Best Practices for Noise-Based Augmentation to Improve the Performance of Deployable Speech-Based Emotion Recognition Systems

Anonymous ACL submission

Abstract

Emotion recognition models are a key component of several downstream applications, such as mental health assessments. These models are usually trained on small, clean, and synthetically controlled datasets, which leads to high failure rates in presence of 'unseen' background noises, promoting noise-overlay based adversarial attacks. Noisy data augmentation has aided robustness of speech recognition and classification models, wherein, the ground truth label remains consistent even in the presence of noise which, isn't always true for subjectively perceived emotion labels. In this work, we create realistic noisy samples of IEMOCAP, using multiple categories of environmental and synthetic noise. We evaluate how ground truth labels (human) and predicted labels (model) change as a function of these noise source introductions. We show that some commonly used noisy augmentation techniques, impact human perception of emotion, thus, falsifying the 'clean' ground truth label. Our experiments show that the performance of both, baseline, and even denoised emotion recognition models significantly declines on noisy samples as compared to that on the clean set. This performance degradation prevails when model is trained on a combination of clean and test set mismatched noisy samples. We investigate how using the above found 'human-perceptible' noise overlays can lead to inaccurate metrics when testing the model for robustness or vulnerability to adversarial attacks. Finally, we present a set of recommendations for noise-based augmentation of speech emotion datasets and for deploying the models trained using those datasets.

1 Introduction

Emotion recognition models are often used for downstream applications such as advertising, mental health monitoring (Khorram et al., 2018), hiring (emo) and surveillance (Martin, 2019). Therefore, accuracy and adversarial robustness is critical, as incorrect emotion predictions can have devastating consequences. However, robust and generalizable emotion recognition systems remain out of reach due to the complexity and variability that is inherent in the production of emotion. Further, emotion recognition is a low-resourced domain, and hence, the performance of emotion recognition models is often improved when techniques such as data augmentation or transfer learning are applied (Pappagari et al., 2021; Deng et al.). But, data augmentation, used incorrectly, has the potential to be a further source of noise and may lead to vulnerabilities against maliciously manipulated input data. In this paper we investigate noise based data augmentation techniques, their effects on labels obtained from both, human and emotion recognition models, an assessment of how these augmentation techniques could introduce errors in evaluation of targeted misclassification attacks, and finally provide recommendations for how to move forward.

One of the prime weaknesses of emotion recognition techniques centers on the very nature of the data; common datasets are generally collected in controlled laboratory environments where one of the goals is to avoid as much variation due to noise as possible. This collection method ensures that the collected data are variable only due to the presence of emotion or predefined control variables. The benefit is that the trained models generally avoid spurious correlations between noise and emotion labels, correlations that may exist in real-world data (e.g., data from frustrated users occur in an environment with traffic noise and from non-frustrated users occur in a quiet environment). However, these clean and small datasets promote a mismatch between training and testing conditions allowing an adversary to effectively probe a model, which may be overfitted on a small corpus data due to spurious and learned correlations.

Researchers have addressed these challenges separately. For emotion recognition, low resource challenge is handled by augmenting the controlled collected datasets of emotion via signal manipulation (e.g., adding background noise) aiming for enhanced generalizability to real-world noise conditions or mitigating the risk of overfitting to variations particular to a small dataset (Zheng et al.,

2016). These manipulations have included distribution shift (Abdelwahab and Busso, 2018), physiological factors (Jaiswal et al.), or pink/white noise (Parada-Cabaleiro et al., 2017; Scharenborg et al.). In speech recognition and speaker identification, the scope of these manipulations is larger, encouraging models to learn task-salient features (Mošner et al., 2019). Biberger and Ewert (2016) analyzed robust augmentation methods for automatic speech recognition including the omission of letters, changes in the speed of speech, and the presence of environmental noise through psycho-acoustic masking. They found that humans have surprisingly robust speech processing systems that can easily overcome these modulations. However, the resulting effect of similar modulations on gestalt emotion perception is understudied. The challenge associated with adversarial attacks (intentionally forcing the model to misclassify samples), has been addressed by addressed by various methods, such as noisy augmentation (Pappagari et al., 2021). But, without actually linking these two challenges together it is unlikely that emotion recognition models will be sufficiently robust and resilient due to the inherent link between them.

