reduce the mismatch between train and test sets (see "Domain Adversarial Network for Generalized SER"). An appealing method to achieve

Domain Adversarial Network for Generalized SER

A domain adversarial neural network is trained with labeled data from the source domain and unlabeled data from the target domain. The network consists of a shared feature representation layer, followed by two pathways for the domain classifier and the SER task. There is a gradient reversal layer in between the feature representation layers and the domain classifier. A concise illustration of the network architecture is shown in the bottom left corner of Figure 3. The implementation of this method follows the following four steps:

1) *Step 1*: Train the network with labeled data from the source domain (database A) and unlabeled data from the target domain (database B). For example, database A can be a speech emotional corpus, and database B can be spontaneous recordings collected from the domain where the SER system will be used.
2) *Step 2*: The main task classifier is trained with labeled data from the source domain. This can be a categorical emotion classification task or an emotional attribute prediction task. The domain classifier is trained with data from both the source and target domains. The emotional labels are not needed for this auxiliary task.
3) *Step 3*: Both the main task and the domain classifiers are trained in parallel using the objective function described in (9). With the minmax nature of the loss function, the feature representations learned will be effective across the source and target domains.
4) *Step 4*: The training of a network with an adversarial loss function using a gradient reversal layer may be unstable if the hyperparameters are not properly adjusted, especially λ, which combines both of the losses [(4)]. A recommended strategy is to initialize λ to zero for the first 10 epochs and slowly increase its value until it reaches $\lambda = 1$ by the end of the training. As a result, the models will be properly initialized before the introduction of the gradient reversal layer.

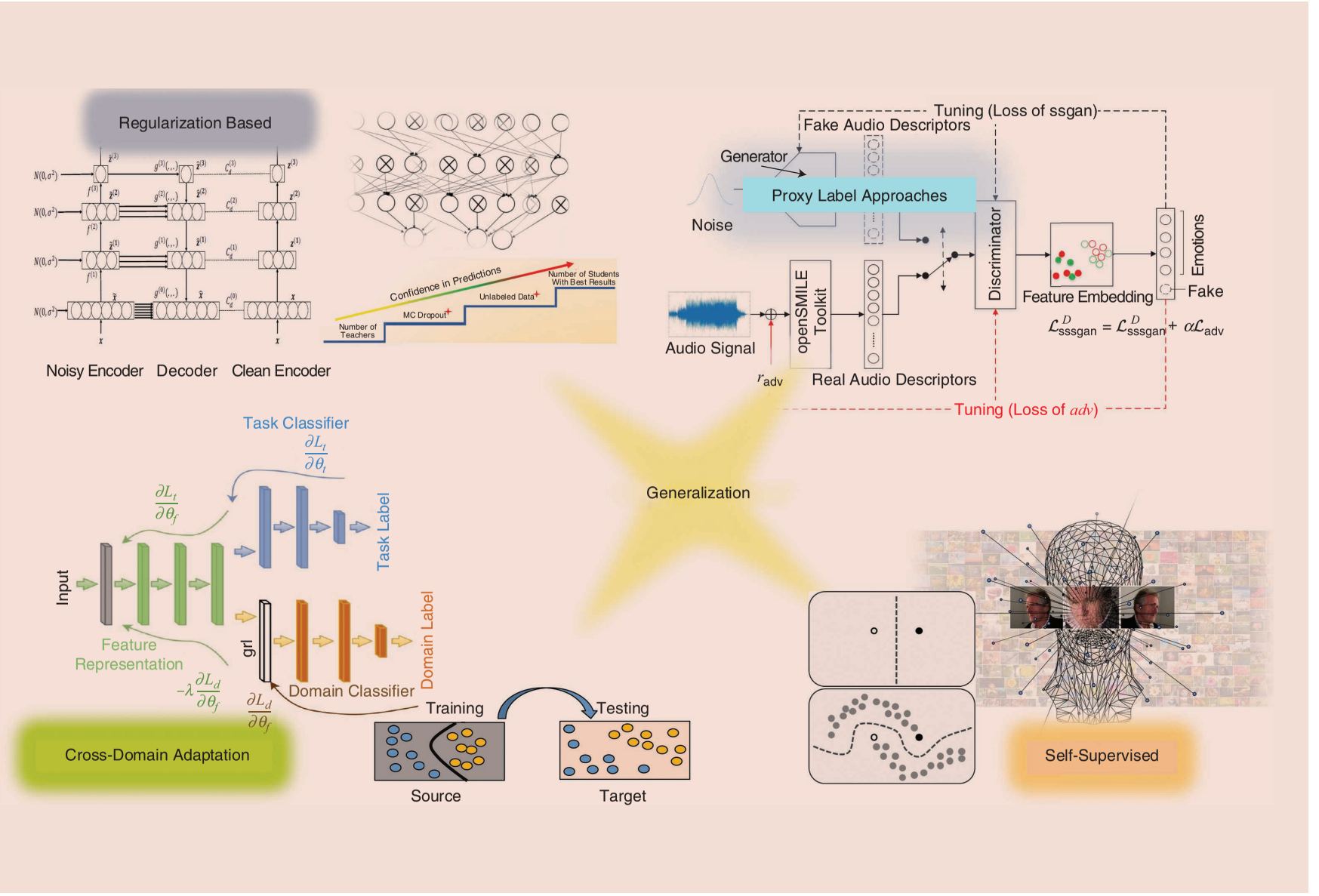


FIGURE 3. A generalization can be achieved through different strategies: Regularization-based methods, such as DNN architectures, layer reconstruction and MC dropout techniques; cross-domain adaptation, where DNN models are adapted efficiently to reduce domain mismatches; and self-supervised learning methods, which derive supervision from the data itself without additional labels, improving generalization and proxy label approaches, thus producing pseudolabels for unlabeled data to improve generalization in DNNs. adv: adversarial.

this goal is to use ladder networks. In contrast to simple autoencoders, the ladder network architecture uses skip connections between the encoder and decoder layers, attenuating the information overload from encoder to the decoder layers. Parthasarathy and Busso [20] explored this architecture, using multitask learning (MTL) as a regularization. The prediction of emotional attributes (arousal, valence, and dominance) was the primary task, and the reconstruction of feature representations at various layers in a DNN was the auxiliary task. By simultaneously solving the primary and auxiliary tasks, the models were regularized by finding more general high-level feature representations that are discriminative for the primary task. The unsupervised nature of the auxiliary task (i.e., reconstruction of intermediate feature representations) helped the SER system to improve its generalization by adding more unlabeled data from the target domain:

$$C_{\text{Lad+MTL}} = \alpha C_{\text{aro}} + \beta C_{\text{val}} + (1-\alpha-\beta) C_{\text{dom}} + \sum_{l} \lambda_l C_d^{(l)}. \quad (7)$$

The implementation of this approach is fairly straightforward, with the loss function shown in (7). The overall MTL loss of the ladder network consists of C_{aro}, C_{val}, and C_{dom}, which are the individual losses for the prediction of arousal, valence, and dominance, respectively. The loss function also includes $C_d^{(l)}$, which is the reconstruction loss at layer l in the network. The hyperparameters α, β, and λ_l are used to weigh these losses with $(\alpha, \beta) \in [0, 1]$ and $\alpha + \beta \leq 1$.

