frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 28 March 2023
DOI 10.3389/fcomp.2023.1039261

OPEN ACCESS

EDITED BY
Kun Qian,
Beijing Institute of Technology, China

REVIEWED BY
Shahin Amiriparian,
University of Augsburg, Germany
Yu Tsao,
Academia Sinica, Taiwan

*CORRESPONDENCE
Shruti Kshirsagar
✉ shruti.kshirsagar@inrs.ca

SPECIALTY SECTION
This article was submitted to Human-Media Interaction, a section of the journal Frontiers in Computer Science

RECEIVED 08 September 2022
ACCEPTED 13 March 2023
PUBLISHED 28 March 2023

CITATION
Kshirsagar S, Pendyala A and Falk TH (2023) Task-specific speech enhancement and data augmentation for improved multimodal emotion recognition under noisy conditions. *Front. Comput. Sci.* 5:1039261. doi: 10.3389/fcomp.2023.1039261

Task-specific speech enhancement and data augmentation for improved multimodal emotion recognition under noisy conditions

Shruti Kshirsagar[1]*, Anurag Pendyala[2] and Tiago H. Falk[1]

[1]Institut National de la Recherche Scientifique, University of Quebec, Montréal, QC, Canada, [2]International Institute of Information Technology, Bangalore, India

Automatic emotion recognition (AER) systems are burgeoning and systems based on either audio, video, text, or physiological signals have emerged. Multimodal systems, in turn, have shown to improve overall AER accuracy and to also provide some robustness against artifacts and missing data. Collecting multiple signal modalities, however, can be very intrusive, time consuming, and expensive. Recent advances in deep learning based speech-to-text and natural language processing systems, however, have enabled the development of reliable multimodal systems based on speech *and* text while only requiring the collection of audio data. Audio data, however, is extremely sensitive to environmental disturbances, such as additive noise, thus faces some challenges when deployed "in the wild." To overcome this issue, speech enhancement algorithms have been deployed *at the input signal level* to improve testing accuracy in noisy conditions. Speech enhancement algorithms can come in different flavors and can be optimized for different tasks (e.g., for human perception vs. machine performance). Data augmentation, in turn, has also been deployed *at the model level* during training time to improve accuracy in noisy testing conditions. In this paper, we explore the combination of task-specific speech enhancement and data augmentation as a strategy to improve overall multimodal emotion recognition in noisy conditions. We show that AER accuracy under noisy conditions can be improved to levels close to those seen in clean conditions. When compared against a system without speech enhancement or data augmentation, an increase in AER accuracy of 40% was seen in a cross-corpus test, thus showing promising results for "in the wild" AER.

KEYWORDS

multimodal emotion recognition, BERT based text features, modulation spectrum features, data augmentation, speech enhancement, context-awareness

1. Introduction

Affective human-machine interfaces are burgeoning as they provide more natural interactions between the human and the machine (Zeng, 2007). Automated emotion recognition (AER) systems have seen applications across numerous domains, from marketing, smart cities and vehicles, to call centers and patient monitoring, to name a few. In fact, the COVID-19 pandemic has resulted in a global mental health crisis that will have long-term consequences to society, economy, and healthcare systems (Xiong et al., 2020). Being able to detect changes in affective states in a timely and reliable manner can allow

individuals and organizations to put in place interventions to prevent, for example, burnout and depression (Patrick and Lavery, 2007).

AER systems can rely on a wide range of modalities, including speech, text, gestures/posture, and physiological responses (e.g., *via* changes in heart/breathing rates). For so-called "in the wild" applications, multimodal systems are preferred in order to compensate for certain confounds and to improve overall AER accuracy by providing the system with some redundancy and complementary information not available with unimodal systems (Naumann et al., 2009; Parent et al., 2019). Multimodal systems, however, can be very time consuming to implement, costly to run, and potentially intrusive to the users (e.g., requiring on-body sensors with physiological data collection) and their privacy (Sebe et al., 2005). Notwithstanding, with audio inputs, one may be able to devise a multimodal speech-and-text system with the use of an advanced speech-to-text system, thus relying on a single input modality. As such, text and speech have emerged as two popular AER modalities.

Recent advances in deep learning architectures, such as transformers (Vaswani et al., 2017), have redefined the performance envelope of existing AER systems. In fact, most state-of-the-art systems today rely on deep neural network architectures in some way. For example, for text-based systems, self attention and dynamic max pooling has been proposed by Yang et al. (2019). The widely-used Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2018), in turn, has been used to detect cyber abuse in English and Hindi texts (Malte and Ratadiya, 2019). The work by Lee and Tashev (2015) and Kratzwald et al. (2018), in turn, relies on recurrent neural networks (RNN) to better consider long-range contextual effects and to better model the uncertainty around emotional labels. For speech-based AER systems, in turn, mel-spectral features combined with a convolutional neural networks (CNNs) have been extensively explored, specially with self-attention mechanisms to extract emotionally-informative time segments (e.g., Chen et al., 2021). Long-short term memory networks (LSTM) have also been extremely popular (e.g., Haytham et al., 2017; Tripathi et al., 2018; Zhao et al., 2019) and end-to-end solutions have also been explored (Tzirakis et al., 2017).

As mentioned previously, one major advantage of the audio modality is that recent advances in automated speech-to-text conversion have allowed for multimodal speech-and-text-based systems to emerge while requiring the collection of just one signal modality (Chuang and Wu, 2004). Text and speech have been shown to be very useful modalities for multimodal AER systems (Patamia et al., 2021). In this regard, attention-based bidirectional LSTM models (Li et al., 2020), bi-directional RNNs (Poria et al., 2017), transformer-based models (Siriwardhana et al., 2020), multi-level multi-head fusion attention mechanisms (Ho et al., 2020), graph-based CNNs (Zhang et al., 2019), gated-recurrent units (Poria et al., 2019), early and late fusion strategies (Jin et al., 2015), and cross-modal attention (Sangwan et al., 2019) have been explored as strategies to optimally combine information from the two modalities.

One major disadvantage of speech-based systems (either uni- or multi-modal), however, is their sensitivity to environmental factors, such as additive and convolutional noise (e.g., room reverberation). These factors can be detrimental to AER systems (Patamia et al., 2021; Maithri et al., 2022). Commonly, speech enhancement algorithms are applied at the input level stage to minimize environmental factors for in-the-wild speech applications. Enhancement methods can range from more classical methods, such as spectral subtraction and Wiener filtering (Cauchi et al., 2015; Braun et al., 2016), to more recent deep neural network (DNN) based ones (e.g., Parveen and Green, 2004; Lu et al., 2013; Pascual et al., 2017; Zhao et al., 2018). The use of speech enhancement for AER in-the-wild has shown some benefits (e.g., Avila et al., 2021).

Speech enhancement methods can have two very different purposes. If aimed at improving intelligibility/ quality, for example, human perception becomes the main driving factor and quality/intelligibility improvements are typically used as a figure of merit (e.g., Fu et al., 2021). However, if enhancement is used to improve downstream speech recognition applications then other machine-driven outcome measures, such as word error rate improvements, are more appropriate. As such, depending on the final task, the enhancement procedure can be very different. The work by Bagchi et al. (2018), for example, showed that mimic loss-based enhancement was optimal for automatic speech recognition (ASR) downstream tasks. Having this said, it is hypothesized that for multimodal speech-and-text AER systems the use of two different enhancement procedures will be useful, with a quality-driven one used for the speech branch (mimicking how humans perceive emotions from speech) and a machine-driven one for the speech-to-text branch. We will test this hypothesis herein.

Lastly, with deep learning based approaches showing the latest state-of-the-art results, data augmentation has emerged as a useful technique to make systems more robust to in-the-wild distortions at the model training stage (e.g., Hannun et al., 2014). With data augmentation, the training set is increased multi-fold by applying certain transformations to the available training signals, including time-reversal, time-frequency masking, pitch alterations, background noise addition and reverberation corruption, to name a few. For AER specifically, the work by Etienne et al. (2018) showed that vocal track length perturbations served as a useful data augmentation strategy. In this paper, we further explore the advantages that data augmentation can provide, in addition to speech enhancement, for multimodal in-the-wild AER.

The remainder of this paper is organized as follows. Section 2 describes the proposed system. Section 3 describes the experimental setup. Experimental results and a discussion are presented in Section 4 and conclusions in Section 5.