In this paper, we address these two challenges jointly. We first augment a subset of the Interactive Emotional Dyadic MOtion Capture (IEMOCAP) dataset (Busso et al., 2008), a common speech emotion recognition corpus by introducing noise, both by manipulating the signal and adding real-world noise. We analyze how both human perception and state-of-the-art emotion recognition algorithms, including noise-robust methods, perform differently given this noise. We then look into methods to offset these effects by first augmenting the training dataset with examples of the noise category and then introducing a denoising method to the pipeline. Finally, we discuss an often overlooked effect of the choice of noise categories for data augmentation i.e., unreliable metrics during model evaluation, specifically for the case of evaluating adversarial attack accuracy. We end the paper with a set of recommendations for emotion dataset augmentation techniques and and deployment of these emotion recognition algorithms. We investigate the following research questions:

Q1: Does the presence of noise effect emotion perception as evaluated by *human raters*? Is this effect dependent on the utterance length, loudness and type of the added noise, and the original emotion?
Q2: How does the presence of noise affect the performance of *emotion recognition models*? Does this effect vary based on the type of the added noise in agreement with changes in human perception?
Q3: Does dataset augmentation or sample denoising help improve the robustness of emotion recognition models to unseen noise?
Q4: Does the separation of noise additions into 'desirable' and 'perceptible' effect the accuracy of an attacker aiming to force the model to misclassify?
Q5: What are the recommended practices for speech emotion dataset augmentation and model deployment based on insights from the aforementioned crowdsourcing experiments?

Our findings suggest that human emotion perception is usually unaffected by the presence of environmental noise or common modulations such as reverberations and fading loudness. However, it is affected by signal manipulations such as speed, pitch change, and even the presence of pauses and fillers, which should therefore be used with caution. We show that the performance of the emotion recognition models given noisy data is improved by either dataset augmentation or sample denoising, or both. However, the performance drops when an *unseen* noise category is tested. Further, we show that these approaches are brittle to the introduction of certain common types of noise that do not change human perception, especially reverberation. Critically, we show how noise augmentation practices can lead to a falsifiable improvement when measuring efficiency of adversarial attacks, thus leading to an unreliable metric for developing and choosing methods that are best suited for avoiding these attacks. Finally, we end the paper with a set of augmentation and deployment suggestions. These types of signal modulation are common;y observed in virtual interactions. The fragility of emotion recognition models to such distortions can pose a security risk, a risk that must be well understood before these technologies are widely deployed.

2 Relation to Prior Work

Previous work focused on understanding how noise impacts machine learning models can be classified into three main directions: (a) robustness in automatic speech recognition or speaker verification; (b) noise-based adversarial example generation; and, (c) improvement in performance of machine learning models when augmented with noise.

Existing work has looked at how to built speech recognition systems that are robust to various kinds and levels of noise (Li et al., 2014). The common themes in these papers are a concentration on either

data augmentation or gathering more real-world data to produce accurate transcripts (Zheng et al., 2016). Other lines of work have looked into preventing various attacks, e.g., spoofing or recording playback, on speaker verification systems (Shim et al.). Noise in these systems is usually considered to be caused by reverberations or channel-based modulations (Zhao et al., 2014). Researchers have also looked into building noise robust emotion recognition models, using either signal transformation or augmentation (Aldeneh and Provost, 2017). But data augmentation in this analysis was formulated in a similar manner to those for speech recognition, or, speaker identification systems, which have a fixed ground truth and are independent of human perception. Previous works have also investigated adversarial example generation that aim to create audio samples that change the output of the classifier. However, these methods assume white-box access to the network and create modifications in either feature vectors or actual wav files (Carlini and Wagner; Gong and Poellabauer, 2017). This generally results in samples that either have perceptible differences when played back to a human, or are imperceptible to a human but fail to attack the model when played over-air (Carlini and Wagner).

The line of work closest to ours is data augmentation aimed for training more generalizable models. Noise and manipulation-based data augmentation is a prime research field in the space of vision and text. Sohn et al. investivated how rotating and blackening out random pixels in images to create augmented datasets leads to better performance on the test set for digit and object recognition (Sohn et al., 2020). They also investigated how data augmentation can help with zero shot object recognition in sparse classes of the dataset, providing new options for sub sampling of classes. In the field of Natural Language Processing (NLP), researchers studied how replacing words with their semantic synonyms can lead to better performance on tasks such as Part of Speech (POS) tagging and semantic parsing (Wallace et al., 2019). Researchers have explored techniques for noise based data augmentation (Kim and Kim), mostly focusing on how model training can be improved to yield better performance. However, these augmentation techniques are mostly used for acoustic event detection, speaker verification or speech recognition (Ko et al., 2015) , and has been sparingly used in audio based para-linguistic classification tasks.

The common way to deal with noise in any audio signal is to use denoising algorithms. Hence, it is a valid counterpoint to understand how machine learning models trained to recognize emotions perform if they are tested on denoised samples. To this end, we look at two major approaches to denoising algorithms in the audio space: Denoising Feature Space (Valin, 2018) and Speech Enhancement (Chakraborty et al., 2019). Denoising feature space algorithms seek to remove noise from the extracted front end features. Speech enhancement algorithms seek to convert noisy speech to more intelligible speech. Both techniques are associated with challenges, such as signal degradation (Valin, 2018) to harmonic disintegration (Valin, 2018).

