This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3086050, IEEE Transactions on Affective Computing

You're Not You When You're Angry: Robust Emotion Features Emerge by Recognizing Speakers

Zakaria Aldeneh, *Member, IEEE,* and Emily Mower Provost, *Senior Member, IEEE*

Abstract—The robustness of an acoustic emotion recognition system hinges on first having access to features that represent an acoustic input signal. These representations should abstract extraneous low-level variations present in acoustic signals and only capture speaker characteristics relevant for emotion recognition. Previous research has demonstrated that, in other classification tasks, when large labeled datasets are available, neural networks trained on these data learn to extract robust features from the input signal. However, the datasets used for developing emotion recognition systems remain significantly smaller than those used for developing other speech systems. Thus, acoustic emotion recognition systems remain in need of robust feature representations. In this work, we study the utility of speaker embeddings, representations extracted from a trained speaker recognition network, as robust features for detecting emotions. We first study the relationship between emotions and speaker embeddings and demonstrate how speaker embeddings highlight the differences that exist between neutral speech and emotionally expressive speech. We quantify the modulations that variations in emotional expression incur on speaker embeddings and show how these modulations are greater than those incurred from lexical variations in an utterance. Finally, we demonstrate how speaker embeddings can be used as a replacement for traditional low-level acoustic features for emotion recognition.

Index Terms—emotion recognition, speaker recognition, speaker embeddings, speaker representations, transfer learning

1 Introduction

The automatic detection of emotion from the acoustics of speech can help improve human-machine interaction by providing the necessary context to interpret and appropriately react to the behaviors of users. The features used to describe the acoustic signal are a crucial aspect of any emotion recognition model. Consequently, various features have been proposed in the literature for the task of speech emotion recognition (e.g., [1], [2], [3]). However, the extraction of many of these proposed features is susceptible to distortions due to variations in lexical content, the presence of environmental noise, or domain shifts. As a result, there remains a need for robust paralinguistic features that abstract extraneous low-level variations present in the acoustic signal and only capture speaker characteristics that are necessary for predicting emotions.

Previous research has shown that neural networks trained discriminatively on large and diverse datasets learn to extract robust features that are invariant to noise and domain-shifts (e.g., [4], [5]). These features are obtained from intermediate representations that the trained networks extract from the input signal. However, the main requirement for learning powerful features using neural networks is access to large labeled datasets; a requirement that is still challenging to fulfill in the affective computing community in general, and in the emotion recognition community in particular. The challenges associated with finding media sources that provide varied emotional data as well as the challenges associated with annotating the data with accurate emotion labels are the driving reasons behind the data sparsity problem in emotion recognition.

- *Z. Aldeneh and E. Mower Provost are with the University of Michigan, Ann Arbor, MI, USA.*
  Email: {aldeneh, emilykmp}@umich.edu

Many paralinguistic tasks are closely related, and thus, the representations extracted while solving one paralinguistic task can be used for solving other tasks [6], [7]. Specifically, previous research showed that representations learned while solving the emotion recognition task are useful for solving other paralinguistic tasks, such as gender detection and speaker identification [7]. However, unlike emotion recognition, speaker recognition does not suffer from the problem of data sparsity; there are multiple large-scale datasets with speaker labels (e.g., [8], [9], [10]). In this work, we ask if speaker recognition can help emotion recognition by attenuating the challenges that come with having limited amounts of labeled emotion data. We hypothesize that we can improve emotion recognition performance by leveraging speaker embeddings, feature representations trained for the speaker recognition task. Our work complements previous research by demonstrating that speaker embeddings can be used as a replacement to common paralinguistic features in emotion recognition applications.

We propose to study the relationship between emotion and speaker embeddings and assess the embeddings' utility as general paralinguistic features. First, we quantify the effect of emotion on speaker embeddings to determine if the speech characteristics that the embeddings capture can be used for recognizing emotional expression. We hypothesize that emotionally charged vocal expressions change speech characteristics that are captured by speaker embeddings (i.e., speakers sound less like themselves when their vocal expressions are emotionally charged). This hypothesis is supported by existing work, which studied the effect of emotion on speaker representations (e.g., i-vector), focusing on changes in the equal error rate (EER) in speaker verification tasks [11], [12], [13], [14]. However, the focus on the EER metric obfuscated the utility

of speaker embeddings as paralinguistic features because the EER metric, as used in speaker verification tasks, measures the performance as a function of a general population of test speakers. In other words, although emotionally charged vocal expressions might change how identity is encoded for a certain speaker, individual speakers might still sound more like themselves when compared to a general population of other test speakers. In this work, we instead quantify the effect of emotion on speaker representations by hypothesizing that representations extracted from neutral speech, as a group, has more intra-group similarity, compared to the similarity between neutral and emotional speech. We test this hypothesis using a novelty detection framework, implemented using autoencoders, with reconstruction error as a proxy for similarity. The benefit to this paradigm is that it allows us to ask not whether the emotional speech belongs to a different speaker, but instead if the differences in emotional speech are captured by speaker representations. Our results suggest that emotional speech significantly changes speaker embeddings from their neutral representation and that these changes can be utilized in a novelty detection framework for detecting non-neutral speech.

Next, we assess the effectiveness of speaker embeddings for detecting emotions by comparing them not only to state-of-the-art paralinguistic features but also to emotion recognizers from the literature. We expect speaker embeddings to be more robust to the variations introduced by domain shifts compared to common paralinguistic features used in the emotion recognition literature. This is because neural networks used for extracting the speaker embeddings are trained on large and in-the-wild datasets, which make the extracted embeddings invariant to changes in recording conditions and background noises. As a result, we expect speaker embeddings to capture high-level speaker characteristics that can be beneficial for recognizing emotionally expressive speech while abstracting any low-level variations present in the acoustic signal. We test this hypothesis by running both within-corpus and cross-corpus emotion recognition experiments. Cross-corpus setups make the emotion recognition task more challenging as trained models cannot rely on spurious correlations that exist within a dataset to make predictions. Our results demonstrate that emotion recognition models that use speaker embeddings as features outperform those that use state-of-the-art paralinguistic features, especially in cross-corpus settings.

To summarize, the novelty of this work is three-fold: (1) we demonstrate how speaker embeddings highlight the differences that exist between neutral speech and emotionally expressive speech; (2) we show how speaker embeddings can be used in a novelty detection framework for establishing a baseline of how a speaker sounds in the neutral state, and for detecting deviations from this baseline neutral state; and (3) we demonstrate how speaker embeddings provide a robust replacement to general paralinguistic features for recognizing emotional expression. The remainder of this paper is organized as follows. Section 2 covers related work. Section 3 covers the proposed approach. Section 4 covers the datasets used in our work. Sections 5 and 6 cover the experiments and results. Finally, Section 7 includes concluding remarks and proposed future directions.