Regularization with MC dropout

A powerful approach to regularize a model is with dropout, where nodes in the network are randomly removed during training in each epoch. Dropout helps regularize a DNN by training thinner networks at each iteration, avoiding coadaptations among nodes of the network. Coadaptation appears when different hidden units in a neural network have highly correlated behavior. It is better for computational efficiency and the model's ability to learn a general representation if hidden units can independently detect the features of one another. Therefore, this coadaptation among nodes is detrimental to the models. Dropout is an effective technique to avoid these coadaptations. The studies on SER have used dropout in an MC fashion to develop generalizable models for SER. Sridhar and Busso [7] used knowledge distillation to learn acoustic representations under a teacher–student paradigm to improve consistency in the predictions, generalizing the models to diverse input conditions. In this process, an ensemble of teacher models was created with MC dropout. The study used multiple teacher models implemented with different dropout rates to increase the diversity of the ensemble. The learned feature embeddings of the teachers were used to train an ensemble of student models. This method was found to increase the consistency of the models by decreasing the uncertainty in the prediction of emotional attributes provided by the student ensemble.

Coadaptation appears when different hidden units in a neural network have highly correlated behavior.

The strategy from Sridhar and Busso [7] used MC dropout as a way to implement variational inference within a Bayesian deep learning framework to tackle the SER problem and achieved significant performance gains. This approach also points to a new direction of research for SER, where Bayesian learning methods can be successfully used.

Cross-domain adaptation

A cross-corpora evaluation of SER models often leads to a decrease in performance due to poor generalization across different conditions. The challenge of source-target mismatch is best solved by training models on large amounts of labeled data from the target domain. However, obtaining large-labeled data sets is expensive and time consuming. Learning-domain-invariant representations are a viable direction that helps with increasing SER performance.

Fine-tuning a front-end network

Another procedure for achieving generalization in SER is by fine-tuning the representations learned by a front-end network to model the emotional content of a target corpus. Lu et al. [21] used this technique instead of transfer learning between emotional corpora. To avoid the expensive human-annotation process and to cover a wide range of information, a speech front-end network, referred to as *AV-SpNET*, was trained with large-scale media data collected in the wild, which includes multiple languages across different domains. To assign pseudolabels to this unlabeled data, Lu et al. [21] proposed a rule-based method for assigning arousal labels based on prosodic information. The valence scores were derived from transcription of the recordings using sentiment analysis. AV-SpNET is built with an MTL structure using a CNN, where the primary task is to recognize the pseudolabels ($\mathcal{L}_{\text{emo}}$), and the secondary task is to reconstruct the inputs with an autoencoder ($\alpha \mathcal{L}_{\text{auto}}$). The model is optimized using a combination of reconstruction and proxy label-recognition losses, as shown in (8). Once the model is trained, the parameters of the AV-SpNET network are frozen. The outputs of this network are then used as the input of an SER system built for the target-domain problem. Therefore, this procedure is implemented in two steps. Although this approach needs no labeled data from the source and target domains, it requires a huge amount of data to achieve better generalization:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{auto}} + (1-\alpha) \mathcal{L}_{\text{emo}}. \quad (8)$$

Domain adversarial training

Domain adaptation is another technique to reduce the mismatch between train and test conditions. Abdelwahab and Busso [9] used an MTL framework with gradient reversal to achieve generalization. The auxiliary task is to create a domain classifier that recognizes whether the speech is from either the source or target domains. A shared feature representation between domains is learned, preserving discriminative information

for the SER task. This process is implemented using a gradient reversal layer, where the gradient produced by the domain classifier is reversed and propagated to the shared layers such that the model learns an indistinguishable representation for both domains. An added benefit of this method is that no labeled data from the target domain is necessary to train the domain classifier. The losses related to the attribute prediction and domain classification tasks are $\mathcal{L}_y$ and $\mathcal{L}_d$, respectively. The overall cost function can be obtained using (9):

Recently, generative models have been used to improve model generalization, especially with generative adversarial networks.

$$E(\theta_f, \theta_y, \theta_d) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_y^i(\theta_f, \theta_y) - \lambda\left(\frac{1}{m}\sum_{i=1}^{m}\mathcal{L}_d^i(\theta_f, \theta_d)\right), \quad (9)$$

where θ_f, θ_y, and θ_d represent the parameters of the shared layers, layers related to the main prediction task, and the layers of the domain classifier, respectively. Variable n is the number of labeled samples, and m is the number of labeled and unlabeled data samples. The model is trained using a minmax objective, with λ controlling the tradeoff between the two losses. The optimal setting is achieved when the prediction accuracy for the SER task is maximized and the performance for the domain classifier is at chance (i.e., the feature representation does not distinguish between the source and target domains).

Another example is the study from Chao et al. [3], which focused on learning an encoder to derive speech representations between domains that would mitigate the issue of emotional semantic inconsistency (i.e., the difference in the emotion labeling between two different domains). Specifically, the authors [3] designed an adversarial discrepancy learning strategy. This method involves an iterative minmax approach. First, they trained two source emotion regressor networks to result in a maximal recognition difference in the target domain while using the fixed-source speech representation encoder. Then they minimized this recognition difference by updating the source speech representation encoder while keeping these trained regressors fixed. This process then iterates until convergence. This particular strategy of minimizing the maximum difference does not simply learn an encoder to obtain common speech representation between domains, but also encourages such an encoded space to minimize the potential emotional semantic distortion between source and target. Although domain adversarial training techniques are very successful for generalization, the minmax nature of the losses used during training makes these methods unstable and sensitive to hyperparameter tuning.