Model robustness to noise or an adversarial attack can be evaluated by adding noise to the dataset and testing performance of the model. This method is commonly used for various tasks, such as, speech recognition, or speaker identification (Abdullah et al., 2019), whose perception is independent of noise as discussed before. The closest task to ours where the ground truth varies based on noise introduction is sentiment analysis. In this case, the adversarial robustness is usually tested by flipping words to their synonyms, such that the meaning of the text remains the same, and analyzing how the predictions of the model change (Ebrahimi et al., 2017). Unlike the lexical modality, speech cannot be broken down into discrete components with obvious replacements while maintaining speech feature integrity. Hence, introduction of noise for emotion recognition while assuring that the perception remains the same is more difficult.

To the best of our knowledge, this is the first work that has studied the effect of different kinds of real-world noise and varying amounts of noise contamination on the human perception of emotion and the implication of training on these datasets from the perspective of machine performance and robustness to adversarial attacks.

3 Emotion Dataset

For our study, we use the IEMOCAP dataset (Busso et al., 2008), created to explore the relationship between emotion, gestures, and speech. 10 actors, in pairs of two, one male and one female, were recorded over five sessions (either scripted or improvised). The data were segmented by speaker turn, resulting in a total of 10,039 utterances (5,255 scripted turns and 4,784 improvised turns). It contains audio, video, and associated manual transcriptions. We train the emotion classification models on the full IEMOCAP dataset (Section 6.1).

4 Noise

4.1 Environmental Noise

We define environmental noises (ENV) as additive background noises, obtained from the ESC-50 dataset(Piczak)[1], which is used for noise contamination and environmental sound classification. These environmental sounds are representative of many noises in real world deployments, especially in the context of virtual and smart home conversational agents. We use the following categories:

- Natural soundscapes (Nat), e.g., rain, wind.
- Human, non-speech sounds (Hum), e.g., sneezing, coughing etc.
- Interior/domestic sounds (Int), e.g., door creaks, clock ticks etc

We manipulate two factors in addition to noise:

- ENVPos: We vary the position of the introduction of sound that (i) starts and then fades out in loudness(St), or (ii) occurs during the entirety of the duration of the utterance(Co). Complete additive background would represent a consistent noise source in real world, e.g., fan rotation.
- ENVdB: We vary the signal to noise ratio (SNR) of additive background noise for the Co method at levels of 20dB, 10dB and 0dB.
- ENVLen: We vary the length of the introduced background noise, from a short blip to the length of the entire clip. We denote this by: Sh: Short, Me: Medium, and, Co: Complete. A background noise over the entire clip emulates a consistent real world noise source (e.g. fan rotation).

4.2 Synthetic Noise

We define synthetic noise as modulations that aren't additive in background. These kinds of noises in audio signal can occur from linguistic/paralinguistic factors, room environment, internet lags, or the physical locomotion of the speaker. We use the ten following categories:

- SpeedSeg: We speed up a random segment of the utterance by 1.25×.
- Fade: We fade the loudness of the utterance by 2% every second, which emulates the scenario of a user moving away from the speaker. We increase the loudness for fade in, and decrease for fade out.
- Filler: Insertion of non-verbal short fillers such as 'uh', 'umm' (from the same speaker) in the middle of a sentence. The insertion is either just the filler (S) or succeeded and preeceeded by a long pause (L).
- DropW: Dropping all non-essential word belonging to the set :{a, the, an, so, like, and}.
- DropLt: Phonological deletion or dropping of letters has been widely studied as a part of native US-English dialect (pho; Yuan and Liberman). Hence, we drop letters in accordance with various linguistic styles chosen from the set:{/**h**/+vowel, vowel+/n**d**/+consonant(next word), consonant+/**t**/+consonant(next word), vowel+/**r**/+consonant, /ihn**g**/}.
- Laugh/Cry: We add "sob" and "short-laughter" sounds obtained from AudioSet (Gemmeke et al., 2017) to the end of utterance.
- SpeedUtt: Speed up the entire utterance by 1.25× or 0.75×.
- Pitch: Change the pitch by ± 3 half octaves.
- Rev: Add room reverbration to the utterance.

[1] https://github.com/karoldvl/ESC-50

5 User study

We first analyze the effects of noise on human perception using crowdsourcing. We use insights from this experiment to guide the machine learning-centric analyses that follow.

5.1 Sampling and Noise-Perturbations

We randomly select 900 samples from the IEMOCAP dataset for the human perception part of the study, bounded by some constraints. The sample size is far larger than the ones used for previous perception studies (Parada-Cabaleiro et al., 2017; Scharenborg et al.). We select 100 samples from each activation and valence pair bin, i.e., 100 samples from the bin with activation: *low*, valence: *low*; 100 samples from the bin with activation: *low*, and valence: *mid*, and so on. This ensures that the chosen 900 samples cover the range of emotions expressed. We impose another constraint on these 100 samples from each bin, 30 of them are shorter than first quartile or greater than fourth quartile of utterance length in seconds to cover both extremities of the spectrum, and the remaining 70 belong in the middle. We also ensure that the selected samples had a 50-50 even split amongst gender. We introduce noise to 900 samples (Section 3). Each sample is modulated in ten ways: four randomly chosen types of environmental noise and six randomly chosen synthetic noise modulations, giving us a total of 9000 noisy samples.