2. Proposed method

Figure 1 depicts the block diagram of the proposed multimodal AER pipeline. In the case of interest here, speech $S(i)$ is assumed to be corrupted by additive background noise $N(i)$, resulting in noisy speech signal $Y(i) = S(i) + N(i)$. With the multimodal AER system, the top branch focuses on extracting emotion-relevant features directly from the speech component, whereas the bottom branch relies on a state-of-the-art automatic speech recognizer (ASR) to generate text from the noisy speech signal. Features are then extracted from the text transcripts. We concatenated Speech

and text features, then these concatenated features are input to a deep neural network for final emotion classification. As noisy speech is known to corrupt AER/ASR performance, here we also include a speech enhancement step, one optimized for speech quality improvement (top branch) and another for ASR. Each sub-block is described in detail in the subsections to follow:

2.1. Speech enhancement

Enhancement and noise suppression has been widely used across many different speech-based applications. In human-to-human communications, the goal of enhancement is to improve the quality of the noisy signal, not only to increase intelligibility, but also to improve paralinguistic characterization that humans do so well, such as emotion recognition. In human-to-machine interaction (e.g., ASR), however, improving quality may not be the ultimate goal, and instead, improvement in downstream system accuracy could be regarded as a better optimization criterion. Here, we explore the use of a quality-optimized enhancement algorithm for the speech branch of the proposed method and an ASR-optimized algorithm for the text generation branch. The two algorithms used are described next:

2.1.1. MetricGAN+: A quality-optimized enhancement method

MetricGAN+ is a recent state-of-the-art deep neural network specifically optimized for quality enhancement of noisy speech and shown to outperform several other enhancement benchmarks (Fu et al., 2019, 2021). In particular, two networks are used. The discriminator's role is to minimize the difference between the predicted quality scores (given by the so-called PESQ, perceptual evaluation of speech quality, rating Rix et al., 2001) and actual PESQ quality scores. PESQ is a standardized International Telecommunications Union full-reference speech quality metric that maps a pair of speech files (a reference and the noisy counterpart) into a final quality rating between 1 (poor) and 5 (excellent). PESQ has been widely used and validated across numerous speech applications.

The generator's role, in turn, is to map a noisy speech signal into its enhanced counterpart. The discriminator and generator models are trained together to enhance the noisy signal in a manner that maximizes the PESQ score of the enhanced signal. MetricGAN+ builds on the original MetricGAN (Fu et al., 2019) *via* two improvements for the discriminator and one for the generator. More specifically, for the discriminator training, along with the enhanced and clean speech signals, the noisy speech was also used to minimize the distance between the discriminator and target objective metrics. The second improvement is that the speech generated from the previous epochs is reused to train the discriminator to avoid the catastrophic forgetting of the discriminator. For the generator, in turn, the learnable sigmoid function was used for mask estimation. The interested reader is referred to Fu et al. (2019, 2021) for more details on the MetricGAN and MetricGAN+ speech enhancement methods.

2.1.2. Mimic loss: An ASR-optimized enhancement method

Spectral mapping-based speech enhancement is an enhancement method specifically optimized for downstream ASR applications (Bagchi et al., 2018). We refer henceforth to this method as 'mimic loss based enhancement' as the model uses mimic loss instead of student-teacher learning, thus the speech enhancer is not jointly trained with a particular acoustic model. We use this enhancement model as it has been shown to be a useful pre-processing method for many ASR systems, thus offers some flexibility on the choice of ASR model to use (Bagchi et al., 2018). The overall system is comprised of two major components: a spectral mapper and a spectral classifier which are trained in three steps.

First, a spectral classifier is trained to predict senone labels from clean speech with a cross-entropy criterion, resulting in a classification loss L_C between predicted and actual senones. The weights of this spectral classifier are then frozen and used in the last step. Second, a spectral mapper is pre-trained to map noisy speech features to clean speech features using a mean squared error (MSE) criterion. This results in a fidelity loss L_F between the denoised features and features from the clean speech counterpart. Bagchi et al. (2018) relied on log-spectral magnitude components extracted over 25ms windows with a 10-ms shift as features and a deep feed-forward neural network for mapping.

Lastly, noisy speech is input to the pre-trained spectral mapper, resulting in a denoised version, which is input to the "frozen" spectral classifier, resulting in a predicted senone. In parallel, the clean speech counterpart is also input to the frozen spectral classifier, resulting in a soft senone label and a mimic loss L_M between the soft senone label and the predicted senone. The spectral mapper is then retrained using joint loss (L_F and L_M), thus allowing the enhancer to emulate the behavior of the classifier under clean conditions while keeping the projection of noisy signal closer to that of the clean signal counterpart. The same hyperparameters described by Bagchi et al. (2018) were used herein. The interested reader is referred to Bagchi et al. (2018) for more details on the mimic loss enhancement method.

2.2. Automatic speech recognition

In order to generate text from speech, a state-of-the-art automatic speech recognizer is needed. Here, wav2vec 2.0, an end-to-end speech recognition system is used (Baevski et al., 2020). A complete description of the method is beyond the scope of this paper, hence only an overview is provided; the interested reader can obtain more details from Baevski et al. (2020). Wav2vec 2.0 relies on the raw speech waveform as input. This 1-dimensional data then passes through a multi-layer 1-d CNN to generate speech representation vectors. Vector quantization is then used on these latent representations to match them to a codebook. Half of the available speech data is masked and the remaining quantized data is fed into a transformer network. By using contrastive loss, the model attempts to predict the masked vectors, thus allowing for

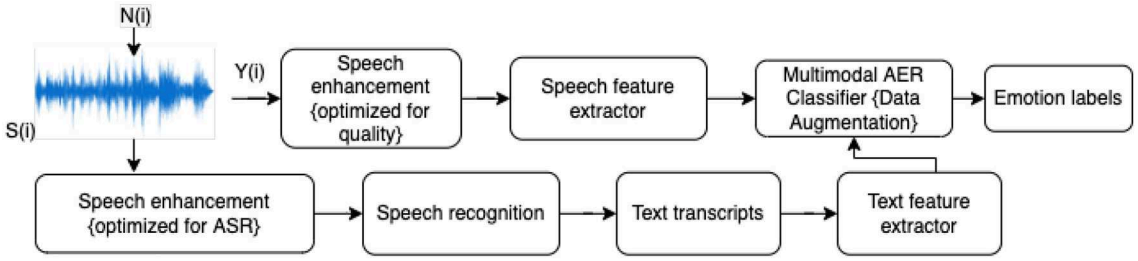

FIGURE 1
Experimental pipeline for AER using audio and text features.

pre-training on unlabeled speech data. The model is then fine-tuned on labeled data for the subsequent down-streaming ASR task.

2.3. Speech feature extractor

Several AER systems have been proposed recently, and they have relied on different speech feature representations. Here, we focus on the three most popular representations, namely: prosodic, eGeMAPS, and modulation spectral features. In particular, prosody features include fundamental frequency (F0), intensity measures, and voicing probabilities, as these have been widely linked to emotions (Banse and Scherer, 1996). Next, the so-called extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) (Eyben et al., 2016), which has been widely used in many recent emotion recognition challenges (e.g., Valstar, 2016; Ringeval et al., 2019; Xue et al., 2019), is also explored and contains a set of 88 acoustic parameters relating to pitch, loudness, unvoiced segments, temporal dynamics, and cepstral features. Lastly, modulation spectral features are explored as they capture second-order periodicities in the speech signal and have been shown to convey emotional information (Wu et al., 2011; Avila et al., 2021). Modulation spectral features (termed MSFs) were extracted using a window size of 256 ms and a frame step of 40 ms. The interested reader is referred to Falk and Chan (2010b) and Avila et al. (2021) for complete details on the computation of this representation.

2.4. Text feature representations

Text has also been used to infer the emotional content of written material and several state-of-the-art methods and techniques exist. Here, we explore three recent methods, namely BERT (Bidirectional Encoder Representations from Transformers), TextCNN, and Bag-of-Words (BoW). A brief overview of each method is given below:

2.4.1. BERT-bidirectional encoder representations from transformers

BERT is based on a transformer network and attention mechanism (Devlin et al., 2018) that also learns contextual relations between words in the text (Tenney et al., 2019). BERT comes in two flavors: BERTBase and BERTLarge. The BERTBase model uses 12 layers of transformers block with a hidden dimension of 768 and 12 self-attention heads; overall, there are approximately 110 million trainable parameters. On the other hand, BERTLarge uses 24 layers of transformers block with a hidden size of 1024 and 16 self-attention heads, resulting in approximately 340 million trainable parameters. Here, we employ the BERTBase model for text feature extraction. The BERT hidden state vector is used as input to the AER system. The interested reader is referred to Devlin et al. (2018) for more details on BERT.