Manifold subspace learning

Transfer learning can be effective to minimize mismatches between source and target domains. Zhang et al. [8] used a manner inspired by manifold and subspace learning to transfer knowledge from domains in a cross-corpus SER evaluation. The procedure constructs a neighbor graph that can measure the differences in the feature distributions between the source and target data sets by preserving the geometrical structure of the data. During training, they learn a corpus-invariant projection matrix that aligns the features from different corpora into a common discriminative subspace, which leads to improvements in SER performance. This is a joint framework that combines a discriminative subspace learning, a graph-based distance metric and a feature-selection method. It is not completely unsupervised, which helps the model to avoid learning irrelevant feature representations that may not be discriminative for the SER classifier. However, this technique is implemented in an iterative manner and needs to tune several hyperparameters. The approach also uses the $l_{2,1}$-normalization, which combines in a single function the l_1 and l_2 constraints. This optimization is implemented in an iterative manner, first the l_2 constraint and then the l_1 constraint, which makes it computationally expensive.

Generative models

Recently, generative models have been used to improve model generalization, especially with generative adversarial networks (GANs). GANs use an adversarial approach between a generator, which creates synthetic samples matching a target distribution, and a discriminator, which determines whether the samples are real or created by the generator (i.e., fake). The key idea in using generative models is to create rich representations, especially when the training data are limited. Sahu et al. [22] used a model relying on GANs to classify emotions based on the low-dimensional feature bottleneck layer of an autoencoder. The lower-dimensional encoding space is matched with a simple prior distribution p_z by using a GAN formulation. They used samples from this lower-dimensional subspace as the input to the decoder of the autoencoder to obtain synthetic feature vectors, which were further used for SER. Sahu et al. developed the following three different variations of this autoencoder-GAN framework:

1) a GAN framework used to match the encoding space to p_z
2) an additional GAN to match the output of the decoder (synthetic features) to the real feature vectors (input features)
3) a similar framework to the first two methods, but conditioning the GAN with an emotion class label.

They trained the models with the MSE loss as the reconstruction loss to update the autoencoder weights, cross-entropy loss for the GAN, and an additional mutual information loss for the conditional GAN. This approach was evaluated using cross-corpus experiments with low-resource conditions on the training data, demonstrating the improvement in generalization by adding synthetic samples. The study trained SER models with very few samples from the source corpus. They progressively added more synthetic samples, observing higher improvements in recognition accuracy on the target corpus as they added more synthetic data.

Bao et al. [19] presented another study that uses a generative model to address the problem of data scarcity in SER. This study used CycleGAN to generate synthetic feature representations that aim to reduce the distance between synthetic and real data and increase the emotional discrimination of the feature representation. CycleGAN can map the source and target domains without paired training data. CycleGAN implements a bidirectional mapping between the source and target domains, where an adversarial discriminator generates synthetic samples indistinguishable from the real samples. A classifier is added to discriminate emotions between the generated data to learn a generalized distribution from real source samples, avoiding that the model only reconstructs the original data. The overall loss function is given by (10):

Multiview learning aims to learn one function to model each view and jointly optimizes all the functions to improve the generalization performance.

$$\mathcal{L} = \sum_i \mathcal{L}_i^{\text{GAN}} + \lambda^{\text{cyc}} \sum_i \mathcal{L}_i^{\text{cyc}} + \lambda^{\text{cls}} \mathcal{L}^{\text{cls}}, \tag{10}$$

where $\mathcal{L}_i^{\text{GAN}}$ is the minmax GAN loss for each emotional class i, $\mathcal{L}_i^{\text{cyc}}$ is the cycle-consistency loss, and $\mathcal{L}^{\text{clc}}$ is the softmax cross-entropy loss of the classifier added to the model. The cycle-consistency loss accounts for translating back the synthetic data from the target to the source domain, calculating the MSE between the real and generated samples. The λ^{cyc} and λ^{cls} weights are the hyperparameters of the model.

Su et al. [18] proposed a novel approach to achieve cross-corpus emotion recognition. They trained a CycleGAN with two generators to learn a bidirectional mapping between source (S) and target (T) corpora with the goal of generating synthetic source domain samples that are target aware. They used labeled data from the source domain and unlabeled data from the target domain. To achieve their objective, they conditioned the generator to learn the mapping from the target to the source domain ($G_{T\to S}$) on the emotional labels from the source domain. They enforced two regularization constraints: identity and cycle-consistency loss. Identity loss maintains the source-domain information after transformations. This objective is achieved by transforming the source samples using the $G_{T\to S}$ transformation back to the source domain. Then the loss function imposes that the transformed samples should be similar to the actual source samples. This loss also ensures that the samples in the target domain that are similar to the samples in the source domain are not heavily transformed. The cycle-consistency loss ensures that the samples undergoing bidirectional transformation ($G_{S\to T}$ to $G_{T\to S}$) are identical to the original samples.

The methods presented by Sahu et al. [22], Bao et al. [19], and Su et al. [18] need a two-stage training procedure where first, the generator of the GAN is trained to generate fake samples and second, the augmented samples are used to train the SER classifier. Another disadvantage is the use of the minmax loss function, which makes the GAN training very unstable and sensitive to hyperparameter tuning.

Self-supervised learning methods

An appealing approach to improve the generalization of the SER models is the use of self-supervised learning, where the key idea is to derive labels for auxiliary tasks directly from the data. The aim of this formulation is that by solving these auxiliary tasks, the model has to extract general patterns from the data, which leads to feature representations that generalize better for the main task. An example of self-supervised learning in other domains includes masking a word in a sentence, expecting that the network would predict the missing word. The label here is the missing word, which is freely available. Another example in natural language processing is changing a word in a sentence for a random word and asking the neural network to identify the mistake. In SER, this process can be used with predictive and contrastive models.

Predictive models

A class of self-supervised models are predictive models, where the loss function is computed in the output space by estimating performance between predicted and ground-truth labels. The losses include the self-supervised tasks or the main target task. An example of predictive models in SER was proposed by Pascual et al. [12]. This study developed a problem-agnostic speech encoder to learn general speech representations to tackle different downstream supervised tasks (e.g., SER or speaker recognition). They designed a single-neural encoder followed by several small subnetworks to jointly solve multiple self-supervised tasks. These subnetworks consist of self-supervised tasks, including reconstruction of the waveform, and acoustic features such as Mel-frequency cepstral coefficients and prosody features. The general representations created with these self-supervised tasks are used for SER and other speech tasks by either freezing the encoder or fine-tuning the encoder and task classifier.