5.2 Crowdsourcing Setup

We recruited workers using Amazon Mechanical Turk belonging to the United States who are native English speakers, to reduce the impact of cul-

tural variability. We ensured that each worker had $> 98\%$ approval rating and more than 500 approved Human Intelligence Tasks (HITs). We ensured that all workers understood the meaning of activation and valence using a qualification task that asked workers to rank emotion content similar to Jaiswal et al. (2019). All HIT workers were paid a minimum wage ($\$9.45$/hr), pro-rated to the minute. Each HIT was annotated by three workers.

For our main task, we created original and modulated sample pairs, and asked three workers to annotate if they perceived the pair to have the same emotion. If they said *yes* for both activation and valence, the noisy sample was labeled *same* and they could directly move to the next HIT. If they said *no*, the noisy sample was labeled *different*. In this case, they were asked to assess the activation and valence of the noisy sample using Self Assessment Manikins (Bradley and Lang, 1994) on a scale of [1, 5] (similar to original IEMOCAP annotation). We ensured the quality of the annotations by paying bonus based on time spent, not just number of HITs, and by disqualifying annotators if their level of agreement with our attention checks was low.

We created final labels for the noisy examples by taking the majority vote over each pair. The final label was either *same* emotion perception or *different* emotion perception and an averaged valence and activation score. The inter annotator agreement was 79% for activation and 76% for valence. All the samples along with the paired noisy examples and their annotations will be made available for further research.

5.3 Human Perception Study Results

Based on crowdsourcing experiment, we find that:

(a) The presence of environmental noise, even when loud, rarely affects human perception, verifying our initial judgement that humans are able to psycho-acoustically mask the background noise in various cases, as also shown in prior work (e.g., (Stenbäck, 2016)).

(b) Addition of laughter or crying does, as one might expect, change human perception.

(c) Addition of laughter or crying increases perceived activation while laughter increases perceived valence and crying decreases it.

(d) Increasing the pitch usually increases perceived activation and vice-a-versa.

(e) Increases in the speed of an utterance increases activation and vice-versa (Busso et al., 2009).

(f) The impercebtility of noise for emotion perception is not statistically correlated to dataset variables, such as, original emotion of the utterance, the gender of the speaker, and the length of the utterance.

We report our human perception results below, focusing on the five synthetic modulations that significantly change human perception of either activation or valence. The values represent the ratio of the utterances that were marked as different based on majority vote amongst all the utterances introduced with that noise. The results are as follows:

	Act	Val
`Filler(L)`	0.10	0.06
`Filler(S)`	0.06	0.03
`Laugh`	0.16	0.17
`Cry`	0.20	0.22
`SpeedUtt(0.75x)`	0.13	0.03
`SpeedUtt(1.25x)`	0.28	0.06
`Pitch(1.25x)`	0.22	0.07
`Pitch(0.75x)`	0.29	0.10

This suggests that the modifications presented in Table 1 that are imperceptible by humans for emotion perception, should not impact performance of an *ideal* machine learning model.

6 Methods

We first outline our baseline state-of-the-art emotion classifier (Section 6.1). We describe methods to denoise the data (Section 6.3) and to augment training with both the noise augment samples and the denoised samples (Section 6.2). Finally, we describe the setup and evaluation of the use case, the adversarial attack (Section 6.4).

6.1 Network

To focus on the effect of types of noise contamination on performance of emotion classifiers, we initially use the state-of-art single utterance emotion classification model which has been used in previous research (Khorram et al., 2017; Krishna et al.). In the sections that follow, we introduce noise-robust methods to understand their ability to mitigate the problems that arise.

Acoustic Features. We extract 40-dimensional Mel Filterbanks (MFB) features using a 25-millisecond Hamming window with a step-size of 10-milliseconds. Each utterance is represented as a sequence of 40-dimensional feature vectors. We z-normalize the acoustic features by each speaker.

Emotion Labels. The target emotion labels represented in a dimensional format are binned into three classes to represent `{low, medium, high}` for both activation and valence.

Architecture. We use a well known architecture (Aldeneh et al., 2017) in emotion recognition to be ensure a consistent baseline for comparison

of performance over different setups. The extracted MFBs are processed using a set of convolution layers and Gated Recurrent Units (GRU), which are fed through a mean pooling layer to produce an acoustic representation which is then fed into a set of dense layers to classify activation or valence.