2.4.2. TextCNN

TextCNN is a deep learning model for short text classification tasks and has been used as a baseline model for text classification (Zhang et al., 2018). TextCNN transforms a word into a vector using word embeddings, which are then fed into a convolutional layer, followed by a max-pooling layer, and a fully connected output layer. In our experiment, TexCNN embeddings were extracted using the model described by Poria et al. (2018). We used three convolutional layers with 64 filter and kernel sizes of 3, 4, and 5 respectively in each layer, followed by max-pooling and finally 150 dense layers to extract the final text features. Specifically, with pre-trained 300-dimensional GloVe vectors (Pennington et al., 2014), we first extracted the semantic vector space representation and then fed them to a 1-D-CNN to extract 100-dimensional text features vector.

2.4.3. Bag-of-Words

The bag-of-words (BOW) method is commonly employed in natural language processing (Alston, 1964). The approach is straightforward and flexible and can be used in many ways to extract features from documents. BOW represents the text by describing the occurrence of words within a document. It consists of two parts: a vocabulary of known words and a measure of the presence of these words. It is called a "bag" of words because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not wherein the document. In this method, first, a word histogram is generated within a text document. Next, the frequencies of each word from a dictionary are computed, and finally the resultant vector is fused and used as the

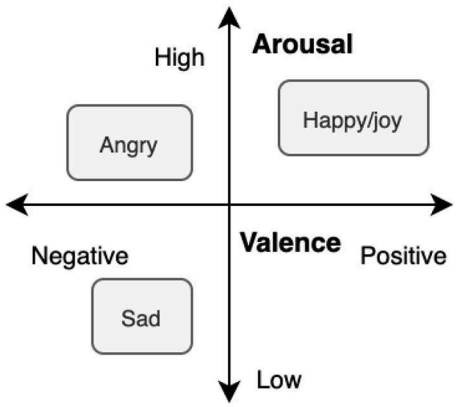

FIGURE 2
Valence-arousal emotional space with the three discrete emotions considered here.

text features. For our experiment, we used CountVectorizer from the sklearn library. A 652-dimensional feature vector was used for each utterance and the unigram model was used to generate the BOW representation.

2.5. Multimodal AER classifier

Here, we rely on a fully connected deep neural network for multimodal emotional recognition. Three dense layers (of dimensions 256, 128, 32) were used, plus a final classification layer. A dropout rate of 0.6 was used, batch normalization was performed after every layer, and class weights of [1, 1.8] were assigned during training. Grid search was performed on the validation set to obtain the optimal hyperparameters. Rmsprop, Adam, and SGD optimizers were explored, and learning rates of 0.01, 0.001, and 0.0001 were tested to find the optimal combination. Once the best parameters were found with the validation set, we reported the best performance on our test data. Experimentation codes are available on github[1]. The network is trained with and without data augmentation in order to explore its effect on in-the-wild AER performance.

3. Experimental setup

In this section, we present the setup used in our experiments.

3.1. Datasets used

The dataset used for experimentation is the Multimodal EmotionLines Dataset (MELD) (Poria et al., 2018). It is a multimodal emotion classification dataset which has been created by extending the EmotionLines dataset (Chen et al., 2018). MELD contains approximately 13,000 utterances from 1,433 dialogues from the TV series 'Friends'. Each statement is annotated with emotion and sentiment labels and encompasses audio, visual, and textual modalities. The MELD dataset contains conversations, where each dialogue has utterances from multiple speakers. EmotionLines was created by crawling the discussions from each episode and then grouping them based on the number of statements in conversation into four groups of utterances. Finally, 250 dialogues were sampled randomly from each group, resulting in the final dataset of 1,000 dialogues. The utterances in each dialogue were annotated with the most appropriate emotion category.

For this purpose, the six universal emotions (joy, sadness, fear, anger, surprise, and disgust) were considered. This annotation list was extended with two additional emotion labels: neutral and non-neutral. Each utterance was annotated by five workers from the Amazon Mechanical Turk platform. A majority voting scheme was applied to select a final emotion label for each utterance. While the MELD dataset has labels for several emotions, here we focus on two specific binary tasks to gauge effects across the valence and arousal dimensions. More specifically, we first focus on two tasks. Task 1 comprises anger vs. sad classification to explore the benefits of the proposed tool for low/high arousal classification (Mower et al., 2010; Metallinou et al., 2012). Task 2, in turn, comprises joy vs. sad classification for positive-valence-high-arousal and negative-valence-low-arousal characterization (Park et al., 2013; Li et al., 2019). Figure 2 depicts the arousal-valence emotional space and the three discrete emotions considered. As such, the MELD dataset was split into three disjoint sets: training, test, and development. These were split as follows:

1. Training: angry (1,109 samples), joy (1,743 samples), and sad (682 samples);
2. Validation: angry (153 samples), joy (163 samples), and sad (111 samples);
3. Testing: angry (345 samples), joy (402 samples), and sad (208 samples).

To test the robustness of the proposed methods to in-the-wild conditions, the MELD dataset is corrupted by multi-talker babble noise, cafeteria noise, and noise recorded inside a commercial airplane at different SNR levels: −10,−20, 0, 5, 10, 15, and 20 dB. The AURORA (Hirsch and Pearce, 2000) and DEMAND noise datasets (Thiemann et al., 2013) are used for this purpose. Note that only a subset of these conditions are used during augmentation, including airport and babble noise and SNR levels of 0, 10, and 20 dB. The remainder are left as unseen conditions during testing.

Next, we utilized the IEMOCAP dataset to show the generalizability of the proposed model. The IEMOCAP dataset has 12 h of audio-visual data from 10 actors where the recordings follow the dialogue between a male and a female actor in both scripted or improvised topics. After the audio-video data was collected, it was divided into small utterances of length between 3 and 15 s, which were then labeled by evaluators. Each utterance was evaluated by 3–4 assessors. The evaluation form contained ten options (neutral, happiness, sadness, anger, surprise, fear, disgust, frustration, excitement, and others). We consider only three: anger, sadness, and happy so as to remain consistent with the previous MELD data experiments and to be able to directly test the models

1 https://github.com/shrutikshirsagar/Speech-enhancement-Audio-Text-ER

trained on the MELD dataset. To this end, the dataset was split into three disjoint sets: training (70%), development (10%) and test (20%). IEMOCAP contains utterances from the disjoined speakers in training and testing. More specifically, we used sessions 2,3,4, and 5 for training and session 1 for testing purposes. These were split as follows:

1. Training: angry (772 samples), happy (416 samples), and sad (758 samples);
2. Validation: angry (111 samples), happy (60 samples), and sad (110 samples);
3. Testing: angry (220 samples), happy (119 samples), and sad (216 samples).

Finally, we also utilized the spontaneous "in the wild" English-language Emoti-W database (Dhall et al., 2017). It was made available through the 2017 Emotion Recognition in the Wild Challenge. Some level of background noise was present in the recording as Emoti-w is "in the wild" dataset. The labels for the EMoti-W challenge dataset were created from the closed captions available in movies and TV series. Complete details about the Emoti-W dataset can be found in Dhall et al. (2017). The data is available in a sampling frequency of 48 kHz; videos are available in MPEG-2 format with 25 frames per second. Emotion labels are available for seven emotion categories: anger, disgust, fear, happiness, neutral, sadness, and surprise were available in this dataset. We consider only three: anger, sadness, and happy/joy so as to remain consistent with the previous MELD data experiments and to be able to directly test the models trained on the MELD dataset as mentioned earlier. Again, we use only the labeled training and development subsets in our experiments. Training, testing and validation split of Emoti-W dataset are as follows:

1. Training: angry (110 samples), happy (120 samples), and sad (90 samples);
2. Validation: angry (11 samples), happy (24 samples), and sad (17 samples);
3. Testing: angry (64 samples), happy (60 samples), and sad (61 samples).

3.2. Benchmark systems

To gauge the benefits of the proposed method, two benchmark systems are used, namely BcLSTM and DialogueRNN and results are reported in Table 1 for task 1 and 2. BcLSTM is bi-directional RNN proposed by Poria et al. (2017). It is comprised of a two-step hierarchical training process. First, it extracts embeddings from each modality. For text, GloVe embeddings (Pennington et al., 2014) were used as input to a CNN-LSTM model to extract contextual representations for each utterance. For audio, Openmsile based features (Eyben, 2013) were input to an LSTM model to obtain audio representations for each audio utterance. Next, contextual representations from the audio and text modalities are fed to the BcLSTM model for emotion classification.