2 Related Works

2.1 Speaker Representations and Emotional Speech

There are several works that studied the relationship between speaker representations and emotional speech. In this section, we cover works that looked at this relationship as it relates to traditional (e.g., i-vectors) and neural speaker representations. i-vectors are common representations used in speaker identification and verification applications [15]. They capture several sources of variation (e.g., identity, age, gender) present in the acoustic signal as represented by the Gaussian mixture model (GMM) mean supervector. More recently, neural representations have outperformed their i-vector counterparts (e.g., [16], [17], [18]). These representations are extracted from the intermediate layers of a neural network that was discriminatively trained to classify speakers. Some common neural representations introduced in the literature include the d-vector and x-vector representations [9], [17], [17], [18], [19].

One question with these representations is how other modulations (e.g., emotion) change their ability to recognize speakers. Previous research used degradation in the EER metric in a speaker verification task as a proxy for quantifying the effect of emotion on speaker representations [12], [13], [14], [20]. However, one limitation with the use of the EER metric for this purpose is that it measures both the inter- and intra-speaker variations in the representations. In other words, the negative samples used when evaluating the EER for a speaker always came from a different speaker (i.e., a speaker is always compared to other speakers). So the metric will only be affected if the variations due to emotions are bigger than those due to changes in speaker identity. In contrast, we study how emotion modulates speaker representations by treating neutral speech from a given speaker as a group, and determining if this group has more intra-group similarity, compared to the similarity between neutral and emotional speech. The similarity measure is used as a proxy for the amount of modulation that emotion incurs on the speaker representations. The benefit of this approach is that it allows us to determine if differences in emotional speech are captured by speaker recognition features.

2.2 Speech Representations for Emotion Recognition

Many of the contributions in the early works in emotion recognition came from engineering features to reflect emotion variations in speech. Several of these features were borrowed from acoustic analysis studies done on speech segments extracted from individuals displaying different emotions while performing various tasks [21], [22]. One of the most common paradigms for extracting acoustic features for emotion recognition involves two steps. First, low-level-descriptors (LLDs) are extracted using a short sliding window (e.g., extracted every 25 milliseconds) applied to the acoustic signal. Then, a set of statistical functions are applied to these LLDs to get a feature representation of an utterance. Some popular feature sets that were developed include the INTERSPEECH 2009 (IS09) Emotion Challenge features, the INTERSPEECH 2013 Computational Paralinguistics ChallengeE (ComParE), the Geneva Minimalistic Acoustic Parameter Set (GeMAPS), and the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [1], [2], [3], [23]. The benefit of this paradigm is that it allows for a description of how properties of the low-level acoustic features change over the course of an utterance, obviating the need for

a detailed focus on the short-time dynamical properties of the features.

Yet, the short-time dynamical properties of acoustic features convey critical cues into an individual's emotions. Previous research has demonstrated how the application of statistical functionals in the feature extraction process can obfuscate these cues and has shown that modeling the acoustic features directly in neural networks can alleviate this problem [24]. However, one challenge with using low-level acoustic features (e.g., MFCCs, pitch, etc.) to directly predict emotion is that their extraction can be significantly affected by variations in the recording conditions or variations in the lexical content of the utterance [25]. In other words, the features extracted from an utterance can look different depending on what the speaker said or depending on the environment of the speaker during the recording.

Representation learning, through the use of neural networks, has been shown to be an effective way to learn powerful features that are invariant to lexical content and recording conditions (e.g., [26], [27]). As a result, more recent approaches to emotion recognition from speech have focused on using neural networks to train recognition models that rely on minimally engineered features. These works have used spectrograms, filterbanks, or raw-waveforms for building emotion recognition models [28], [29], [30]. However, datasets used for building speech emotion recognition models remain significantly smaller than those used for building other speech models (e.g., speaker recognition). This hinders the ability of neural networks trained for the task of emotion detection to extract robust representations from the acoustic signal to be used in other domains or applications.

In this work, we show that features extracted from a neural network that was trained for speaker recognition can be used as general features for detecting emotions. We demonstrate how emotional expression modulates speaker embeddings from their neutral representation, and demonstrate how these modulations can be used for detecting emotional expression. Finally, we show that speaker embeddings can outperform traditional state-of-the-art features in challenging cross-corpus emotion recognition tasks.

3 Method

In this work, we propose the use of speaker embeddings as a replacement to traditional paralinguistic acoustic features for the recognition of emotions. We introduce speaker embeddings and the model used to extract them in this section.

3.1 Speaker Embeddings

Speaker embeddings are fixed-size vector representations of variable-length utterances. They are typically used in speaker recognition and diarization tasks [9], [17], [18], and can also be used for adapting acoustic models in automatic speech recognition systems [31]. The current standard for extracting robust speaker embeddings is by taking the outputs from an intermediate layer of a neural network that was discriminatively trained to identify speakers from a large set of individuals. Common speaker embeddings from the literature include d-vectors, x-vectors, and embeddings extracted from the VGG-M speaker identification network [9], [17], [18].

Speaker recognition neural networks map low-level acoustic features (e.g., Mel-filterbanks, MFCCs, spectrograms) extracted from utterances to speaker identities present in the training set. The representations (i.e., transformation) that such networks learn in the process can be used for extracting general embeddings to represent utterances from new speakers not seen in the training phase. These representations encode speech characteristics needed for recognizing speakers but abstract low-level variations that are not needed for recognizing speakers.

3.2 The x-vector Model

We focus our work on speaker embeddings extracted from the x-vector model as described in [18], [26]. We choose to work with the x-vector system because it has been demonstrated that it provides state-of-the-art embeddings for speaker recognition and diarization applications [18], [32], [33], [34], and because it is built on top of the open-source Kaldi toolkit [35]. The network used for extracting x-vectors is summarized in Table 1, and it consists of three parts: (1) frame-level feature extraction sub-network; (2) statistics pooling layer; and (3) and utterance-level classification sub-network.

The frame-level feature extraction sub-network takes in a sequence of 30-dimensional MFCC frames, where each frame represents 25 millisecond and outputs a sequence of 512-dimensional features. It consists of five layers with a time-delay architecture. The first layer stacks the current frame at t with context frames from the previous two and the next two time steps. The second and third layers stack the current frame at t with context frames $t\pm2$ and $t\pm3$, respectively. The fourth and fifth layers do not add any context frames and only transform the representations at the current frame. The statistics pooling layer summarizes the frame-level features by taking the mean and standard deviation across the time dimension. Finally, the utterance-level classification sub-network consists of two fully-connected layers and a softmax layer for classifying speakers.

Given a variable-length utterance by a speaker that was not seen in the training phase, a fixed-size representation for this utterance can be obtained by taking the output of the "segment6" layer (before the non-linearity) from the neural network summarized in Table 1. We use the outputs of "segment6" layer as our embeddings for two reasons. First, previous research has suggested that they encode information relating to emotion, speaking style, and speaking rate [32], [36]. Second, previous research found that they are better equipped than other outputs for capturing speaker characteristics in speaker verification tasks [26].