Contrastive models

Using contrastive losses in the feature representation space is another self-supervised approach. A technique for achieving this goal is by adding data perturbations to create different views from the data and then training the network to minimize the mismatches between the views. The concept of "views" in this context comes from the multiview training strategy, where multiple inputs are used to train a system. The inputs could be perturbations of a single modality (e.g., adding noise, adversarial changes through gradient reversal, and speech rate manipulations) or different features extracted from the signal (e.g., extracting features from statistics over frame level features, spectral features, and temporal features). Using information from different views can significantly improve the model performance. Multiview learning aims to learn one function to model each view and jointly optimizes all the functions to improve the generalization performance. Combining contrastive and reconstruction losses has been an effective technique in SER tasks.

Jiang et al. [13] used different views of the input data by introducing augmentations such as random pitch shift, speed perturbation, room reverberation, and additive noise to the raw speech waveforms and spectrograms. They learned speech representations using encoders. For a given input sample, they used two different augmentations to provide two correlated views of the sample and learned an encoded representation by simultaneously training two encoders. They used the MSE loss between the encoded representations and the input raw feature. The encoders were built using the transformer layers. By maximizing agreement between the learned feature representations of the two augmented samples via a contrastive loss in the latent space, they extracted meaningful feature representations for an SER downstream task. The contrastive loss objective used here is called the *normalized temperature-scaled cross-entropy loss*. Equation (11) shows this loss, where $sim(z_i, z_j)$ represents the cosine similarity between the encoded feature representations. The function $\mathbb{1}_{[k=i]} \in [0, 1]$ is an indicator function that is one if $k \neq i$ and zero otherwise. The variable τ denotes a temperature parameter. Equation (11) represents the loss function for a positive pair of examples (i, j). The final loss is computed across all positive pairs [both (i, j) and (j, i)] within a minibatch:

$$\mathcal{L}_{i,j} = -\log \frac{\exp\left(\frac{sim(z_i, z_j)}{\tau}\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp\left(\frac{sim(z_i, z_k)}{\tau}\right)}. \tag{11}$$

Proxy label approaches

Proxy label methods are a particular class of semisupervised learning (SSL) algorithms that focus on producing pseudolabels on unlabeled data, augmenting the training set. These proxy labels are produced by the model itself, or variants of it, without any additional supervision. Some of the prominent SSL methods relying on proxy label include self-teaching, self-ensembling, and multiview training.

Generative modeling with label smoothing

Zhao et al. [23] proposed a semisupervised GAN (SSGAN) with proxy labels to learn powerful feature representations to achieve state-of-the-art results on publicly available emotional corpora. In addition to classifying the inputs as real or fake, the SSGAN discriminator is also able to predict the emotional class of a sample. They used a divergence-probability measurement to smooth the conditional label distribution given the inputs using adversarial and virtual adversarial training methods. This distribution smoothing process using virtual labels for the unlabeled set serves as a proxy label to train the model. This method does not need to include complex architectures. Abdelwahab and Busso [9] implemented each of the three blocks (the feature representation layers and the domain and task classifiers) with a two-layer DNN, each with 256 nodes.

Affective speech modeling: Usability

Performance should not be the only metric considered when deploying SER systems for real-world applications. There are also major usability constraints that need to be addressed, including scalability, privacy, and ethics requirements, where representation learning methods can be used to mitigate the potential issues associated with these considerations. The current application systems are often set up as cloud servers with local end users, as shown in Figure 4. Several issues arise when providing SER services with this setting. For example, transferring local data over the Internet may result in delayed responses or even worse: in the leakage of private information. If we decide to perform model prediction in the local devices

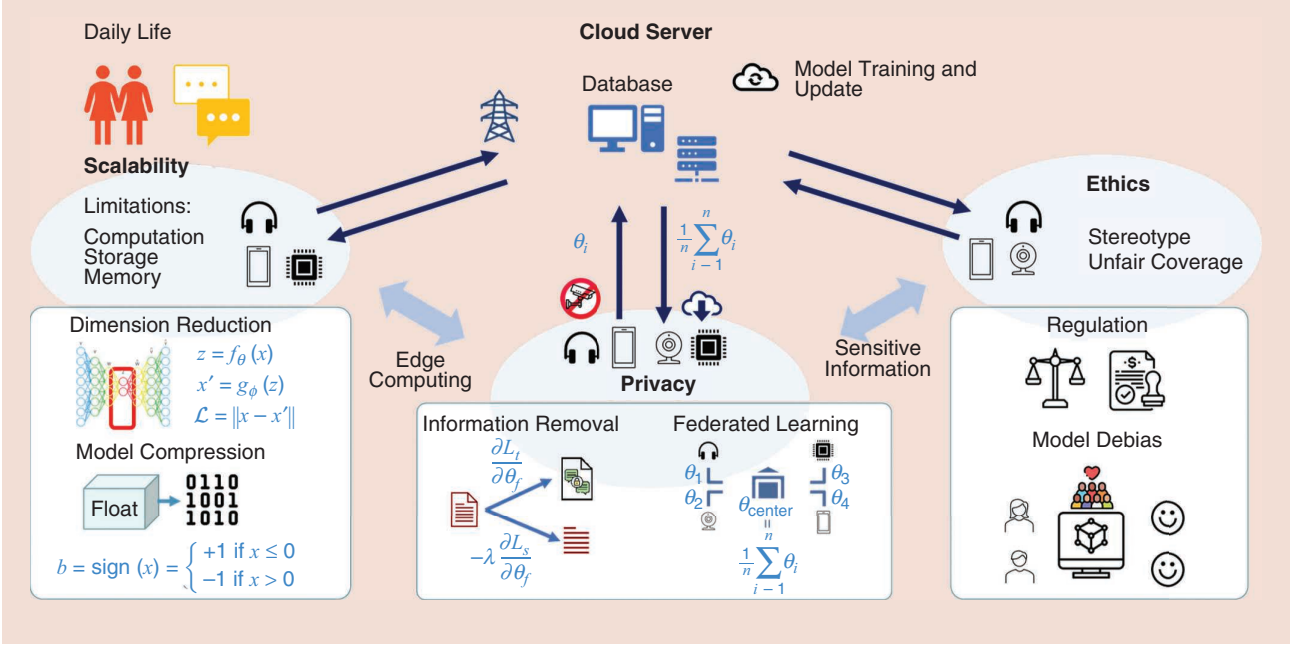


FIGURE 4. An overview of SER usability challenges and current solutions.

(e.g., edge computing), we need to dramatically condense the model and the data representation. Another related problem is how to train or adapt the model in the server when new information is available from the local users. Another usability issue is unintentional bias in SER systems, which can result in important negative societal consequences. Biased results may reflect issues related to imbalanced representation in the data or underlying social stereotypes reflected on the labels. The following sections discuss some of the representation learning approaches intended to mitigate these usability issues for SER systems.