DialogueRNN, in turn, employs three stages of gated recurrent units (GRU) to model emotional context in conversations (Poria et al., 2019). The spoken utterances are fed into two GRUs: global and party GRU, to update the context and speaker state, respectively. In each turn, the party GRU updates its state based on i) the utterance spoken, ii) the speaker's previous state, and iii) the conversational context summarized by the global GRU through an attention mechanism. Finally, the updated speaker state is fed into the emotion GRU, which models the emotional information for classification. The attention mechanism is used on top of the emotion GRU to leverage contextual utterances by different speakers at various distances. Lastly, our proposed system comprises a feedforward DNN model and a 768- dimensional BERT(base) text feature vector fused (at the feature level) with a 311-dimensional vector comprised of eGEMAPs and MSF features.

TABLE 1 Benchmark system performance for the two AER tasks based on the MELD dataset.

Model	Task 1		Task 2	
	F1-score	BA	F1-score	BA
bcLSTM	0.70	0.72	0.82	0.83
DialogueRNN	0.72	0.72	0.84	0.85
Proposed system	0.74	0.73	0.87	0.87

3.3. Figures-of-merit

Balanced accuracy and F1-score are used as figures of merit to assess the performance of the proposed emotion classifier. In summary, precision shows us how many positive samples classified by the model are actually positive, i.e.,

$$Precision = \frac{TP}{TP + FP}, \tag{1}$$

Where TP corresponds to true positives and FP to false positives. Recall, in turn, calculates how many of the true positives are captured by the model. This is also called true positive rate or sensitivity and given by

$$Recall = \frac{TP}{TP + FN}, \tag{2}$$

Where FN corresponds to false negatives. Moreover, F1-score represents the harmonic mean of precision and recall and is useful in binary tasks where classes are unbalanced and is given by:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}. \tag{3}$$

Lastly, balanced accuracy is given as the arithmetic mean of sensitivity (true positive rate or recall) and specificity (true negative rate) which, in turn, is given by:

$$Specificity = \frac{NP}{TN + FP}, \tag{4}$$

Where TN corresponds to true negatives. As such, balance accuracy (BA) is given as:

$$BA = \frac{Sensitivity + Specificity}{2}. \tag{5}$$

The interested reader is referred to Powers (2020) for more details on these classical performance metrics.

3.4. Quality scores

To gauge the improvements in quality and intelligibility of the enhancement algorithms, two objective speech quality measures are used, namely, PESQ and the short-term objective intelligibility (STOI) (Taal et al., 2011). While PESQ estimates the perceived speech quality on a 5-point mean opinion score scale ranging from bad to excellent, STOI measures the intelligibility of the signal on a 0–1 scale, with higher values suggesting greater intelligibility. Both methods are termed "intrusive" as they require access to the enhanced and a reference signal. More details on the PESQ and STOI measurement algorithms can be found in Rix et al. (2001).

4. Experimental results and discussion

In this section, we present and discuss the obtained experimental results.

4.1. Ablation study 1

In this first ablation experiment, we wish to explore the optimal set of text and speech features to include in the final system. We consider speech and text modalities separately in this study. We start with clean speech to find the best feature per modality and, subsequently, test the robustness of such set under unseen noisy conditions. In this study, babble and airport noises are considered. In both cases, the emotion classifier is trained on clean speech only. Table 2 shows the performance obtained for each modality individually for task 1. In the table, the feature termed 'fusion' corresponds to the fusion of MSF and eGeMAPS features.

As can be seen, for clean speech conditions and text-only AER, BERT-based text features resulted in the best performance across all metrics, hence corroborating previous reports (Yang et al., 2019; Stappen et al., 2021; Yang and Cui, 2021). As such, only BERT features are explored in the unseen noisy conditions. Babble noise is shown to degrade overall performance more severely than airport noise. Overall, BERT based features under 0 dB noise conditions are shown to achieve accuracy inline with that achieved by textCNN features under clean conditions, thus further suggesting improved robustness of the BERT text features. Given this finding, the final proposed system shown in Figure 1 will rely on BERT based text features.

As for speech features, under clean conditions eGeMAPS showed the highest overall performance of the three tested feature sets, thus corroborating findings by Eyben et al. (2013). Further gains could be seen with the fused feature set, however, thus suggesting the complementarity of spectral and modulation spectral features. As such, only the fused feature set is explored in the noisy mismatch condition. Moreover, similar to the text features, at low SNR levels, babble noise degraded performance more drastically compared to airport noise. Overall, the achieved performance with text-based features only was higher than what was achieved with audio features alone, thus corroborating the results reported by Patamia et al. (2021).

4.2. Ablation study 2

This second ablation study is an oracle experiment in which one modality in the multimodal system is kept clean and the other is corrupted by noise at varying levels and types. This study will allow us to gauge which modality is most sensitive to environmental factors and would benefit the most from speech enhancement. In all cases, the emotion classifier is trained on clean speech only. Table 3 show the performance obtained for Task 1 and Task 2, respectively.

As can be seen, the fusion of speech and text features in the clean condition (first row in the tables) showed improvements relative to each modality alone (i.e., Table 2) by as much as 2% for text and 7% for audio in terms of F1 score for Task 1. Furthermore, using noisy speech to generate "noisy" text resulted in more severe performance degradations for both Tasks, thus suggesting that more powerful machine-tuned enhancement algorithms may be useful for in-the-wild applications to assure the highest possible quality for text generation. Overall, on average, over the two types of noise, a drop of 32, 24, and 21% in F1 score was observed at 0, 10, 20 dB SNR levels relative to clean conditions, respectively, for Task 1. On the other hand, corrupting only the speech content had a less pronounced effect. Overall, on average, over the two types of noise, a drop of 16%, 13%, and 9% in F1 score was observed at 0, 10, 20 dB SNR levels over clean conditions for Task 1, respectively.

For Task 2, similar findings were observed. Overall, on average, over the two types of noise, a drop of 65, 33, and 28% in F1 score has been observed at 0, 10, and 20 dB SNR levels relative to clean conditions, respectively, when only text was corrupted. The drops in accuracy when the audio was corrupted were of 41, 27, and 25%, respectively. These findings corroborate those by Kessous et al. (2010) and Patamia et al. (2021) who showed that text modality achieved higher performance than audio in clean conditions. The drops in accuracy, however, under noisy conditions motivate the need for strategies to improve accuracy in the wild, as in the proposed system.

4.3. Ablation study 3

This third ablation study is an oracle experiment in which we wanted to test the hypothesis if we need two separate enhancement for improving ASR accuracy. As mentioned earlier, we used quality-(MetricGAN+) and ASR-optimized (mimic loss) enhancement algorithms for the speech and text branches shown in the proposed model in Figure 1. This study will allow us to gauge which combination of speech enhancement is better suited for this task. In all cases, the emotion classifier is trained on clean speech only. Table 4 show the performance obtained for Task 1 and Task 2.

As can be seen, for both Task 1 and Task 2, the best combination comprised the use of a quality-optimized enhancement algorithm for the top speech branch and an ASR-optimized (mimic loss) method for the bottom text branch. This combination resulted in the best accuracy for very extreme conditions (i.e., 0 dB SNR levels) and emphasizes the need for task-specific enhancement algorithms for AER.

TABLE 2 Ablation study 1: Performance comparison of different features for each individual modality.

Noise type	Feature	F1-score	BA	Feature	F1-score	BA
	Text			Audio		
Clean	BERT	0.72	0.76	Prosodic	0.62	0.61
Clean	TextCNN	0.56	0.54	eGEMAPS	0.69	0.67
Clean	BoW	0.62	0.59	MSF	0.66	0.67
Clean				Fusion	0.69	0.71
Airport (0 dB)	BERT	0.54	0.52	Fusion	0.51	0.51
Airport (10 dB)	BERT	0.60	0.57	Fusion	0.53	0.50
Airport (20 dB)	BERT	0.62	0.59	Fusion	0.55	0.52
Babble (0 dB)	BERT	0.58	0.56	Fusion	0.51	0.52
Babble (1 dB)	BERT	0.61	0.58	Fusion	0.52	0.51
Babble (20 dB)	BERT	0.61	0.58	Fusion	0.52	0.51

Feature termed "fusion" corresponds to the fusion of eGeMAPS and MSFs.