4 Datasets

4.1 IEMOCAP

The interactive emotional dyadic motion capture (IEMOCAP) dataset was collected to study audio-visual emotional expression in dyadic interactions [37]. Interactions in the dataset were recorded from five dyadic sessions, each between a male and a female actor. In each session, the actors perform a series of scripted and improvised scenarios designed to elicit emotion expression. The dataset contains approximately 12 hours of speech from 10 speakers (five males and five females).

The recordings from each interaction were manually segmented into utterances based on speaker turns in the dialogue.

TABLE 1: The network architecture used in the speaker identification task taken from [26]. Speaker embeddings are extracted from the segment6 layer. N is the total number of speakers used in the training phase. T is the total number of frames in an utterances. The input size of 150 for the frame1 layer is the result of stacking five context frames, each with a size of 30. The input sizes of 1536 for the frame2 and frame3 layers are a result of stacking three context frames, each with a size of 512.

Layer	Layer context	Total context	Input × Output
frame1	$[t-2, t+2]$	5	150×512
frame2	$\{t-2, t, t+2\}$	9	1536×512
frame3	$\{t-3, t, t+3\}$	15	1536×512
frame4	$\{t\}$	15	512×512
frame5	$\{t\}$	15	512×1500
stats. pooling	$[0, T)$	T	$1500T \times 3000$
segment6	$\{0\}$	T	3000×512
segment7	$\{0\}$	T	512×512
softmax	$\{0\}$	T	$512 \times N$

Each utterance was then annotated for emotion using both the categorical and dimensional definitions of emotion. We use the categorical representation of emotion in this work and focus on utterances that received majority agreement. The final subset that we use contains 1,636 *happy*, 1,103 *angry*, 1,708 *neutral*, and 1,084 *sad* utterances, giving a total of 5,531 utterances. We combine *happy* and *excited* utterances to form our *happy* category.

4.2 MSP-IMRPOV

The MSP-IMRPOV dataset was collected to study audio-visual emotional expression in dyadic interactions while maintaining partial control over lexical content [38]. Interactions in the dataset were recorded from six dyadic sessions, each between a male and a female actor. The actors in each dyadic interaction improvise scenarios that lead one of them to utter a target sentence in a specific emotion. This approach of eliciting emotion was designed to maintain the spontaneous nature of the interaction while controlling for lexical content. Overall the dataset contains approximately 9 hours of speech from 12 speakers (six males and six females).

Similar to the recordings in IEMOCAP, the ones in MSP-IMPROV were manually segmented into utterances based on speaker turns from each dialogue. The resulting utterances were then annotated for emotion following both the categorical and dimensional definitions of emotion. We use the categorical representation of emotion and utterances that received majority agreement by annotators. The final subset that we use contains 2,644 *happy*, 792 *angry*, 3,477 *neutral*, and 885 *sad* utterances, giving a total of 7,798 utterances.

4.3 VESUS

One limitation with both the IEMOCAP and MSP-IMPROV datasets is that they have very few utterances where the lexical content is the same but the emotion varies; making it difficult to study the influence of emotion and lexical variations on the embeddings independently. Studying these two variables independently is necessary since emotions modulate not only speech acoustics, but also language [39]. These modulations can influence the sequence of phonemes that are uttered, which can then affect the extracted speaker embeddings. As a result, the Varied Emotion in Syntactically Uniform Speech (VESUS) dataset was collected to provide the research community with a lexically controlled emotional dataset [40]. Over 250 distinct phrases were uttered by 10 actors (five males and five females) while portraying five emotional states (*neutral*, *angry*, *happy*, *sad*, and *fear*). The phrases were chosen such that they are semantically neutral, i.e., they don't carry any emotional connotation.

Overall the dataset contains approximately six hours of speech. The utterances in the dataset contain labels assigned based on the intended emotion by the actors and labels assigned based on the perceived emotion collected from 10 crowd-sourced annotators. We run two separate analyses whenever we use the VESUS dataset: one using the intended (i.e., instructed) emotion labels, and another using utterances that achieved at least 50% consistency among the crowd-sourced annotators (i.e., at least five annotators agreed).

5 Experiments

This section describes the experiments used to assess the utility of speaker embeddings in emotion recognition tasks. The first experiment quantifies the effect of emotion variation on speaker embeddings; teasing out the effects on the embeddings due to emotion variations from those due to lexical variations. The second experiment compares the performance of an emotion recognition model trained and evaluated with speaker embeddings as features to the performance of recognition models trained and evaluated with state-of-the-art features used in the emotion literature.

The speaker embeddings that we use in all of our experiments were extracted using a pre-trained[1] x-vector model that was discriminatively trained to identify speakers in the combined VoxCeleb1 and VoxCeleb2 datasets [9], [10]. The combined VoxCeleb datasets contain more than 2,000 hours of speech (more than 1 million utterances) from more than 7,000 speaker identities. The x-vector model, summarized in Table 1, takes in the voiced frames of an utterance as an input and gives a speaker identity as an output. The input features to the x-vector model are 30-dimensional Mel-frequency cepstral coefficients (MFCCs) extracted from 16kHz utterances using a 25 millisecond sliding window. All utterances are mean normalized using a three-second window before being fed into the speaker identification network. A more detailed training recipe for the speaker identification network can be found in [18].

5.1 Experiment 1: Speaker Embeddings and Emotions

In this experiment, we quantify the effect emotion has on speaker embeddings to examine whether or not the embeddings capture speech characteristics that are changed by emotion. We hypothesize that embeddings are modulated by emotional speech, allowing us to either preserve or enhance the differences that exist between a neutral expression and the expression of emotion. We formulate this problem by asserting that neutral speech, as a group, has more intra-group similarity, compared to the similarity between neutral and emotional speech. We test this hypothesis using a novelty detection framework, implemented using autoencoders, with reconstruction error as a proxy for similarity. The use of the reconstruction

1. https://kaldi-asr.org/models/m7

error of an autoencoder for novelty detection tasks has been studied for other applications by several works in the literature (e.g., [41], [42], [43]). To the best of our knowledge, we are the first to propose the use of autoencoders to analyze the effect of the variations in emotion and in lexical content on speaker embeddings.

We address the following two questions about the relationship between speaker embeddings and emotion in our experiments:

- Q1: Do variations in emotion significantly modulate speaker embeddings from their neutral representation?
- Q2: Are the modulations on the neutral embeddings due to emotion variation larger or smaller than those due to lexical variation?

Answering these two questions is necessary not only for understanding how emotions affect speaker embeddings, but also for assessing their utility as general paralinguistic features in emotion recognition tasks.

We rely on two datasets to pre-train and evaluate our networks. All of the autoencoders that we use in this analysis were first trained on embeddings extracted from the 100-hour clean version of the LibriSpeech dataset and validated on the clean development set of LibriSpeech [44]. The 100-hour clean version of LibriSpeech contains a total of 28,539 utterances from 585 speakers (284 males and 301 females). The pre-training was performed to ensure that the parameters of our autoencoders are properly tuned for encoding and decoding speaker embeddings for general *neutral* speaker population and to provide the same starting point for all speaker-specific autoencoders. We use the VESUS dataset to test emotional similarity because it provides us with the means to control for both emotion and lexical content without compromising the total number of samples available for each speaker [40].