A light model not only requires fewer computational resources but also less memory, which can be an important requirement for edge computing.

Scalability with compact representation

State-of-the-art representation learning frameworks utilize various forms of reconstruction loss-optimization methods to minimize the difference in the response of full, original and squeezed models. The goal is to fit the high-dimensional DNN into a light network that can be used in a local mobile device. Several studies have proposed techniques to simultaneously take care of both the recognition performance and the model complexity [10], [24]. In the following, we discuss two ways of addressing this issue: dimension reduction and model compression.

Reducing the dimension of the feature representation

The methods of dimensional reduction have been used for years since the handcrafted acoustic representations used in the early works in SER tasks led to overexpanded dimension problems. The studies have proposed more targeted feature sets rooted in the externalization of expressive speech. A prominent work is the eGeMAPS feature set [1], a minimal selected acoustic set for paralinguistic tasks based on psychoacoustic knowledge and statistical results. Aside from the manual-selection approaches used to reduce feature dimension, studies have widely used traditional statistical feature-selection algorithms such as principal component analysis (PCA) and forward feature selection as a preprocessing step to reduce the feature dimension.

With DNNs, feature reduction can be effectively implemented with autoencoders, using a bottleneck layer trained utilizing either supervised or unsupervised strategies. The autoencoder aims to reconstruct input features at the output layer, thereby preserving meaningful information in the latent space. An autoencoder can be considered a nonlinear generalization of a PCA. Advanced networks cast the problem as a latent distribution modeling problem and solve it with generative models, such as variational and adversarial autoencoders, and GANs. These networks add different loss functions to constrain the latent layer to follow a given distribution. The studies have reported that these low-dimensional latent representations can not only reduce memory consumption but also improve the robustness by mitigating the overfitting problem [24]. An important observation is that there is a tradeoff between model compactness and system performance. An open challenge is to balance this tradeoff achieving a useful system with reasonable performance.

Model compression

A straightforward approach for reducing model complexity is by squeezing a complex model into a light architecture with a small number of parameters. Training methods such as the teacher-student model can be very effective, where the objective for the light model (i.e., student) is to mirror the representation response of the full model (i.e., teacher). A light model not only requires fewer computational resources but also less memory, which can be an important requirement for edge computing. Another way to compress the model is to focus on the architecture itself by pruning the number of deep network nodes or layers to fundamentally reduce the memory usage. Specifically, several studies have designed low-rank layers, 1×1 convolutional filters, or pooling mechanisms to shrink the node numbers in deep network layers. These models often achieve a comparable performance to the full model.

An alternative technique used to reduce the memory requirement is to quantize the network weights, compressing the size of every neural network node unit. Zhao et al. [10] proposed to binarize the weights of the SER model, successfully achieving a prominent compression rate. The method uses the sign function to perform the binarization. The idea is to minimize the l_2 loss computed between the float-value weight matrix and the binarized target matrix with a scaling factor. Take a convolutional layer as an example: The weights W and inputs I_s are processed with a binarized layer with $H = sign(W)$ and $B = sign(I_s)$. The optimized objective function is expressed as

$$\alpha^*, \beta^* = \underset{\alpha,\beta}{\operatorname{argmin}} \|\mathbf{I}_\mathbf{s}^\mathbf{T}\mathbf{W} - \alpha\beta\mathbf{H}^\mathbf{T}\mathbf{B}\|. \tag{12}$$

We can regard $\alpha^*\mathbf{H}$ and $\beta^*\mathbf{B}$ as the binarized approximation of $\mathbf{I_s}$ and $\mathbf{W}$, respectively. After backpropagation with the recognition loss, we obtain the optimized rescaling factors α and β. We represent the original model with these binarized parameters, which significantly reduce each value into only 1 b. The experimental results using the binarized convolutional recurrent neural network achieved only a 1% loss in accuracy in two SER databases, with a model that was 26 times smaller in its memory requirement.

Privacy-aware speech representation

Privacy concerns have rapidly emerged with the growing integration of SER into everyday life. Speech representation is known to contain sensitive personal attributes. From acoustic features, it is possible to infer personal information such as identity, gender, and age. Developing methods to preserve user privacy is an important usability goal. The studies have focused on two major directions: 1) securing communication between cloud service and edge devices, where the inference happens in the cloud and the edge device only needs to upload the feature and wait for the prediction output from the cloud service; and 2) training methods without the collection of

sensitive information from edge-based devices, where the inference directly happens at the edge device. In the following, we discuss these two scenarios.

Cloud-edge service (inference on the cloud service)

In deploying SER as a cloud service, a user's private information is exposed to risk during communication between the edge and the cloud. In this scenario, studies have proposed representation learning procedures to mask sensitive attributes. The approaches are often based on adversarial training with a gradient reversal layer and regularized optimization. These methods can be used to handle a single or a few sensitive attributes, where the model is built with parallel paths, one for the main SER task and another for the sensitive attribute that we want to mask (identity, gender, age, or even the location). The model essentially erases the sensitive attributes using the reverse gradient layer [25] or flexibly aligns the privacy attribute in a specific order through a layered dropout mechanism [11] when deriving the representation while maintaining high accuracy for the main SER task. If the loss function for the main task is $\mathcal{L}_t$, the loss function for the sensitive attributes task is $\mathcal{L}_s$, the current parameters of the network is θ_f, and the gradient that is backpropagated (Δw_{total}) is computed in (13). The parameter λ weighs the loss of the gradient reversal layer to control the tradeoff between privacy preservation and the main task performance:

$$\Delta w_{\text{total}} = \frac{\partial L_t}{\partial \theta_f} + \left(-\lambda \frac{\partial L_s}{\partial \theta_f}\right). \tag{13}$$

They trained a CycleGAN with two generators to learn a bidirectional mapping between source (S) and target (T) corpora with the goal of generating synthetic source domain samples that are target aware.

An alternative procedure is to use regularized optimization, where we can introduce a loss to guide the model to remove sensitive information from the user. For example, we can preserve privacy-related attributes by exploiting contrastive losses. Arora and Chaspari [26] proposed a training strategy using the Siamese network, which creates uniform random pairing for multiplicative perturbation of the data. The approach repeatedly applies the Gombertz function, a nonlinear transformation that can prevent the inversion of the sensitive information (i.e., the speaker information in this article) and limits the growth of the input space. The use of the contrastive loss maximizes the emotion-related discriminative distance and effectively reduces the growth rate of the input sensitive attribute information by a repeated Gombertz function in the learned representation.