TABLE 3 Ablation study 2: Performance comparison of multimodal oracle system for Task 1 and Task 2.

		Task 1		Task 2	
Audio	Text	F1-score	BA	F1-score	BA
Clean	Clean	0.74	0.73	0.87	0.87
Clean	Airport (0 dB)	0.57	0.58	0.55	0.53
Clean	Airport (10 dB)	0.61	0.58	0.67	0.62
Clean	Airport (20 dB)	0.62	0.59	0.68	0.63
Clean	Babble (0 dB)	0.58	0.59	0.50	0.51
Clean	Babble (10 dB)	0.61	0.58	0.63	0.58
Clean	Babble (20 dB)	0.61	0.58	0.67	0.62
Airport (0 dB)	Clean	0.65	0.62	0.60	0.66
Airport (10 dB)	Clean	0.65	0.63	0.68	0.68
Airport (20 dB)	Clean	0.68	0.65	0.70	0.68
Babble (0 dB)	Clean	0.62	0.60	0.63	0.67
Babble (10 dB)	Clean	0.65	0.62	0.68	0.67
Babble (20 dB)	Clean	0.68	0.65	0.69	0.67

TABLE 4 Ablation study 3: Performance comparison of enhancement system for Task 1 and Task 2.

			Task 1		Task 2	
Noise	Enhancement-1	Enhancement-2	F1-score	BA	F1-score	BA
Airport (0 dB)	MetricGAN+	MetricGAN+	0.60	0.60	0.53	0.50
	MetricGAN+	Mimic-loss	0.65	0.64	0.56	0.51
	Mimic-loss	MetricGAN+	0.61	0.59	0.54	0.52
	Mimic-loss	Mimic-loss	0.61	0.60	0.55	0.52
Babble (0 dB)	MetricGAN+	MetricGAN+	0.59	0.59	0.56	0.51
	MetricGAN+	Mimic-loss	0.62	0.59	0.57	0.51
	Mimic-loss	MetricGAN+	0.60	0.60	0.56	0.51
	Mimic-loss	Mimic-loss	0.61	0.61	0.56	0.52

10.3389/fcomp.2023.1039261

4.4. Overall system performance

This last study explores the performance of the proposed system described in Figure 1, combining speech enhancement optimized for each branch (speech and text), as well as data augmentation to provide robustness at the model training level. Data augmentation methods are useful to solve imbalanced data problems. It also helps the model to learn the complex distribution of the data and helps prevent overfitting. The work by Hu et al. (2018) showed that adding noisy versions of the clean speech data to the training set improved speech recognition accuracy in mismatched noisy conditions. Therefore, in this work, we utilized the same strategy. Table 5 show the obtained results in rows labeled 'Data augmentation only' for Task 1 and Task 2. As can be seen, data augmentation alone already improved AER results, thus corroborating findings by Trinh et al. (2021); Neumann and Vu (2021), and Kshirsagar and Falk (2022a,b).

Next, we gauge the benefits of using speech enhancement alone. As before, AER models are trained solely on clean speech. During run time, we pre-process the test data with the MetricGAN+ algorithm for the speech branch and the mimic loss enhancer for the text branch, as described in Section 2. Table 5, show the obtained results in rows labeled 'Enhancement only'. As can be seen, applying speech enhancement improves overall performance relative to the noisy conditions, but the final results are still below what was achieved in clean conditions, as well as what was achieved with data augmentation. The gains observed were typically more substantial at low SNR values, thus corroborating results by Triantafyllopoulos et al. (2019).

In an attempt to better understand the reason behind the poor AER performance with speech enhancement alone, Figure 3 depicts an average modulation spectrogram, from top to bottom, for clean, noisy (airport at 0 dB SNR), MetricGAN+, and mimic-loss enhanced speech for angry (left) and sad (right) emotions, respectively. Modulation spectrograms are a frequency-frequency representation where the y-axis depicts acoustic frequency and the x-axis modulation frequency. From the clean plot, we can see the typical speech modulation spectral representation with most modulation energy lying below 16 Hz (Falk and Chan, 2010a) and a slowing of the amplitude modulations with the sad emotion (Wu et al., 2011). Noise, in turn, is shown to affect the modulation spectrogram by smearing the energy across higher acoustic and modulation frequencies, as suggested by Falk et al. (2010). The enhancement algorithms, however, are not capable of completely removing these environmental artifacts and seem to be introducing other types of distortions that can make the AER task more challenging. Combined, these factors result in the reduced gains reported in the Tables. This was in fact confirmed by listening to the outputs of the MetricGAN+ enhancement algorithm. We have also presented the PESQ, and STOI scores in Table 6. This verifies the significance of having task-specific enhancement for improving the AER performance in noisy conditions.

Finally, we test the combined effects of speech enhancement and data augmentation, as in the proposed system, to gauge the benefits of noise robustness applied at both the input and model levels, respectively. For Task 1, gains (relative to

TABLE 5 Performance comparison of the proposed method in different noisy test conditions for Task 1 and Task 2.

	Task 1		Task 2	
Signal	F1-score	BA	F1-score	BA
Clean	0.74	0.73	0.87	0.87
Noisy-Airport (−20 dB)	0.49	0.49	0.43	0.53
Data augmentation only	0.51	0.49	0.53	0.52
Enhancement only	0.56	0.52	0.51	0.48
Proposed	0.56	0.54	0.52	0.50
Noisy-Airport (−10 dB)	0.53	0.46	0.44	0.57
Data augmentation only	0.53	0.52	0.52	0.50
Enhancement only	0.57	0.52	0.54	0.51
Proposed	0.59	0.54	0.57	0.56
Noisy-Airport (0 dB)	0.57	0.55	0.50	0.50
Data augmentation only	0.67	0.68	0.61	0.61
Enhancement only	0.65	0.64	0.56	0.51
Proposed	0.65	0.63	0.62	0.59
Noisy-Airport (10 dB)	0.59	0.57	0.55	0.51
Data augmentation only	0.69	0.70	0.66	0.66
Enhancement only	0.68	0.65	0.61	0.55
Proposed	0.71	0.69	0.65	0.62
Noisy-Airport (20 dB)	0.60	0.58	0.60	0.55
Data augmentation only	0.69	0.68	0.67	0.66
Enhancement only	0.67	0.65	0.62	0.56
Proposed	0.71	0.69	0.67	0.65
Noisy-Babble(−20 dB)	0.52	0.49	0.49	0.49
Data augmentation only	0.52	0.49	0.54	0.54
Enhancement only	0.57	0.52	0.54	0.51
Proposed	0.58	0.58	0.56	0.51
Noisy-Babble (−10 dB)	0.54	0.51	0.52	0.51
Data augmentation only	0.56	0.51	0.55	0.52
Enhancement only	0.59	0.54	0.54	0.51
Proposed	0.56	0.52	0.59	0.54
Noisy-Babble (0 dB)	0.59	0.57	0.54	0.51
Data augmentation only	0.66	0.66	0.58	0.59
Enhancement only	0.62	0.59	0.57	0.51
Proposed	0.64	0.61	0.61	0.58
Noisy-Babble (10 dB)	0.60	0.58	0.58	0.54
Data augmentation only	0.72	0.71	0.63	0.62
Enhancement only	0.68	0.66	0.61	0.55
Proposed	0.70	0.68	0.66	0.64
Noisy-Babble (20 dB)	0.61	0.58	0.61	0.56
Data augmentation only	0.74	0.72	0.67	0.67
Enhancement only	0.70	0.67	0.66	0.60
Proposed	0.70	0.69	0.67	0.64