The autoencoders that we use consist of five hierarchical down-sampling stages and five hierarchical up-sampling stages. The hierarchical architecture of our autoencoders is similar to the models used in [28]. Each down-sampling layer in our autoencoders reduces the dimensionality of its input by two while each up-sampling layer increases the dimensionality of its inputs by two. This reduces the effective size of speaker embeddings to 16 features from their original 512 features before being up-sampled. Each block (but the last) in our autoencoders consists of a fully-connected layer followed by a *Tanh* activation. The last block only includes a fully-connected layer with no activation units. We use the mean squared error (MSE) loss function and train our autoencoders using the ADAM optimizer with a learning rate of 0.001 and batch sizes of 256. We run the training for a total of 100 epochs and apply early stopping once the validation loss does not improve for five consecutive epochs. For fine-tuning, we use a batch size of 32 and use the same learning rate and loss functions used for training the autoencoders. We run the fine-tuning for a total of 50 epochs and apply early stopping once the loss on a held-out validation set does not improve for five consecutive epochs.

5.1.1 Question 1 Experimental Setup

We first pre-train the autoencoders using neutral speech from the LibriSpeech corpus. Then, for each speaker in our test corpus (VESUS), we partition their data into three categories: (1) neutral training, (2) neutral testing, and (3) emotional testing. We use the neutral training data (which consist of 70% of a speaker's total neutral data) to fine-tune the autoencoder for each speaker. This allows us to construct a *baseline* model for each speaker. We then create a distribution using the reconstruction error associated with the neutral testing data and compare the reconstruction errors obtained from the emotional testing data to this distribution. If, in general, the reconstruction error on the neutral speech is lower than that of the emotional speech, this will support the hypothesis that embeddings are modulated by emotion.

We analyze the effect emotion has on the reconstruction errors using linear mixed effect models (LMEMs), implemented via the $lme4$ package [45] in R [46]. We set the reconstruction error as a response variable in our linear models and set the emotion (*neutral* vs. *non-neutral*) and the gender as dependent binary variables. We set random intercepts for speaker_ids and utterance_duration (discretized into 3-quantiles), as well as per-speaker random slopes. In case the linear model fails to converge, we simplify the model by removing the per-speaker random slope and only retain the random intercepts, as suggested in [47]. We use likelihood ratio tests to test for statistical significance and test a full model (with the emotion fixed effect) against a null model (without the emotion fixed effect).

5.1.2 Question 2 Experimental Setup

We first pre-train the autoencoders using neutral speech from the LibriSpeech corpus. Then, for each speaker in our test corpus (VESUS), we partition their data into four categories: (1) neutral training, (2) neutral testing-a, (3) neutral testing-b, and (4) emotional testing. Further, we filter utterances in partition (4) such that we only retain those that can be matched based on lexical content with utterances in partition (2). Note that due to the lexically controlled nature of VESUS, the lexical content of each utterance is unique. As a result, each neutral partition contains utterances with unique content. We use the data in the neutral training partition to fine-tune the autoencoder for each speaker, allowing us to construct a *baseline* model for each speaker. We then create a distribution using the reconstruction error associated with the neutral testing-a data, and compare the reconstruction error of the neutral testing-b and emotional testing data to this distribution. If the difference in reconstruction errors between partition (2) and partition (4) is bigger than the error between and partition (2) and partition (3), then this will support the hypothesis that modulation on the speaker embeddings due to variations in emotion are larger than those due to variations in lexical content.

We run a series of LMEMs to analyze the effect of emotion variation and lexical content variation on reconstruction errors. We set the reconstruction error as a response variable in our linear models, and set a binary value (i.e., *neutral* vs. *non-neutral with same content* or *neutral* vs. *neutral with different content*) and the gender as dependent binary variables. We set random intercepts for speaker_ids and utterance_duration (discretized into 3-quantiles), as well as per-speaker random slopes. We follow the same process described in Section 5.1.1 to fit the LMEMs and test for significance.

5.2 Experiment 2: Speaker Embeddings as General Paralinguistic Features

While the previous experiment investigates whether or not speaker embeddings capture speech characteristics that are changed by variations in emotion, this experiment investigates if these disturbances can be used for recognizing emotions. We compare the emotion recognition performance obtained with speaker embeddings to the performance obtained with state-of-the-art features used in the literature. We hypothesize that emotion recognizers that use speaker embeddings as features will outperform those that use common features from the paralinguistics literature. Our hypothesis is based on the fact that speaker embeddings are extracted from models that were trained on much bigger and diverse datasets compared to the commonly used emotion features. Specifically, the large and in-the-wild nature of the datasets used for training the speaker recognition models encourages the models to extract robust representations that capture speaker characteristics from a given audio signal, regardless of the acoustic conditions or environmental noise present. This experiment allows us to understand the relationship between the speaker and emotion recognition tasks and helps us assess the prospects of replacing low-level paralinguistic features with speaker embeddings in emotion recognition models.

We compare the performance obtained using the extracted speaker embeddings to baselines obtained from common paralinguistics feature sets. The first category is the same 30-dimensional MFCCs used by the speaker identification model to extract the speaker embeddings. This allows us to ask how the transformation introduced by the speaker embeddings improves our ability to recognize emotion. The second category includes feature sets broadly grouped based on their use of statistics to characterize the patterns in low-level acoustic features. These feature sets include: the INTERSPEECH 2009 (IS09) Emotion Challenge features [1] (384 parameters), the INTERSPEECH 2013 Computational Paralinguistics ChallengE (ComParE) [2] features (6,373 parameters), the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [3] (62 parameters), and the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [3] (88 parameters). The features for all categories were z-normalized using the training set statistics while the utterances for the speaker embeddings were mean-normalized using a three-second window applied to the MFCC features of each utterance.

We assess the utility of the features in emotion recognition by running both within-corpus and cross-corpus recognition experiments with the IEMOCAP and MSP-IMPROV datasets. For the within-corpus experiments, we follow a leave-one-speaker-out evaluation scheme. For the cross-corpus experiments, we train our models using the labeled samples from one dataset and evaluate on the other dataset (and vice versa). The cross-corpus setup limits the effect of spurious correlations that a trained model can use in the evaluation process. We use unweighted average recall (UAR), which takes an average of the recall of each emotion class, as our evaluation metric in this experiment. This metric allows us to account for the class imbalance in the two datasets we use in this experiment.