On-edge service (inference on the edge device)

In the scenario of edge-based applications, the goal is to evaluate the SER model on the edge device without transmitting raw data. This on-device training strategy directly prevents private information leakage during transmission. This paradigm is referred to as *federated learning* (*FL*) [27] and enables the system to locally derive a representation, sending only model information from those edge devices that are aggregated on the cloud. During inference, the network parameters are distributed back to the edge devices so that the local models can perform as well as the equivalent models trained with all the data.

We assume that we have n local users $\{E_1, E_2, \ldots, E_n\}$, locally training machine learning models with their own data $\{D_1, D_2, \ldots, D_n\}$. Conventionally, the most common method is to train a final model M_{all} by using the union of all the data $\mathcal{D} \in \{D_1 \cup D_2 \cup D_3 \cup \ldots \cup D_n\}$. However, studies have developed FL strategies to train a model without sharing the data, addressing privacy issues. The ultimate goal is to have a performance similar to the performance of a model trained with all the data. Each edge device learns a lightweight representation, locally trained using data collected on the edge. The cloud model needs only either the estimated gradients or the parameters of the edge models, without the need of the original data or the learned representation. Equations (14) and (15) show the averaged FL variables, where Δw_i stands for the gradient of the model i, and θ_i stands for the parameters of the model i. The final cloud model is updated with these aggregated gradients or parameters. The edge devices received the updated model during inference:

$$\theta_{\text{center}} = \frac{1}{n}\sum_{i=1}^{n} \theta_i \tag{14}$$

$$\Delta w_{\text{center}} = \frac{1}{n}\sum_{i=1}^{n} \Delta w_i. \tag{15}$$

There are three commonly used types of FL methods: horizontal FL (HFL), vertical FL, and federated transfer learning. The selection of these algorithms depends on the target scenario. Here we discuss HFL, the most common FL technique, in more detail. HFL corresponds to when the tasks and input features used across edge devices are both the same, but the set of users are different (e.g., devices A and B both predict emotional categories through the use of acoustic features, but each device belongs to different users). It is often implemented with the averaging integration approach, where the parameters of the model in the cloud are updated according to the average encrypted gradients from the edge devices [28].

FL provides a privacy-preserving strategy to locally build a robust and strong model. FL has the advantage of facilitating the integration of edge user's data without privacy infringement (see "A Federated Learning Framework for SER"). By aggregating the parameters or gradients, this method can still achieve a performance similar to conventional methods trained with all the data without the risk of data leakage because the local data are never shared with the cloud. These advantages are appealing for SER, given the applications that are relevant to this area (e.g., health care). Recently, Latif et al. [28] has demonstrated that FL is an effective approach in SER tasks. A limitation of using FL is the increase of the computation consumption imposed on the edge devices. This perspective also

A Federated Learning Framework for SER

We provide an example on how to implement a federated learning solution for SER. We focus on horizontal federated learning. Figure S2 illustrates the process, where E_i represents the device of user i. We assume that each user is using the SER service on the device. The following four steps of this iterative strategy are the distributed center model, edge-model optimization, aggregation, and center model update:

1) *Step 1*: The first step is the distributed center model, where each device directly clones the parameter from the cloud service model. Accordingly, the edge models are identical at the beginning.
2) *Step 2*: The second step is edge-model optimization, where the edge models predict the emotion output of each utterance. The edge models compute the cross-entropy loss and directly optimize the local models. Note that each model is different at this stage, as dictated by the local data used in each device.
3) *Step 3*: The third step is aggregation, where the cloud service averages all the parameters from the edge models optimized with local data [(14) and (15)].
4) *Step 4*: The fourth step is center model update, where the SER model in the cloud is updated with the aggregated model parameters.

By repeating these four steps, the center model is optimized without getting data from edge devices, properly preventing the disclosure of private information.

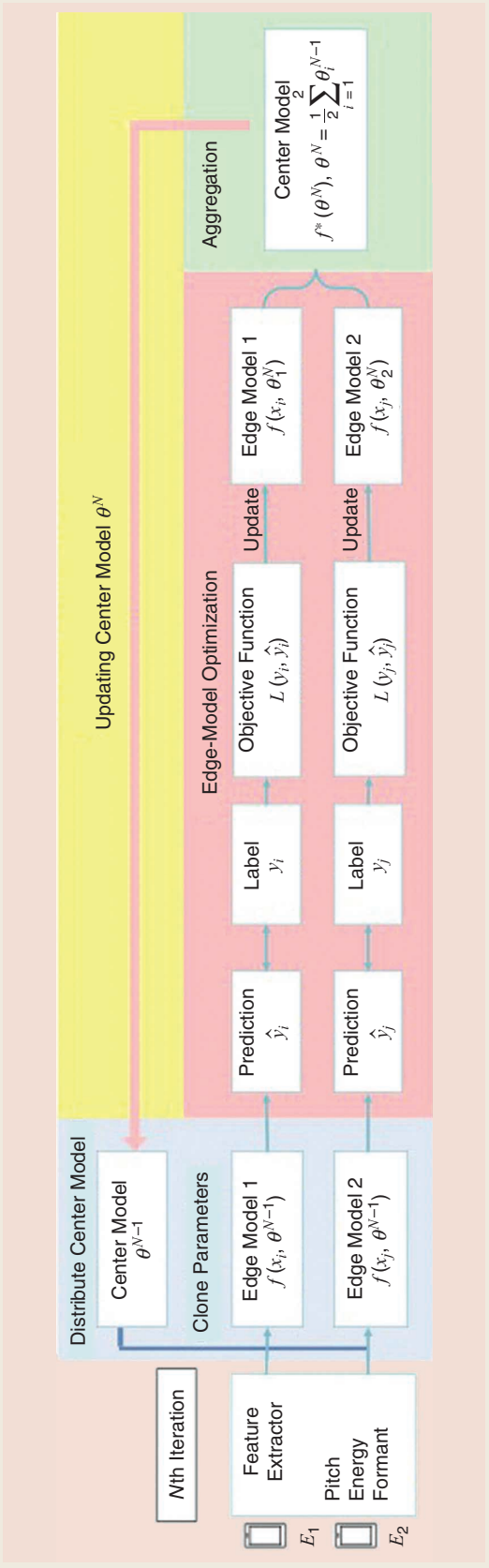


FIGURE S2. The flowchart of a federated learning algorithm example. E_n is the edge device of user n, then f means the local model distributed from the center model, and θ, x is the model parameters and features, respectively. Further, L is the loss function, and in this example, we assume it is a cross-entropy loss function.

requires more memory. These limitations will be mitigated as more powerful portable devices are introduced into the market.