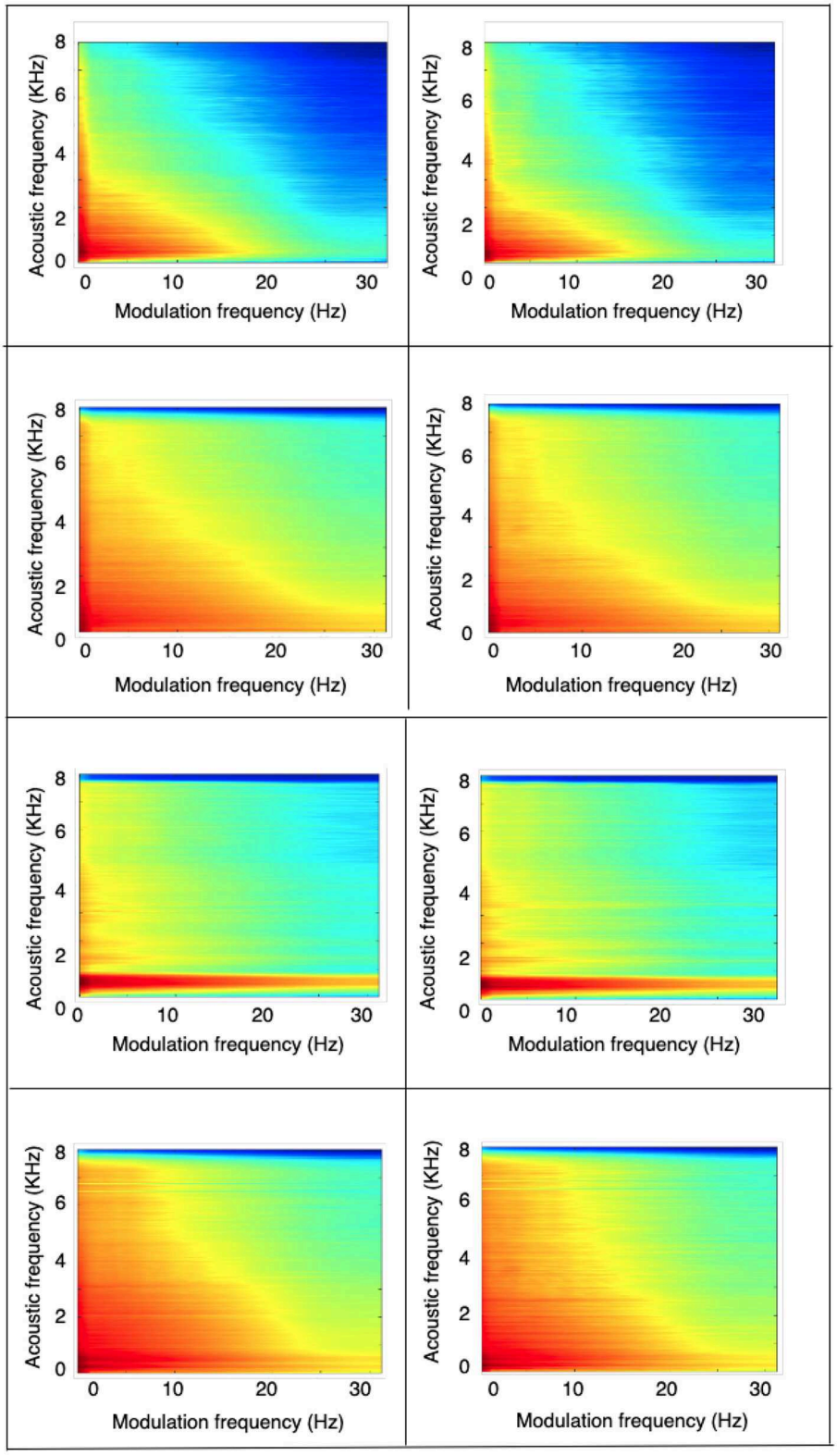

FIGURE 3
Modulation spectrogram for different conditions, from **(top–bottom)**: clean, (airport) noisy at 0 dB, MetriGAN+, and mimic-loss enhanced speech. **(Left)** plots correspond to angry and **(right)** plots to sad emotion.

TABLE 6 Performance comparison of PESQ and STOI score.

Signal	PESQ	STOI
Noisy-Airport (−20 dB)	1.107	0.219
MetricGAN+	1.130	0.312
MimicLoss	1.054	0.244
Noisy-Airport (−10 dB)	1.094	0.367
MetricGAN+	1.132	0.412
MimicLoss	1.112	0.368
Noisy-Airport (0 dB)	1.102	0.620
MetricGAN+	1.225	0.657
MimicLoss	1.112	0.627
airport (10 dB)	1.583	0.791
MetricGAN+	1.899	0.812
MimicLoss	1.622	0.800
Noisy-Airport (20 dB)	2.800	0.885
MetricGAN+	2.979	0.895
MimicLoss	2.894	0.886
Noisy-abble (−20 dB)	1.100	0.188
MetricGAN+	1.154	0.254
MimicLoss	1.038	0.229
Noisy-Babble (−10 dB)	1.103	0.342
MetricGAN+	1.151	0.363
MimicLoss	1.138	0.356
Noisy-Babble (0 dB)	1.139	0.576
MetricGAN+	1.229	0.639
MimicLoss	1.180	0.591
Noisy-Babble (10 dB)	1.577	0.764
MetricGAN+	1.939	0.789
MimicLoss	1.605	0.768
Noisy-Babble (20 dB)	2.792	0.871
MetricGAN+	2.968	0.880
MimicLoss	2.799	0.871

using each strategy individually) were seen for the airport noise condition at higher and lower SNR conditions. In fact, with data augmentation alone, accuracy inline with what was achieved with clean speech was obtained. For Task 2, in turn, the proposed model showed improvements over the other methods for almost all tested conditions in terms of F1 score, thus showing the importance of the proposed method to classify between opposing emotions in extremely noisy scenaerios; in the case here, joy vs. sad. Notwithstanding, for Task 2 a gap of 23% remained between the best achieved performance and the clean speech accuracy. Furthermore, we also tested the generalization ability of the proposed system using unseen Cafeteria noise type and unseen SNR levels such as 5 dB and 15 dB. As can be seen in Table 7 the model was able to generalize across mismatched

TABLE 7 Performance comparison of the proposed method in unseen noise and SNR levels for Task 1.

	Task 1		Task 2	
Signal	F1-score	BA	F1-score	BA
Noisy - Cafeteria (5dB)	0.58	0.57	0.52	0.47
Data augmentation only	0.63	0.60	0.67	0.64
Enhancement only	0.66	0.64	0.64	0.59
Proposed	0.68	0.65	0.67	0.65
Noisy - Cafeteria (15dB)	0.61	0.58	0.54	0.48
Data augmentation only	0.65	0.62	0.68	0.63
Enhancement only	0.69	0.66	0.67	0.61
Proposed	0.70	0.69	0.71	0.70

TABLE 8 Cross-corpus performance on unseen IEMOCAP and Emoti-W datasets for Tasks 1 and 2.

		Task 1		Task 2	
Experiment	Dataset	F1-score	BA	F1-score	BA
1	IEMOCAP	0.94	0.94	0.85	0.85
2		0.49	0.55	0.50	0.60
3		0.64	0.69	0.72	0.70
1	Emoti-W	0.67	0.66	0.61	0.62
2		0.46	0.52	0.48	0.53
3		0.58	0.60	0.56	0.56

noise types and noise levels with significant performance gain with the proposed methodology. For comparison purposes, the state-of-the-art DialogueRNN system achieved an F1 score of 0.59 and 0.55 for Task 1 and Task 2, respectively, when corrupted with airport noise at 0 dB. The proposed system, in turn, was able to outperform this benchmark by 10 and 12%, respectively. Overall, the obtained results suggest that data augmentation combined with speech enhancement can be a viable alternative for robust in-the-wild automatic multimodal emotion recognition while requiring access to only one signal modality: audio.

4.5. Generalizability of proposed method

To test the generalizability of the proposed method, six additional experiments have been conducted on IEMOCAP and Emoti-W datasets. First, we retrain the proposed AER model using the IEMOCAP training dataset partition and test it on the IEMOCAP test set to obtain an upper bound on what can be achieved on this particular dataset. Next, to gauge the advantages brought by the proposed system, we retrain the AER system shown in Figure 1 but without the enhancement and data augmentation steps. Training was done on the MELD dataset and the model was then tested on the unseen IEMOCAP test data and the unseen Emoti-W testset. This gives us an idea of how challenging the cross-corpus task is when the proposed innovations are not present

and should give us a lower bound on what could be achieved cross-corpus. Finally, we tested the full proposed method trained on the MELD dataset and tested on the unseen IEMOCAP and Emoti-W test data. Experimental results are reported in Table 8. As can be seen, cross-corpus testing is an extremely challenging task where performance accuracy can drop to chance levels if strategies are not put in place. The proposed innovations, on the other hand, provides some robustness, and gains of 30% and 44% on IEMOCAP and 26% and 17% on Emoti-W could be seen with the proposed system for Tasks 1 and 2, respectively, over a system without task-specific speech enhancement and data augmentation. The gaps to the upper bound obtained with Experiment 1 suggest that there is still room for improvement and emotion-aware enhancement and/or alternate data augmentation strategies may still be needed.