Training. We use a subject independent five-fold cross validation scheme to select our train, test and validation sets. We generate noisy samples for each sample in the test as described in Section 4. We use a weighted cross-entropy loss function for each task and learn the model parameters using the RMSProp optimizer. We implement models using the Keras library (Chollet, 2015). We train our networks for a maximum of 50 epochs but use early stopping to stop the training if the validation loss does not improve after 5 consecutive epochs. Once the training process ends, we revert the network's weights to those that achieved the lowest validation loss on the emotion classification task. We use Unweighted Average Recall (UAR) (chance is 0.33) as our metric and report average over all test samples in 5 runs.

6.2 Noise Augmented Model Training

We then augment the dataset with noise. We test environmental and synthetic noise separately. We perform leave-one-noise-out cross-validation by augmenting each fold with one kind of noise from: {Hum, Int, Nat}. We train on two kinds of environmental noise and test on the third. We do the same with synthetic noise by augmenting each fold with the most impactful augmentations from: {SpeedSeg, Fade and Reverb}. We then repeat this experiment, including both the noise augmented data and the same data after denoising (Section 6.3).

6.3 Denoising

Our procedure to make a 'noise-robust' emotion recognition model is two fold: first we pass all the noisy (and non-noisy samples) in the dataset through a common denoising algorithm, Recurrent Neural Network Noise Suppression (RNNNoise, denoising feature space) (Valin, 2018), proposed in 2017 for noise suppression, and create 'noise-suppressed' samples in the training, validation and testing sets. RNNNoise is trained on environmental noises, and these noises overlap considerably with those in our dataset. We use the default parameters used in the algorithm and use it on an 'as-is' basis for our experiments. We assume that the system does not have the knowledge of 'which' noise is introduced and therefore do not compare with other denoising algorithms that assume a-priori knowledge of noise category (e.g., (Tibell et al., 2009)).

Table 1: The original activation (Act) and valence (Val) performance is 0.67 and 0.59, respectively. The table shows the change in UAR on IEMOCAP (Iem) and the Augmented IEMOCAP (Iem(Aug)) datasets. The rows are augmentation methods that were evaluated as imperceptible for emotion by human annotators.

		UAR			
		Iem		Iem(Aug)	
		Act	Val	Act	Val
Environmental Noise					
NatSt		-.25	-.24	.22	.18
NatdB (Co)	20dB	-.33	-.34	.22	.13
	10dB	-.37	-.41	.33	.31
	0dB	-.40	-.41	.27	.26
HumSt		-.22	-.24	.15	.13
HumdB (Co)	20dB	-.33	-.37	.16	.16
	10dB	-.37	-.42	.21	.26
	0dB	-.40	-.42	.25	.21
IntSt		-.21	-.25	.15	.18
IntdB (Co)	20dB	-.31	-.39	.20	.22
	10dB	-.34	-.39	.23	.19
	0dB	-.40	-.41	.30	.26
Synthetic Noise					
SpeedSeg		-.09	-.12	.03	.02
Fade	In	-.07	-.10	.05	.04
	Out	-.09	-.14	.02	.06
DropW		-.04	-.05	.02	.00
DropLt		-.03	-.02	.06	.03
Reverb		-.36	-.37	.05	.04

We then train a network with the architecture and hyperparameter ranges described in Section 6.1. The network is trained on features derived from the denoised samples. Therefore, it can learn to recognize emotion, even if the chosen RNNNoise algorithm introduces unusual signal properties.

6.4 Use Case: Adversarial Attack

The methods thus far present techniques to generate noisy data, understand the fragility of the model with respect to these noise modulations, and then to counteract this fragility. Finally, we present methods to investigate the implications of model fragility. We frame the problem as a decision boundary untargeted attack, which tries to find the minimal amount of noise perturbation possible in *k-queries* such that the decision of a model changes, while having no knowledge about the internal workings of the model (probabilities or gradients). This framing allows us to understand what may happen if we deploy a fragile model into the real world.

A decision-based attack starts with a large adversarial perturbation and seeks to reduce this perturbation while staying adversarial. In our case, because we are considering noise augmentation, we define the distance between an adversarial sam-

ple and the original sample as the degradation in signal to noise ratio (SNR). The input to this attack model is a set of permissible noise categories (e.g., white, Gaussian, or categorical), and the original signal. We also modify the decision boundary attack to first use any random sound from four categories at the lowest SNR level. If the attack model is successful, it then shifts to using other additive noise options from that category.

Optionally, we provide information about how much performance degradation of the model is observed given a particular noise type, to the attack model, such that it is more successful given the same limited budget of *k-queries*. The modified sample is considered a successful attack only when both, the noise addition chosen by the algorithm is at SNR $>$ 10dB, and the model changes its output label. The former condition ensures that the sample is not audibly judged as contaminated by humans (Kidd Jr et al., 2016). We use the Foolbox Tooklit (Rauber et al., 2017) to implement a decision boundary attack (Brendel et al., 2017).