The emotion recognition model that we use is based on deep neural networks (DNNs) as previous research has demonstrated their effectiveness when used with state-of-the-art feature sets [48], [49], [50]. For the within-corpus experiments, we follow a leave-one-speaker-out evaluation scheme, where for each test speaker, we use the opposite gender speaker from the test speaker's session as our validation speaker. For the cross-corpus experiments, we use the two speakers from the last session (i.e., session five for IEMOCAP and session six for MSP-IMPROV) as our validation speakers. The hyper-parameters for our DNNs include the number of hidden layers $\{1,2,3\}$ and the width of each hidden layer $\{128,256,512\}$. We use ReLU activation units in all of our experiments. The networks were trained using the ADAM[2] optimizer with a learning rate 10^{-4} on batches with 32 samples. We assign weights to our training samples according to the inverse of their respective frequencies in the training sets and train the models using a weighted cross-entropy loss function for a total of 100 epochs. We use the held-out validation set for hyper-parameter selection and early stopping. We apply the model that yields the highest validation performance to the unseen test data and report the test performance. Finally, we run each setup 30 times to account for variance from random initialization and training.

5.2.1 Additional Baselines

We compare the performance of emotion recognizers that use speaker embeddings to the performance of recognizers from the literature. The baselines that we use are grouped into four categories: end-to-end models [29], [51], convolutional/recurrent models [24], [52], [53], transfer learning models [54], [55], and feature engineering models [56]. For each setup, we indicate: (1) the normalization technique that was used; (2) if data augmentation was used; and (3) if we re-implement the setup or report the performance as indicated by the respective authors. For the setups that we re-implement, we follow the same training and testing protocol described above. For setups that utilize linear support vector machine (SVM) classifiers, we optimize the complexity hyper-parameter using validation data, $C \in \{10^{-5}, 10^{-4}, \ldots, 10^{1}\}$.

6 Results

6.1 Experiment 1: Speaker Embeddings and Emotions

In this experiment, we study how emotion modulates speaker embeddings, measured in terms of reconstruction error. Smaller reconstruction errors indicate that the samples are more similar to the baseline distribution of *neutral* utterances while bigger reconstruction errors indicate that the samples are different from the baseline distribution. We will treat evidence of emotion-centric modulation, measured by reconstruction error, as evidence of the utility of embeddings for emotion recognition.

6.1.1 Question 1 Results

Our first question asked whether or not variations in emotion significantly modulate speaker embeddings from their *neutral* representation. We find that variations in emotion significantly modulate speaker embeddings from their *neutral* representation. In addition, we find that these modulations are consistent across male and female speakers. Figure 1a shows the reconstruction errors associated with 3,032 utterances, grouped by *intended* emotions (758 *neutral*, 758 *happy*, 758 *angry*, 758 *sad*).

2. Default parameters were used ($\alpha = 0.0001, \beta_1 = 0.9, \beta_2 = 0.999$)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3086050, IEEE Transactions on Affective Computing

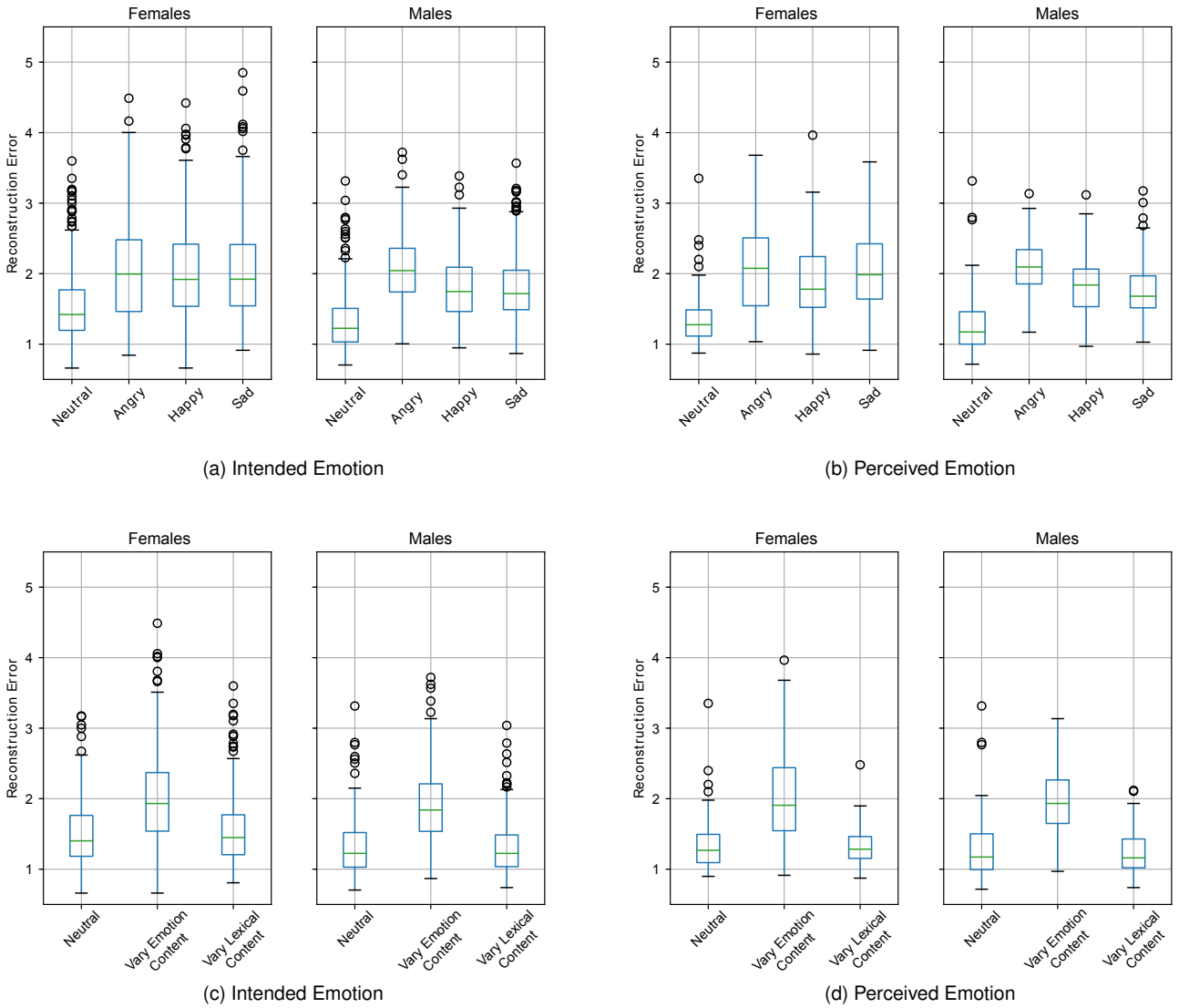


Fig. 1: Reconstruction errors obtained from autoencoders trained with embeddings extracted from *neutral* utterances. Sub-figures (a) and (b) show the reconstruction errors grouped by emotion (*neutral*, *angry*, *happy*, *sad*) and gender (females, males). Sub-figures (c) and (d) compare the reconstruction errors obtained from *neutral* utterances to those obtained from *emotional* utterances with lexical content fixed, and to those obtained from *neutral* utterances but with different lexical content.

Figure 1b shows the reconstruction errors associated with 752 utterances, grouped by *perceived* emotions (188 *neutral*, 188 *happy*, 188 *angry*, 188 *sad*). The perceived emotions group includes utterances whose labels achieved at least 50% agreement between the intended and perceived emotions.