5. Conclusions

This paper has explored the use of task-specific speech enhancement combined with data augmentation to provide robustness to unseen test conditions for multimodal emotion recognition systems. Experiments conducted on the MELD dataset show the importance of BERT for text feature extraction and a fused eGEMAPS-modulation spectral set for audio features. The importance of data augmentation at the training stage and of task-specific speech enhancement at the testing stage are shown on two binary speech emotion classification tasks. Lastly, cross-corpus experiments showed the proposed innovations resulting in 40% gains relative to an AER system without enhancement/augmentation. While the obtained results suggest that task-specific enhancement, combined with data augmentation are important steps toward reliable "in the wild" emotion recognition, speech enhancement algorithms may still be suboptimal and may be removing important emotion information. As such, future work should explore the development of emotion-aware enhancement algorithms that can trade-off noise suppression and emotion recognition accuracy.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

SK and TF: conceptualization. TF: funding acquisition and writing-review and editing. SK: methodology, software, validation, visualization, and writing-original draft. Part of this initial work was done during AP's MITAC Globalink internship, he worked on experimentation in Tables 1–3. Therefore, he is considered a second author in this paper. All authors contributed to the article and approved the submitted version.

Funding

The authors acknowledge funding from the Natural Sciences and Engineering Research Council of Canada and MITACS via their Globalink program.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alston, W. P. (1964). Philosophy of Language. *J. Philos. Logic*. 2, 458–508.

Avila, A., Akhtar, Z., Santos, J., O'Shaughnessy, D., and Falk, T. (2021). Feature pooling of modulation spectrum features for improved speech emotion recognition in the wild. *IEEE Trans. Affect. Comput*. 12, 177–188. doi: 10.1109/TAFFC.2018.2858255

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: a framework for self-supervised learning of speech representations. *arXiv preprint* arXiv:2006.11477. doi: 10.48550/arXiv.2006.11477

Bagchi, D., Plantinga, P., Stiff, A., and Fosler-Lussier, E. (2018). "Spectral feature mapping with mimic loss for robust speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Calgary, AB: IEEE), 5609–5613.

Banse, R., and Scherer, K. (1996). Acoustic profiles in vocal emotion expression. *J. Pers. Soc. Psychol*. 70, 614. doi: 10.1037/0022-3514.70.3.614

Braun, S., Schwartz, B., Gannot, S., and Habets, E. A. (2016). "Late reverberation psd estimation for single-channel dereverberation using relative convolutive transfer functions," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)* (Xi'an: IEEE), 1–5.

Cauchi, B., Kodrasi, I., Rehr, R., Gerlach, S., Jukić, A., Gerkmann, T., et al. (2015). Combination of mvdr beamforming and single-channel spectral processing for enhancing noisy and reverberant speech. *EURASIP J. Adv. Signal Process*. 2015, 61. doi: 10.1186/s13634-015-0242-x

Chen, S., Zhang, M., Yang, X., Zhao, Z., Zou, T., and Sun, X. (2021). The impact of attention mechanisms on speech emotion recognition. *Sensors* 21, 7530. doi: 10.3390/s21227530

Chen, S.-Y., Hsu, C.-C., Kuo, C.-C., Ku, L.-W., et al. (2018). Emotionlines: an emotion corpus of multi-party conversations. *arXiv preprint* arXiv:1802.08379. doi: 10.48550/arXiv.1802.08379

Chuang, Z.-J., and Wu, C.-H. (2004). "Multi-modal emotion recognition from speech and text," in *International Journal of Computational Linguistics Chinese Language Processing, Volume* 9, *Number 2, August 2004: Special Issue on New Trends of*

Speech and Language Processing, 45–62. Available online at: https://aclanthology.org/O04-3004.pdf

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* arXiv:1810.04805.

Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J., and Gedeon, T. (2017). "From individual to group-level emotion recognition: emotiw 5.0," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (Glasgow).

Etienne, C., Fidanza, G., Petrovskii, A., Devillers, L., and Schmauch, B. (2018). Cnn+ lstm architecture for speech emotion recognition with data augmentation. *arXiv preprint* arXiv:1802.05630. doi: 10.21437/SMM.2018-5

Eyben, F. (2013). "Recent developments in openSMILE, the Munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia* (Barcelona).

Eyben, F., Scherer, K., Schuller, B., and Sundberg, J., André, E., Busso, C., et al. (2016). The geneva minimalistic acoustic parameter set GeMAPS for voice research and affective computing. *IEEE Trans. Affect. Comput.* 7, 417. doi: 10.1109/TAFFC.2015.2457417

Eyben, F., Weninger, F., and Schuller, B. (2013). "Affect recognition in real-life acoustic conditions-a new perspective on feature selection," in *Proceedings 14th INTERSPEECH* (Lyon).

Falk, T., and Chan, W. (2010a). Modulation spectral features for robust far-field speaker identification. *IEEE Trans Audio Speech Lang Process.* 18, 90–100. doi: 10.1109/TASL.2009.2023679

Falk, T., and Chan, W. (2010b). Temporal dynamics for blind measurement of room acoustical parameters. *IEEE Trans. Instrum Meas.* 59, 24697. doi: 10.1109/TIM.2009.2024697

Falk, T., Zheng, C., and Chan, W.-Y. (2010). A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Trans. Audio Speech Lang. Process.* 18, 1766–1774. doi: 10.1109/TASL.2010.2052247

Fu, S.-W., Liao, C.-F., Tsao, Y., and Lin, S.-D. (2019). "Metricgan: generative adversarial networks based black-box metric scores optimization for speech enhancement," in *International Conference on Machine Learning (PMLR)* (Long Beach, CA), 2031–2041.

Fu, S.-W., Yu, C., Hsieh, T.-A., Plantinga, P., Ravanelli, M., Lu, X., et al. (2021). Metricgan+: an improved version of metricgan for speech enhancement. *arXiv preprint* arXiv:2104.03538. doi: 10.21437/Interspeech.2021-599

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., et al. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv:1412.5567*. doi: 10.48550/arXiv.1412.5567

Haytham, F., Lech, M., and Lawrence, C. (2017). Evaluating deep learning architectures for speech emotion recognition. *Neural Netw.* 92, 60–68. doi: 10.1016/j.neunet.2017.02.013

Hirsch, H., and Pearce, D. (2000). "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)* (Paris).

Ho, N.-H., Yang, H.-J., Kim, S.-H., and Lee, G. (2020). Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network. *IEEE Access* 8, 61672–61686. doi: 10.1109/ACCESS.2020.2984368

Hu, H., Tan, T., and Qian, Y. (2018). "Generative adversarial networks based data augmentation for noise robust speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Calgary, AB: IEEE), 5044–5048.

Jin, Q., Li, C., Chen, S., and Wu, H. (2015). "Speech emotion recognition with acoustic and lexical features," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (South Brisbane, QLD: IEEE), 4749–4753.

Kessous, L., Castellano, G., and Caridakis, G. (2010). Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *J. Multimodal User Interfaces* 3, 33–48. doi: 10.1007/s12193-009-0025-5

Kratzwald, B., Ilić, S., Kraus, M., Feuerriegel, S., and Prendinger, H. (2018). Deep learning for affective computing: text-based emotion recognition in decision support. *Decis. Support. Syst.* 115, 24–35. doi: 10.1016/j.dss.2018.09.002

Kshirsagar, S., and Falk, T. H. (2022a). Cross-language speech emotion recognition using bag-of-word representations, domain adaptation, and data augmentation. *Sensors* 22, 6445. doi: 10.3390/s22176445

Kshirsagar, S. R., and Falk, T. H. (2022b). Quality-aware bag of modulation spectrum features for robust speech emotion recognition. *IEEE Tran. Affect. Comput.* 13, 1892–1905. doi: 10.1109/TAFFC.2022.3188223

Lee, J., and Tashev, I. (2015). High-level feature representation using recurrent neural network for speech emotion recognition. *Interspeech* 2015, 336. doi: 10.21437/Interspeech.2015-336

Li, C., Bao, Z., Li, L., and Zhao, Z. (2020). Exploring temporal representations by leveraging attention-based bidirectional lstm-rnns for multi-modal emotion recognition. *Inf. Process. Manag.* 57, 102185. doi: 10.1016/j.ipm.2019.102185

Li, T.-M., Chao, H.-C., and Zhang, J. (2019). Emotion classification based on brain wave: a survey. *Hum. Centric Comput. Inf. Sci.* 9, 1–17. doi: 10.1186/s13673-019-0201-x

Lu, X., Tsao, Y., Matsuda, S., and Hori, C. (2013). "Speech enhancement based on deep denoising autoencoder," in *Interspeech* (Lyon), 436–440.