For our purpose, we assume that an attacker has access to a subset of labelled data from the set of users (U) in IEMOCAP and that the training and the testing conditions remain the same as that of the black-box emotion recognition model. We calculate the average accuracy of a successful attack, Acc_{att}, for any sample from user U_a using k queries. We calculate Acc_{att} varied over two variables, each with two possible values: 1) $Corr \in \{Y, N\}$ to represent if the attacker has access to the information about the correlation between the performance deterioration of emotion recognition and the noise category and 2) $Noises \in \{All, Im\}$ to indicate if the evaluation is performed on ***All*** noises or only noises that are ***Im****perceptible* by humans for emotion perception.

7 Analysis

7.1 Baseline (SoTA) Model Performance

We assess the performance of the model on the original IEMOCAP data and find that the model obtains a performance of 0.67 UAR on the activation and 0.59 UAR on the valence task. Next, we augment the test samples of each fold with each of the noise types (Section 4) and investigate how the performance of the model changes. We do not include noise types that were found to often affect human perception (e.g., *Pitch, SpeedUtt, Laugh*).

We find that the machine learning model's performance decreases by an average of 33% for environmental noise, fading, and reverberation. There is also a smaller drop in performance for speeding up parts of the utterance and dropping words, showing the brittleness of these models. Table 1 reports the percentage change in performance when testing on noisy test data as over clean test data (0.67 for activation and 0.59 for valence).

7.2 Baseline Model Performance with Augmentation

We find that data augmentation improves performance on noisy test data over a baseline system trained only on the original IEMOCAP data. This is very pronounced (an increase of 22%) when the environmental noise is introduced at the start of the utterance, e.g., when the test set is introduced with `NatSt` and the train set is introduced with `HumSt` and `IntSt`. We speculate that the network learns to assign different weights to the start and end of the utterance to account for the initial noise.

Though the performance increases on the continuous background noise contaminated test samples, is aided by addition of training samples that have a continuously present background noise as well, the performance is still affected due to introduction of *new* noise in the test set. We also find that it is hard to improve the performance of the system on utterances contaminated with reverberation, a common use case, even when the training set is augmented with other types of noise. This can be because, reverberation adds a continuous human speech signal in the background delayed by a small period of time. None of the other kind of noises have speech in them, and hence augmentation doesn't aid the model to learn robustness to this kind of noise.

7.3 Denoising Algorithms and Model Performance

We find that adding a denoising component to the models leads to a significant improvement (average of $23\% \pm 3\%$ across all environmental noise categories) in continuous noise introduction at 20dB SNR as compared to when there is no denoising or augmentation performed. We observe a decline in performance when using noise suppression algorithms to deal with other signal to noise ratios of continuous noise additions, possibly due to masking of emotional information from the speech to maintain comprehensibility which may flatten the tone in the obtained denoised signal (Spadini and Suyama, 2019). We further show that the addition of a denoising component does not significantly improve performance in presence of signal manip-

Table 2: Attacker efficiency (Acc_{att}) using noise-based adversarial methods over the number of black box queries an attacker can access (k), the knowledge of correlation between performance deterioration of emotion recognition and noise category (*Corr*), and whether the attack efficiency is evaluated only on those noise categories that human perception doesn't change (*Noises*). Higher values are better. Maximum value for each k is in **bolds**. Significance is established using paired t-test, adjusted p-value < 0.05

Corr	*Noises*	k=5	k=15	k=25
No	Im	0.22	0.29	0.39
Yes	Im	0.32	0.38	0.53
No	All	0.31	0.36	0.45
Yes	All	**0.36**	**0.43**	**0.58**

ulation rather than just noise addition, for example, when samples were faded in or out or segments were sped up. While we did see an improvement in performance (an average of $+36\%$) for *unseen* reverberation contaminated samples as compared to data augmentation, the performance is still significantly lower (-28%) than when tested on a clean test set. Finally, we observe an general trend of increase in emotion recognition performance for the combined dataset (noisy and non-noisy samples), as compared to when the model is trained on the clean training set, which supports the findings from previous dataset augmentation researches.

7.4 Use Case: Adversarial Attack Evaluation

In this section, we aim to show how the addition of 'non-allowed' noises can not only impact the brittleness of already trained models, but also lead to inaccurate evaluation metrics. We use the term 'non-allowed' noises to refer to the set of noises that have been shown to change human perception, because they lead to unpredictable shifts in ground truth labels. We use the downstream task of noise contamination for adversarial attacks as an example to show how the noise-augmentation criterion can lead to incorrect evaluation metrics. Table 2 shows the attacker efficiency. We find that an attack is more likely to succeed if the attacker can corrupt the sample with 'non-allowed' noises. For example, the attacker, with no additional knowledge has a success rate of 0.31 with five allowed queries (k=5) when using complete set of noises for corruption as compared to 0.22 where they only use noises that do not change human perception of emotion, demonstrating fragile benchmarking and evaluation of adversarial efficiency and robustness.