Figure 1a. We find that the reconstruction error was significantly increased by 0.667±0.103 when moving from neutral speech to angry speech ($\chi^2(1)$=16.475, p=4.929e-05). Similarly, we find that the reconstruction error was significantly increased by 0.458±0.070 when moving from neutral speech to happy speech ($\chi^2(1)$=16.760, p=4.242e-05). Finally, we find that the reconstruction error was significantly increased by 0.490±0.071 when moving from neutral speech to sad speech ($\chi^2(1)$=17.560, p=2.784e-05). If we use the reconstruction error as a feature and apply a threshold to separate neutral and non-neutral speech, then we obtain an Area Under the Receiver Operating Characteristic curve (AUC) of 0.782. This indicates that the reconstruction errors obtained from an autoencoder that was exclusively trained on neutral speech can be used for detecting non-neutral speech.

Figure 1b. We find that the reconstruction error was significantly increased by 0.645±0.126 when moving from neutral speech to angry speech ($\chi^2(1)$=11.683, p=6.307e-4). Similarly, we find that the reconstruction error was significantly increased by 0.458±0.070 when moving from neutral speech to happy speech ($\chi^2(1)$=11.871, p=5.703e-4). Finally, we find that the reconstruction error was significantly increased by 0.548±0.080 when moving from neutral speech to sad speech ($\chi^2(1)$=14.606, p=1.325e-4). If we use the reconstruction error as a feature and apply a threshold to separate neutral and non-neutral speech, then we obtain an AUC of 0.850. Again, demonstrating the utility of this setup for speech novelty detection applications.

6.1.2 Question 2 Results

Our second question asked whether the modulations due to emotion are larger or smaller compared to those due to lexical variation. We find that variations in the lexical content have a non-significant effect on neutral embeddings, compared to the significant effect observed in the emotional utterances. We compare the reconstruction errors in three partitions of data as described in Section 5.1: (1) control neutral; (2) non-neutral with fixed lexical content; and (3) neutral with varying lexical content.

Figures 1c and 1d show the reconstruction errors, grouped by the aforementioned three partitions, associated with 1,892 and 458 utterances, respectively. Figure 1c displays the results obtained with intended emotion labels while Figure 1d displays the results obtained with perceived emotion labels. As before, the perceived emotions group includes utterances whose labels achieved at least 50% agreement between the intended and perceived emotions.

Figures 1c. We find that the reconstruction error was significantly increased by 0.529±0.0253 when moving from neutral speech to non-neutral speech while keeping content fixed ($\chi^2(1)$=384.450, p<2.2e-16). If we use the reconstruction error as a feature and apply a threshold to separate neutral and non-neutral speech, we obtain an AUC of 0.785. In contrast, we find that the reconstruction error does not significantly change when varying lexical content while keeping emotion fixed.

Figures 1d. We find that the reconstruction error was significantly increased by 0.649±0.091 when moving from neutral speech to non-neutral speech while keeping content fixed ($\chi^2(1)$=14.398, p=1.480e-4). If we use the reconstruction error as a feature and apply a threshold to separate neutral and non-neutral speech, then we obtain an AUC of 0.840. We again find that the reconstruction error does not significantly change when varying lexical content but fixing emotion to neutral.

6.1.3 Experiment 1 Discussion

The findings from this experiment suggest that while neutral speaker embeddings may be invariant to modulations due to variations in lexical content, they are significantly changed by variations in emotions. We find that using utterances with majority emotion agreement yields smaller overlaps between the interquartile ranges (IQRs) of reconstruction errors obtained from the neutral and emotional utterances, for both female and male speakers, compared to those obtained when using all utterances (i.e., intended emotions). One explanation for this is that the intended emotion labels are more subtle than the perceived emotion labels that we use in this work. As a result, we see more pronounced modulations with the perceived labels compared to the modulations we see when using the intended labels. Finally, we note that none of the models we ran yielded significant interaction between emotion and gender, suggesting that the increases in reconstruction error per emotion (for both intended and perceived) are consistent across female and male speakers.

The findings suggest that speaker embeddings can be used for establishing a baseline of how an individual sounds in their neutral state (i.e., normal behavior). Then, disturbances to this speaker model can be used as a proxy for measuring deviations from this normal behavior. This property of speaker embeddings can be beneficial in applications where we have ample baseline data from a speaker in the neutral state, but have limited or no access to outlier or novel data points from the speaker in certain states (e.g., road rage detection applications in vehicles). In the next experiment, we test if we can utilize these observed modulations in speaker embeddings for detecting emotions in challenging settings.

6.2 Experiment 2: Speaker Embeddings as General Paralinguistic Features

In the first experiment, we studied the relationship between emotion and speaker embeddings. In this section, we compare speaker embeddings to state-of-the-art paralinguistic features on the task of emotion recognition. We first demonstrate the relative ability of embeddings, compared to conventional speech emotion features, in a within-corpus experiment. We then repeat the analysis in a cross-corpus experiment. In both cases, we assess the efficacy of the feature sets on the IEMOCAP and MSP-IMPROV datasets.

We first compare the emotion recognition performance of different feature sets within-corpus. Overall, we find that speaker embeddings, when used as paralinguistic features, significantly outperform or perform comparably to the baseline features described in Section 5.1. Specifically, we find that speaker embeddings significantly outperform all baselines in the within-corpus setup on the MSP-IMPROV dataset; and we find that speaker embeddings only significantly outperform MFCCs, GeMAPS, and eGeMAPS baselines in the within-corpus setup on the IEMOCAP dataset (Table 2).

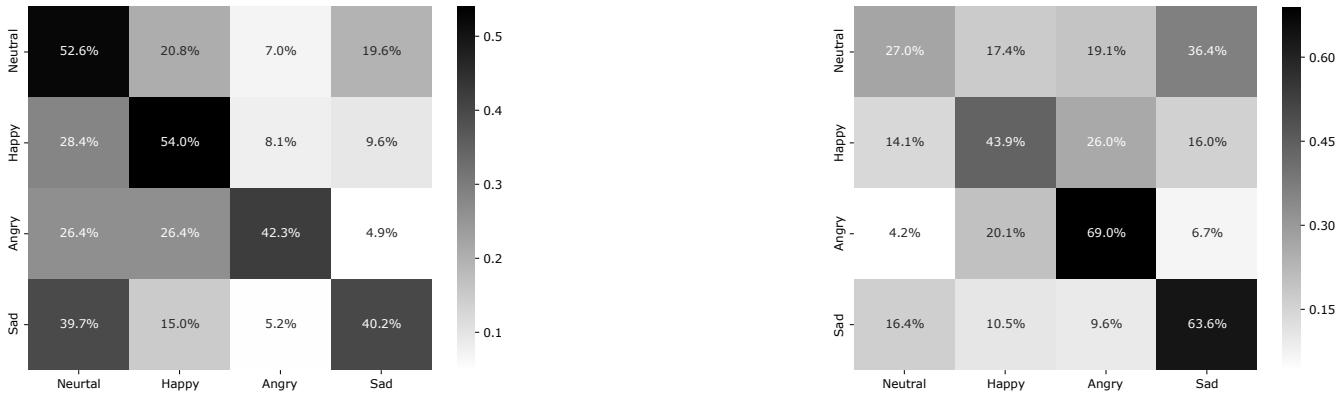
Next, we analyze the performance of these feature sets in a more challenging cross-corpus task. We find that speaker embeddings significantly outperform all other features when evaluating the models on the IEMOCAP and MSP-IMPROV datasets (Table 2). In addition, we observe a higher test performance when we test on the IEMOCAP dataset than we do when we test on the MSP-IMPROV dataset. Among the baselines, we find that the ComParE feature set outperforms all other baselines on the IEMOCAP dataset but performs comparably to IS09 and GeMAPS on the MSP-IMPROV dataset. The results suggest that the embeddings are more robust to domain-shifts than baseline features.