Ethical considerations in building speech representation

The ethical considerations that have to be addressed during SER deployment is another important usability issue. Although there is still an active discussion about the definition of privacy, the data protection scheme for service providers, and the modification and creation of new regulations [29], we can make actionable changes in building representation learning algorithms for SER to attenuate potential ethical issues.

Reducing bias in the SER models

The data used to train a model can be affected by unintentional bias, which will be reflected on the models. The bias appears due to poor representation of underrepresented groups in the data or by social stereotypes reflected on the labels collected with perceptual evaluations. An interesting and immediate method to attenuate bias in the models is by building representation learning styles that intentionally compensate for known unbalanced representation. Domain adaptation is an appealing approach, where the auxiliary task is the recognition of an attribute that we aim our model to correct for bias. For example, Gorrostieta et al. [30] integrated the concept of "equality of odds" to define the fairness of the model, which means that the distribution of sensitive predictions should be equally distributed. The study formulated the protected variable (e.g., gender or age) as an adversary task and trains a main model, which considers the adversarial loss of the protected attributes. The study showed that the proposed adversary representation learning technique was able to explicitly reduce the unwanted bias that exists in a given data set. These tactics address the fairness in the learned speech representation space when modeling the affective speech signal. This is a critical problem as emotion recognition systems have become a key component in the decision-making processes that impact our lives.

Conclusions

The growing body of SER research into using deep representation learning approaches has enabled the possibility of the wide deployment of speech solutions across domains, where we expect the rapid, into-life adoption of SER technology. SER can be readily plugged into a variety of human-centered and service-oriented industries, such as finance, entertainment, sales, marketing, health care, and education, where spoken interaction plays a major role. This article provided a comprehensive summary on the three major affective speech signal modeling challenges that must be addressed to deploy successful SER solutions: robustness, generalization, and usability. This article presented effective deep representation learning architectures that are suitable to address these issues. The deep representation learning methods covered in this article provide appealing solutions to build robust SER systems against unwanted signal and natural human perceptual variations. The solutions improve, in a principled way, the generalization of speech representations that are agnostic to contexts, settings, and domains. The solutions consider usability constraints during the learning of speech representation to improve scalability and trustworthiness, while deploying SER technology. The complexity involved in realizing the value of SER in our everyday lives requires continuous scientific and technical endeavors in modeling, and representing emotional speech signals. Affective speech processing formulations, with carefully designed DNNs to address these key challenges, will undoubtedly provide a core module to deploy SER algorithms built in the laboratory into ubiquitous SER services in the market.

Acknowledgments

The work presented in this article was funded by the National Science Foundation under grants CNS-2016719 and CAREER IIS-1453781, and the Ministry of Science and Technology (Taiwan) under grants 109—2634-F-007—012 and 110—2634-F-007—012.

Authors

Chi-Chun Lee (cclee@ee.nthu.edu.tw) received his Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, California, USA, in 2012. He is an associate professor in the Electrical Engineering Department of the National Tsing Hua University, Hsinchu City, 300044, Taiwan, where he leads the Behavioral Informatics and Interaction Computational Lab. He is a recipient of the Foundation of Outstanding Scholar's Young Innovator Award (2020), Chinese Institute of Electrical Engineering's Outstanding Young Electrical Engineer Award (2020), and Ministry of Science and Technology Taiwan Futuretek Breakthrough Award (2018 and 2019). His research interests are in speech and language, affective multimedia, health analytic, and behavior computing. He is a member of the International Speech and Communication Association and ACM. He is a Senior Member of IEEE.

Kusha Sridhar (kusha.sridhar@utdallas.edu) received his M.S. degree in electrical engineering from the University of Southern California, Los Angeles, California, USA, in 2017. He is currently pursuing his Ph.D. degree in electrical engineering at the University of Texas at Dallas, Dallas, Texas, 75080, USA. His research interests include areas related to affective computing, focusing on emotion recognition from speech, machine learning, and speech signal processing. He is a Student Member of IEEE.

Jeng-Lin Li (cllee@gapp.nthu.edu.tw) received his B.S. degree in the electrical engineering at National Tsing Hua University, Hsinchu City, 300044, Taiwan, in 2016, where he is currently pursuing Ph.D. degree. He was awarded the NTHU Principal Outstanding Student Scholarship (2017–2020), Garmin Scholarship (2018), Yahoo Scholarship (2019), and Novatek Ph.D. Scholarship (2020). His research interests include behavior signal processing, multimodal emotion recognition, and health analytics. He is a student member of the IEEE Signal Processing Society and a member of the International Speech and Communication Association. He is a Student Member of IEEE.

Wei-Cheng Lin (wei-cheng.lin@utdallas.edu) received his M.S. degree in electrical engineering from the National Tsing Hua University, Taiwan, in 2016. He is currently pursuing his Ph.D. degree in electrical and computer engineering at the University of Texas at Dallas, Dallas, Texas, 75080, USA. His research interests include affective computing, deep learning, and multimodal/speech signal processing. He is a student member of the IEEE Signal Processing Society and the International Speech and Communication Association. He is a Student Member of IEEE.

Bo-Hao Su (borrissu@gapp.nthu.edu.tw) is currently pursuing his Ph.D. degree and received his B.S. degree in the Department of Electrical Engineering at National Tsing Hua University, Hsinchu City, 300044, Taiwan, in 2017. He was awarded the NTHU Principal Outstanding Student Scholarship (2018–2022) and the INTERSPEECH 2018 Sub Challenge Championship. His research interests include behavioral signal processing, cross corpus speech emotion recognition, and machine learning. He is a student member of the International Speech and Communication Association. He is a Student Member of IEEE.

Carlos Busso (busso@utdallas.edu) is a professor in the Electrical Engineering Department of the University of Texas at Dallas, Dallas, Texas, 75080, USA. He received his Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, California, USA. He is a recipient of an NSF CAREER Award. In 2014, he received the ICMI Ten-Year Technical Impact Award. He was a corecipient of both the Hewlett-Packard Best Paper Award at the 2011 IEEE International Conference on Multimedia and Expo and the Best Paper Award at the Association for the Advancement of Affective Computing International Conference on Affective Computing and Intelligent Interaction 2017. His research interest include human-centered multimodal machine intelligence and applications; and the broad areas of affective computing, nonverbal behaviors for conversational agents, and machine learning methods for multimodal processing.