8 Recommendations

We propose a set of recommendations, for both augmentation and deployment of emotion recognition models in the wild, that are grounded in human perception. For augmentation, we suggest that:

1. Environmental noise should be added to datasets to improve generalizability to varied noise conditions, whether using denoising, augmentation, or a combination of both.
2. It is good to augment datasets by fading the loudness of the segments, dropping letters or words, and speeding up small (no more than 25% of the total sample length) segments of the complete sound samples in the dataset. But it is important to note that these augmented samples should not be passed through the denoising component as the denoised version loses emotion information.
3. One should not change the speed of the entire utterance more than 5% and should not add intentional pauses or any background noises that elicit emotion behavior, e.g., sobs or laughter.

Regarding deployment, we suggest that:

1. Noisy starts and ends of utterances can be handled by augmentation, hence, if the training set included these augmentations, there is no issue for deployed emotion recognition systems.
2. Reverberation is hard to handle for even augmented emotion recognition models. Hence, the samples must either be cleaned to remove the reverberation effect, or must be identified as low confidence for classification.
3. Deploy auxilary models that classify the noise types, and degree/intent of signal manipulation are useful filters for identifying maliciously manipulated data before the data is sent to be processed by the emotion recognition model.

9 Conclusion

In this work, we study how the presence of real world noise, environmental or synthetic, affects human emotion perception. We identify noise sources that do not affect human perception, such that they can be confidently used for data augmentation. We look at change in performance of models that are trained on the original dataset when tested on these noisy samples and if augmentation of the training set leads to an improvement. We conclude that, unlike humans, machine learning models are extremely brittle to the introduction of many kinds of noise. While the performance of the machine learning model on noisy samples is aided from augmentation, the performance is still significantly lower when the noise in the train and test environments does not match. In this paper, we demonstrate fragility of the emotion recognition systems and valid methods to augment the datasets, which is a critical concern in real world deployment.

10 Ethical Considerations

Data augmentation is often applied to speech emotion recognition to improve robustness. Better augmentation methods are an important way to not only ensure reliability and robustness of these models, but also improve the real-life adoption in high-stakes downstream applications. Knowing when human perception of emotion can change in the presence of noise is needed to design better model unit tests and adversarial tests for verifying the reliability of the system. However, emotion variability is often dependent on multiple factors, such as, culture, race, gender, age etc, some of which are highly protected variables. These models can also encode stereotypical expected behavior from a certain group, and hence have a higher error rate for other groups. It is important to note that this paper considers a small set of crowd-sourced workers as human raters of emotion perception, who are located in the United States and are well versed in English, the language this dataset is collected in, and the model is trained on.

References

How emotion ai can transform large-scale recruitment processes shorturl.at/gcs15, note = Accessed: 10-21-2019.

Phonological history of english consonant clusters= https://bit.ly/3l08zbu, Accessed: 10-21-2019.

Mohammed Abdelwahab and Carlos Busso. 2018. Domain adversarial for acoustic emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Hadi Abdullah, Washington Garcia, Christian Peeters, Patrick Traynor, Kevin RB Butler, and Joseph Wilson. 2019. Practical hidden voice attacks against speech and speaker recognition systems. *arXiv preprint arXiv:1904.05734*.

Zakaria Aldeneh, Soheil Khorram, Dimitrios Dimitriadis, and Emily Mower Provost. 2017. Pooling acoustic and lexical features for the prediction of valence. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 68–72.

Zakaria Aldeneh and Emily Mower Provost. 2017. Using regional saliency for speech emotion recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2741–2745. IEEE.

Thomas Biberger and Stephan D Ewert. 2016. Envelope and intensity based prediction of psychoacoustic masking and speech intelligibility. *The Journal of the Acoustical Society of America*.

Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*.

Wieland Brendel, Jonas Rauber, and Matthias Bethge. 2017. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*.

Carlos Busso, Murtaza Bulut, and Sungbok Lee. 2009. Shrikanth narayanan fundamental frequency analysis for speech emotion processing. *The role of prosody in affective speech*, 97:309.

Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*.

Rupayan Chakraborty, Ashish Panda, Meghna Pandharipande, Sonal Joshi, and Sunil Kumar Kopparapu. 2019. Front-end feature compensation and denoising for noise robust speech emotion recognition. In *INTERSPEECH*, pages 3257–3261.

François Chollet. 2015. keras. https://github.com/fchollet/keras.

Jun Deng, Zixing Zhang, Erik Marchi, and Björn Schuller. Sparse autoencoder-based feature transfer learning for speech emotion recognition. In *2013 humaine association conference on affective computing and intelligent interaction*.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2017. Hotflip: White-box adversarial examples for text classification. *arXiv preprint arXiv:1712.06751*.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA.