Figure 2 shows the confusion matrices obtained when using speaker embedding in cross-corpus emotion recognition settings. When testing on the MSP-IMPROV corpus, we find that the performance of detecting the *neutral* and *happy* emotions is higher than the performance of detecting the *angry* and *sad* emotions. In contrast, when testing on the IEMOCAP corpus, we find that the performance of detecting the *angry* and *sad* emotions is higher than the performance of detecting the *neutral* and *happy* emotions. The trends displayed by the confusion matrix in Figure 2b agree with the trends we saw in Figures 1a and 1b. Specifically, the confusion matrix in Figure 2b shows that we obtain the highest performance when detecting the angry emotion, followed by both the *sad* and *happy* emotions. However, the confusion matrix in Figure 2a shows that the *happy* emotion is the easiest to detect, followed by the *angry* and *sad* emotion. Finally, we find that the improvements gained by using speaker embeddings over ComParE features cannot be attributed to the improvement in recognizing a specific emotion, but instead, can be attributed to a consistent improvement across all emotions.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2021.3086050, IEEE Transactions on Affective Computing

TABLE 2: The unweighted average recall (UAR) obtained for each setup in the within-corpus and cross-corpus experiments. MSP and IEM denote the MSP-IMPROV and IEMOCAP dataset, respectively. Models in the within-corpus experiments are evaluated following a leave-one-speaker-out evaluation scheme. MSP under cross-corpus indicates the performance of a model that is trained on IEMOCAP and evaluated on MSP-IMPROV; IEM under cross-corpus indicates the performance of a model that is trained on MSP-IMPROV and evaluated on IEMOCAP. The results shown are averages (±1 standard deviation) from 30 runs with different random seeds. The best result in each experiment is **bolded**. † indicates that the marked performance is significantly higher than all baselines; ∗ indicates that the marked performance is significantly higher than MFCCs, GeMAPS, and eGeMAPS; Significance is assessed at $p < 0.05$ using the Tukey's honest test on the ANOVA statistics.

Features	**Within-corpus UAR (%)**		**Cross-corpus UAR (%)**	
	MSP	**IEM**	**MSP**	**IEM**
Chance	25.0	25.0	25.0	25.0
MFCCs	45.8 (±0.9)	52.3 (±0.9)	39.2 (±2.7)	43.7 (±3.1)
IS09	47.7 (±0.9)	58.1 (±0.6)	42.1 (±0.9)	43.7 (±2.7)
ComParE	49.0 (±1.1)	**58.2 (±0.7)**	42.0 (±1.1)	48.6 (±3.0)
GeMAPS	45.6 (±0.8)	56.3 (±0.6)	42.2 (±1.1)	38.7 (±2.2)
eGeMAPS	47.5 (±0.8)	57.2 (±0.7)	39.9 (±1.3)	35.9 (±3.1)
Embeddings (this work)	**50.0 (±1.2)**†	57.9 (±1.0)*	**47.3 (±2.1)**†	**50.9 (±2.1)**†



(a) MSP-IMPROV (b) IEMOCAP

Fig. 2: Confusion matrices obtained using speaker embeddings in the cross-corpus setting when (a) training on IEMOCAP and testing on MSP-IMPROV; (b) training on MSP-IMPROV and testing on IEMOCAP.

6.2.1 Additional Baselines

In Table 3, we compare the performance of emotion recognizers that use speaker embeddings to the performance of recognizers from the literature.

End-to-end Models. Our proposed models that use speaker embeddings outperform end-to-end models from Sarma et al. [29] and Latif et al. [51] for the within-corpus evaluation setting. Both of these end-to-end models were trained to detect emotion from raw speech, obviating the need for the manual definition and extraction of acoustic features and enabling the extraction of features that are optimized for the end task in a data-driven way. Our speaker embeddings have the advantage of being trained on large and diverse data, making the embeddings better at capturing the characteristics of speakers compared to features extracted by the end-to-end models that were only trained on emotion data.

Convolutional/Recurrent Models. Our approach outperforms the convolutional-recurrent model (specifically the Vanilla model) proposed by Huang and Narayanan [52] for all within-corpus and cross-corpus setups. Our approach, however, performs comparably to that of Li et al. [53] for the within-corpus evaluation setup on IEMOCAP. The good performance obtained by Li et al. could be attributed to the use of data augmentation when training the emotion classifiers. In contrast, we do not use any data augmentation when training our emotion classifiers. Finally, our approach performs comparably to the approach of Aldeneh and Mower Provost [24] for the within-corpus evaluation setup on MSP-IMPROV but underperforms their approach for the within-corpus evaluation setup on IEMOCAP. We note that the good performance obtained by the authors in [24] is due to the use of speaker identity information during the evaluation through the use of speaker-level normalization. In contrast, we make no assumption about speaker identity in this experiment and only perform utterance-level normalization.

Transfer Learning Models. Our approach outperforms the approach proposed by Neumann and Vu [55] for the cross-corpus evaluation setup on MSP-IMPROV. In their approach, Neumann and Vu integrate features learned by an unsupervised autoencoder that is trained on out-of-domain unlabeled speech data (i.e., Tedlium 2) into a convolutional emotion classifier. Specifically, the features extracted by the unsupervised autoencoder are provided to the emotion classifier as an additional input during the training and testing phases. The results demonstrate that embeddings which were extracted from a speaker recognizer are better suited for emotion classification

TABLE 3: The unweighted average recall (UAR) obtained or reported for each baseline setup in the within-corpus and cross-corpus experiments. MSP and IEM denote the MSP-IMPROV and IEMOCAP dataset, respectively. Models in the within-corpus experiments are evaluated following a leave-one-speaker-out evaluation scheme. MSP under cross-corpus indicates the performance of a model that is trained on IEMOCAP and evaluated on MSP-IMPROV; IEM under cross-corpus indicates the performance of a model that is trained on MSP-IMPROV and evaluated on IEMOCAP. When applicable, we report the standard deviation from 30 runs with different random seeds. The best result in each experiment is **bolded**. "—" indicates unavailability or inapplicability.