Maithri, M., Raghavendra, U., Gudigar, A., Samanth, J., Barua, P. D., Murugappan, M., et al. (2022). Automated emotion recognition: current trends and future perspectives. *Comput. Methods Programs Biomed.* 2022, 106646. doi: 10.1016/j.cmpb.2022.106646

Malte, A., and Ratadiya, P. (2019). "Multilingual cyber abuse detection using advanced transformer architecture," in *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)* (Kochi: IEEE), 784–789.

Metallinou, A., Wollmer, M., Katsamanis, A., Eyben, F., Schuller, B., and Narayanan, S. (2012). Context-sensitive learning for enhanced audiovisual emotion classification. *IEEE Trans. Affect. Comput.* 3, 184–198. doi: 10.1109/T-AFFC.2011.40

Mower, E., Matarić, M. J., and Narayanan, S. (2010). A framework for automatic human emotion classification using emotion profiles. *IEEE Trans. Audio Speech Lang. Process.* 19, 1057–1070. doi: 10.1109/TASL.2010.2076804

Naumann, A. B., Wechsung, I., and Hurtienne, J. (2009). "Multimodal interaction: Intuitive, robust, and preferred?" in *IFIP Conference on Human-Computer Interaction* (Uppsala: Springer), 93–96.

Neumann, M., and Vu, N. T. (2021). "Investigations on audiovisual emotion recognition in noisy conditions," in *2021 IEEE Spoken Language Technology Workshop (SLT)* (Shenzhen: IEEE), 358–364.

Parent, M., Tiwari, A., Albuquerque, I., Gagnon, J.-F., Lafond, D., Tremblay, S., et al. (2019). "A multimodal approach to improve the robustness of physiological stress prediction during physical activity," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)* (Bari: IEEE), 4131–4136.

Park, M. W., Kim, C. J., Hwang, M., and Lee, E. C. (2013). "Individual emotion classification between happiness and sadness by analyzing photoplethysmography and skin temperature," in *2013 Fourth World Congress on Software Engineering* (Hong Kong: IEEE), 190–194.

Parveen, S., and Green, P. (2004). "Speech enhancement with missing data techniques using recurrent neural networks," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1* (Montreal, QC: IEEE), I-733.

Pascual, S., Bonafonte, A., and Serra, J. (2017). Segan: Speech enhancement generative adversarial network. *arXiv:1703.09452*. doi: 10.21437/Interspeech.2017-1428

Patamia, R. A., Jin, W., Acheampong, K. N., Sarpong, K., and Tenagyei, E. K. (2021). "Transformer based multimodal speech emotion recognition with improved neural networks," in *2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML)* (Chengdu: IEEE), 195–203.

Patrick, K., and Lavery, J. F. (2007). Burnout in nursing. *Aust. J. Adv. Nurs.* 24, 43.

Pennington, J., Socher, R., and Manning, C. D. (2014). "Glove: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Doha), 1532–1543.

Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., and Morency, L.-P. (2017). "Context-dependent sentiment analysis in user-generated videos," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (volume 1: Long papers)* (Vancouver, BC), 873–883.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2018). Meld: a multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint* arXiv:1810.02508. doi: 10.18653/v1/P19-1050

Poria, S., Majumder, N., Mihalcea, R., and Hovy, E. (2019). Emotion recognition in conversation: research challenges, datasets, and recent advances. *IEEE Access* 7, 100943–100953. doi: 10.1109/ACCESS.2019.2929050

Powers, D. M. (2020). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint* arXiv:2010.16061. doi: 10.48550/arXiv.2010.16061

Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Tavabi, L., et al. (2019). "AVEC 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop* (Nice).

Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. (2001). "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221), volume 2* (Salt Lake City, UT: IEEE), 749–752.

Sangwan, S., Chauhan, D. S., Akhtar, M., Ekbal, A., Bhattacharyya, P., et al. (2019). "Multi-task gated contextual cross-modal attention framework for sentiment and emotion analysis," in *International Conference on Neural Information Processing* (Vancouver, BC: Springer), 662–669.

Sebe, N., Cohen, I., and Huang, T. S. (2005). Multimodal emotion recognition. *Handbook Pattern Recogn. Comput. Vis.* 4, 387–419. doi: 10.1142/9789812775320_0021

Siriwardhana, S., Kaluarachchi, T., Billinghurst, M., and Nanayakkara, S. (2020). Multimodal emotion recognition with transformer-based self supervised feature fusion. *IEEE Access* 8, 176274–176285. doi: 10.1109/ACCESS.2020.3026823

Stappen, L., Baird, A., Christ, L., Schumann, L., Sertolli, B., Messner, E.-M., et al. (2021). The muse 2021 multimodal sentiment analysis challenge: sentiment, emotion, physiological-emotion, and stress. *arXiv preprint* arXiv:2104.07123. doi: 10.1145/3475957.3484450

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2011). An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Trans. Audio Speech Lang. Process.* 19, 2125–2136. doi: 10.1109/TASL.2011.2114881

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., et al. (2019). What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint* arXiv:1905.06316. doi: 10.48550/arXiv.1905.06316

Thiemann, J., Ito, N., and Vincent, E. (2013). "The diverse environments multi-channel acoustic noise database (demand): a database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics ICA2013, volume 19* (Montreal, QC: ASA).

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

Trinh, V. A., Kavaki, H. S., and Mandel, M. I. (2021). Importantaug: a data augmentation agent for speech. *arXiv preprint* arXiv:2112.07156. doi: 10.1109/ICASSP43922.2022.9747003

Tripathi, S., Tripathi, S., and Beigi, H. (2018). Multi-modal emotion recognition on iemocap dataset using deep learning. *arXiv preprint* arXiv:1804.05788. doi: 10.48550/arXiv.1804.05788

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., and Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE J. Sel. Top. Signal Process.* 11, 1301–1309. doi: 10.1109/JSTSP.2017.2764438

Valstar, M. (2016). "Avec 2016: depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge* (Amsterdam: ACM), 3–10.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Advances in Neural Information Processing Systems* (Long Beach, CA), 5998–6008.

Wu, S., Falk, T., and Chan, W.-Y. (2011). Automatic speech emotion recognition using modulation spectral features. *Speech Commun.* 53, 768–785. doi: 10.1016/j.specom.2010.08.013

Xiong, J., Lipsitz, O., Nasri, F., Lui, L. M., Gill, H., Phan, L., et al. (2020). Impact of covid-19 pandemic on mental health in the general population: a systematic review. *J. Affect. Disord.* 277, 55–64. doi: 10.1016/j.jad.2020.08.001

Xue, W., Cucchiarini, C., van Hout, R., and Strik, H. (2019). "Acoustic correlates of speech intelligibility: the usability of the egemaps feature set for atypical speech," in *Proceedings of 8th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2019)* (Graz), 48–52.

Yang, K., Lee, D., Whang, T., Lee, S., and Lim, H. (2019). Emotionx-ku: BERT-max based contextual emotion classifier. *arXiv preprint* arXiv:1906.11565. doi: 10.48550/arXiv.1906.11565

Yang, Y., and Cui, X. (2021). Bert-enhanced text graph neural network for classification. *Entropy* 23, 1536. doi: 10.3390/e23111536

Zeng, S. (2007). Audio-visual affect recognition. *IEEE Trans. Multimedia* 9, 424–428. doi: 10.1109/TMM.2006.886310

Zhang, D., Wu, L., Sun, C., Li, S., Zhu, Q., and Zhou, G. (2019). "Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations," in *IJCAI* (Macao), 5415–5421.

Zhang, Y., Wang, Q., Li, Y., and Wu, X. (2018). Sentiment classification based on piecewise pooling convolutional neural network. *Comput. Mater. Continua* 56, 285–297.

Zhao, H., Zarar, S., Tashev, I., and Lee, C.-H. (2018). "Convolutional-recurrent neural networks for speech enhancement," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Barcelona: IEEE), 2401–2405.

Zhao, J., Mao, X., and Chen, L. (2019). Speech emotion recognition using deep 1d and 2d cnn lstm networks. *Biomed. Signal Process. Control.* 47, 312–323. doi: 10.1016/j.bspc.2018.08.035