References

[1] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps et al., "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affective Comput.*, vol. 7, no. 2, pp. 190–202, Apr./June 2016. doi: 10.1109/TAFFC.2015.2457417.

[2] H.-C. Chou and C.-C. Lee, "Every rating matters: oint learning of subjective labels and individual annotators for speech emotion classification," in *ICASSP 2019— Proc. IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP)*, 2019, pp. 5886–5890.

[3] G.-Y. Chao, Y.-S. Lin, C.-M. Chang, and C.-C. Lee, "Enforcing semantic consistency for cross corpus valence regression from speech using adversarial discrepancy learning," in *Proc. Interspeech*, 2019, pp. 1681–1685.

[4] S. Alisamir and F. Ringeval, "Into the unknown: Towards self-supervised learning of speech representations for affective computing," *IEEE Signal Process. Mag.*, early access, Feb. 15, 2021.

[5] A. Triantafyllopoulos, G. Keren, J. Wagner, I. Steiner, and B. W. Schuller, "Towards robust speech emotion recognition using deep residual networks for speech enhancement," in *Proc. Interspeech*, 2019, pp. 1691–1695.

[6] K. Sridhar and C. Busso, "Modeling uncertainty in predicting emotional attributes from spontaneous speech," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP 2020)*, Barcelona, Spain, May 2020, pp. 8384–8388.

[7] K. Sridhar and C. Busso, "Ensemble of students taught by probabilistic teachers to improve speech emotion recognition," in *Proc. Interspeech 2020*, Shanghai, China, Oct. 2020, pp. 516–520.

[8] W. Zhang and P. Song, "Transfer sparse discriminant subspace learning for cross-corpus speech emotion recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 307–318, Nov. 2020. doi: 10.1109/TASLP.2019.2955252.

[9] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 12, pp. 2423–2435, Dec. 2018. doi: 10.1109/TASLP.2018.2867099.

[10] H. Zhao, Y. Xiao, J. Han, and Z. Zhang, "Compact convolutional recurrent neural networks via binarization for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP)*, 2019, pp. 6690–6694.

[11] Y.-L. Huang, B.-H. Su, Y.-W. P. Hong, and C.-C. Lee, "An attribute-aligned strategy for learning speech representation," 2021, arXiv:2106.02810.

[12] S. Pascual, M. Ravanelli, J. Serrà, A. Bonafonte, and Y. Bengio, "Learning problem-agnostic speech representations from multiple self-supervised tasks," in *Proc. Interspeech*, ,pp. 161–165. doi: 10.21437/Interspeech.2019-2605.

[13] D. Jiang, W. Li, M. Cao, R. Zhang, W. Zou, K. Han, and X. Li, "Speech simclr: Combining contrastive and reconstruction objective for self-supervised speech representation learning," 2020, arXiv:2010.13991.

[14] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, Mar. 2016, pp. 5200–5204.

[15] H. Li, M. Tu, J. Huang, S. Narayanan, and P. Georgiou, "Speaker-invariant affective representation learning via adversarial training," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process. (ICASSP)*, 2020, pp. 7144–7148.

[16] W.-C. Lin and C. Busso, "Chunk-level speech emotion recognition: A general framework of sequence-to-one dynamic temporal modeling," *IEEE Trans. Affective Comput.*, early access, May 26, 2021.

[17] K. Sridhar and C. Busso, "Speech emotion recognition with a reject option," in *Proc. Interspeech 2019*, Graz, Austria, Sept. 2019, pp. 3272–3276. doi: 10.21437/Interspeech.2019-1842.

[18] B.-H. Su and C.-C. Lee, "A conditional cycle emotion GAN for cross corpus speech emotion recognition," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, 2021, pp. 351–357.

[19] F. Bao, M. Neumann, and N. T. Vu, "Cyclegan-based emotion style transfer as data augmentation for speech emotion recognition," in *Proc. Interspeech*, 2019, pp. 2828–2832.

[20] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2697–2709, Sept. 2020. doi: 10.1109/TASLP.2020.3023632.

[21] C.-C. Lu, J.-L. Li, and C.-C. Lee, "Learning an arousal-valence speech front-end network using media data in-the-wild for emotion recognition," in *Proc. Audio/Visual Emotion Challenge Workshop*, 2018, pp. 99–105.

[22] S. Sahu, R. Gupta, and C. Espy-Wilson, "Modeling feature representations for affective speech using generative adversarial networks," *IEEE Trans. Affective Comput.*, early access, June 2020. doi: 10.1109/TAFFC.2020.2998118.

[23] H. Zhao, Y. Xiao, and Z. Zhang, "Robust semisupervised generative adversarial networks for speech emotion recognition via distribution smoothness," *IEEE Access*, vol. 8, pp. 106,889–106,900, June 2020. doi: 10.1109/ACCESS.2020.3000751.

[24] G. Paraskevopoulos, E. Tzinis, N. Ellinas, T. Giannakopoulos, and A. Potamianos, "Unsupervised low-rank representations for speech emotion recognition," in *Proc. Interspeech*, 2019, pp. 939–943.

[25] M. Jaiswal and E. M. Provost, "Privacy enhanced multimodal neural representations for emotion recognition," in *AAAI*, vol. 34, no. 05, 2020. pp. 7985–7993, doi: 10.1609/aaai.v34i05.6307.

[26] P. Arora and T. Chaspari, "Exploring Siamese neural network architectures for preserving speaker identity in speech emotion classification," in *Proc. 4th Int. Workshop on Multimodal Anal. Enabling Artif. Agents Human–Machine Interaction*, 2018, pp. 15–18.

[27] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Trans. Intell. Syst. Technol. (TIST)*, vol. 10, no. 2, pp. 1–19, 2019. doi: 10.1145/3298981.

[28] S. Latif, S. Khalifa, R. Rana, and R. Jurdak, "Federated learning for speech emotion recognition applications," in *Proc. 19th ACM/IEEE Int. Conf. Inf. Process. Sensor Netw. (IPSN)*, 2020, pp. 341–342. doi: 10.1109/IPSN48710.2020.00-16.

[29] A. Batliner, S. Hantke, and B. W. Schuller, "Ethics and good practice in computational paralinguistics," *IEEE Trans. Affective Comput.*, early access, Sept. 1, 2020.

[30] C. Gorrostieta, R. Lotfian, K. Taylor, R. Brutti, and J. Kane, "Gender de-biasing in speech emotion recognition," in *Proc. Interspeech*, 2019, pp. 2823–2827.

SP