Yuan Gong and Christian Poellabauer. 2017. Crafting adversarial examples for speech paralinguistics applications. *arXiv preprint arXiv:1711.03280*.

Mimansa Jaiswal, Zakaria Aldeneh, Cristian-Paul Bara, Yuanhang Luo, Mihai Burzo, Rada Mihalcea, and Emily Mower Provost. 2019. Muse-ing on the impact of utterance ordering on crowdsourced emotion annotations. In *IEEE ICASSP*.

Mimansa Jaiswal, Zakaria Aldeneh, and Emily Mower Provost. Controlling for confounders in multimodal emotion classification via adversarial learning. In *2019 ICMI*.

Soheil Khorram, Zakaria Aldeneh, Dimitrios Dimitriadis, Melvin McInnis, and Emily Mower Provost. 2017. Capturing long-term temporal dependencies with convolutional networks for continuous emotion recognition. *arXiv preprint arXiv:1708.07050*.

Soheil Khorram, Mimansa Jaiswal, John Gideon, Melvin McInnis, and Emily Mower Provost. 2018. The priori emotion dataset: Linking mood to emotion detected in-the-wild. *arXiv preprint arXiv:1806.10658*.

Gerald Kidd Jr, Christine R Mason, Jayaganesh Swaminathan, Elin Roverud, Kameron K Clayton, and Virginia Best. 2016. Determining the energetic and informational components of speech-on-speech masking. *The Journal of the Acoustical Society of America*.

Hyoung-Gook Kim and Jin Young Kim. Acoustic event detection in multichannel audio using gated recurrent neural networks with high-resolution spectral features. *ETRI Journal*.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Kalpesh Krishna, Liang Lu, Kevin Gimpel, and Karen Livescu. A study of all-convolutional encoders for connectionist temporal classification. In *2018 IEEE ICASSP*.

Jinyu Li, Li Deng, Yifan Gong, and Reinhold Haeb-Umbach. 2014. An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Nicole Martin. 2019. The major concerns around facial recognition technology.

Ladislav Mošner, Minhua Wu, Anirudh Raju, Sree Hari Krishnan Parthasarathi, Kenichi Kumatani, Shiva Sundaram, Roland Maas, and Björn Hoffmeister. 2019. Improving noise robustness of automatic speech recognition via parallel data and teacher-student learning. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6475–6479. IEEE.

Raghavendra Pappagari, Jesús Villalba, Piotr Żelasko, Laureano Moro-Velazquez, and Najim Dehak. 2021. Copypaste: An augmentation method for speech emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Emilia Parada-Cabaleiro, Alice Baird, Anton Batliner, Nicholas Cummins, Simone Hantke, and Björn W Schuller. 2017. The perception of emotions in noisified nonsense speech. In *INTERSPEECH*.

Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*.

Jonas Rauber, Wieland Brendel, and Matthias Bethge. 2017. Foolbox: A python toolbox to benchmark the robustness of machine learning models. *arXiv preprint arXiv:1707.04131*.

Odette Scharenborg, Sofoklis Kakouros, and Jiska Koemans. The effect of noise on emotion perception in an unknown language. In *Proc. 9th International Conference on Speech Prosody 2018*.

Hye-Jin Shim, Jee-Weon Jung, Hee-Soo Heo, Sung-Hyun Yoon, and Ha-Jin Yu. Replay spoofing detection system for automatic speaker verification using multi-task learning of noise classes. In *2018 TAAI*.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*.

Tito Spadini and Ricardo Suyama. 2019. Comparative study between adversarial networks and classical techniques for speech enhancement. *arXiv preprint arXiv:1910.09522*.

Victoria Stenbäck. 2016. *Speech masking speech in everyday communication: The role of inhibitory control and working memory capacity*, volume 1559. Linköping University Electronic Press.

Kajsa Tibell, Hagen Spies, and Magnus Borga. 2009. Fast prototype based noise reduction. In *Scandinavian Conference on Image Analysis*, pages 159–168. Springer.

Jean-Marc Valin. 2018. A hybrid dsp/deep learning approach to real-time full-band speech enhancement. In *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5. IEEE.

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019. Allennlp interpret: A framework for explaining predictions of nlp models. *arXiv preprint arXiv:1909.09251*.

Jiahong Yuan and Mark Liberman. Automatic detection of "g-dropping" in american english using forced alignment. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE.

Xiaojia Zhao, Yuxuan Wang, and DeLiang Wang. 2014. Robust speaker identification in noisy and reverberant conditions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Stephan Zheng, Yang Song, Thomas Leung, and Ian Goodfellow. 2016. Improving the robustness of deep neural networks via stability training. In *Proceedings of CVPR*.