Method	Normalization	Augmentation?	Re-implemented?	Within-corpus UAR (%)		Cross-corpus UAR (%)	
				MSP	IEM	MSP	IEM
Chance	—	—	—	25.0	25.0	25.0	25.0
Sarma et al. [29]	—	✗	✗	—	48.8	—	—
Latif et al. [51]	—	✗	✗	48.5	56.7	—	—
Huang & Narayanan [52]	Utterance-level	✗	✓	39.1 (±1.2)	49.7 (±0.8)	37.8 (±1.8)	36.5 (±3.4)
Li et al. [53]	None	✓	✗	—	57.5	—	—
Aldeneh & Mower Provost [24]	Speaker-level	✗	✗	49.8	**59.5**	—	—
Neumann & Vu [55]	—	✗	✗	—	—	45.8	—
Amiriparian et al. [54]	Utterance-level	✗	✓	31.6	45.6	36.8	29.8
Schmitt et al. [56]	Utterance-level	✗	✓	32.7 (±0.6)	49.7 (±0.4)	35.5 (±0.6)	36.0 (±0.8)
Embeddings (this work)	Utterance-level	✗	✓	**50.0 (±1.2)**	57.9 (±1.0)	**47.3 (±2.1)**	**50.9 (±2.1)**

than those extracted using an unsupervised autoencoder. Our results show that the *deep spectrum* approach, introduced by Amiriparian et al. [54], under-performs our proposed approach by a large margin on all evaluation setups. Deep spectrum features are obtained by running spectrograms of input speech utterances through convolutional networks, which were originally trained on visual tasks, and extracting the resulting intermediate features. Once extracted, deep spectrum features are used by a linear SVM classifier for emotion classification.

Feature Engineering Models. Our approach outperforms the bag-of-audio-words features approach that was introduced by Schmitt et al. [56]. In the bag-of-audio-words approach, acoustic features (i.e., MFCCs) are first quantized according to a vector codebook of audio words and then added to a histogram of audio words for classification by a linear SVM. We used the random sampling technique for generating the codebooks and we used codebooks of size 512 to be consistent with the size of speaker embeddings.

6.2.2 Experiment 2 Discussion

To the best of our knowledge, this is the first work to compare both within-corpus and cross-corpus emotion recognition performance obtained using speaker embeddings to the performance obtained using general features commonly used in the emotion recognition community. Our results suggest that speaker embeddings are highly versatile, and can easily be adapted to other paralinguistic applications such as emotion recognition. We note that speaker embeddings also provide a more compact alternative to some of the features sets (e.g., ComParE). For example, the ComParE feature set contains 6,373 parameters representing energy, spectral, and voicing features [23]. In contrast, speaker embeddings only contain 512 parameters and are extracted from 30-dimensional MFCCs (i.e., spectral). Our results also demonstrate how emotion classifiers that use speaker embeddings outperform other emotion classification methods from the literature when compared under similar training and evaluation conditions.

7 Discussion and Conclusion

In this paper, we proposed the use of speaker embeddings, representations extracted from neural networks trained on a speaker identification task, as paralinguistic features to be used in emotion recognition applications. Speaker embeddings capture high-level speaker characteristics and abstract extraneous low-level variations in the acoustic signal that are not needed for recognizing speakers. The hypothesis that drove our work is that emotionally charged vocal expressions make speakers sound different from how they typically sound.

We first used autoencoders to quantify the effect of emotion on speaker embeddings. We trained our auto-encoders on neutral speech from each speaker, and used the reconstruction errors obtained for test utterances as a proxy for measuring the effect emotion has on speaker embeddings. Our analysis showed that embeddings extracted from expressive speech resulted in significantly increased reconstruction errors compared to neutral speech. In addition, our analysis showed that lexical variation had a non-significant effect on the reconstruction errors obtained from the utterances. Our experiments also demonstrated how the reconstruction errors obtained from the autoencoders can be used as features for detecting deviations from the neutral state. Future work will study techniques for making changes in emotions more pronounced while maintaining speaker discriminative properties in speaker embeddings (e.g., emotion-invariant x-vectors).

We then showed that speaker embeddings can be used as a replacement to common paralinguistic features used in emotion recognition tasks. We demonstrated this by showing not only that speaker embeddings outperform baseline features in cross-corpus emotion recognition tasks, but also that they are more compact (i.e., fewer parameters) than state-of-the-art paralinguistic features. Speaker embeddings outperformed other features despite being extracted from spectral representations (i.e., MFCCs) alone. In contrast, other features used a combination of energy, voicing, and spectral representations. MFCCs used for extracting speaker embeddings were originally designed based on observations from perceptual experiments and thus, may not be optimal for all speech applications. For example, MFCC features smooth the speech spectrum and make it difficult to extract other narrow-band information that is known to be predictive of emotion (e.g., pitch, formants). One extension to the current approach is to train the speaker identification models with representations from which this fine-

grained information is easily extractable (e.g., spectrograms, raw waveform). Another extension to the current approach is to combine speaker embeddings with common emotion features to provide the recognizer access to the fine-grained information present in the speech signal.

In conclusion, this work further contributed to our understanding of the relationship between emotions and speaker representations and demonstrated how variations in emotion manifest themselves in speaker embeddings. These manifestations not only can impact the performance of a verification system, but also can be leveraged for detecting emotions.

ACKNOWLEDGMENTS

This material is based in part upon work supported by the Toyota Research Institute ("TRI"), the IBM PhD Fellowship Award, and by the National Science Foundation (NSF CAREER 1651740). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF, IBM, TRI, or any other Toyota entity. The authors thank Mohamed El Banani for suggesting the title for this work. The authors also thank members of the CHAI lab at the University of Michigan for their advice on this project.

REFERENCES

[1] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Interspeech*, 2009.

[2] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Interspeech*, 2013.

[3] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *Transactions on Affective Computing*, 2016.

[4] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014.

[5] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in neural information processing systems*, 2014.

[6] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps, and B. W. Schuller, "Multi-task semi-supervised adversarial autoencoding for speech emotion recognition," *IEEE Transactions on Affective Computing*, 2020.

[7] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E. M. Provost, "Progressive neural networks for transfer learning in emotion recognition," in *Interspeech*, 2017.

[8] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The speakers in the wild (SITW) speaker recognition database." in *Interspeech*, 2016.

[9] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *Interspeech*, 2017.

[10] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Interspeech*, 2018.

[11] W. Wu, T. F. Zheng, M.-X. Xu, and H.-J. Bao, "Study on speaker verification on emotional speech," in *Ninth International Conference on Spoken Language Processing*, 2006.

[12] S. Parthasarathy and C. Busso, "Predicting speaker recognition reliability by considering emotional content," in *International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017.

[13] S. Parthasarathy, C. Zhang, J. H. Hansen, and C. Busso, "A study of speaker verification performance with expressive speech," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[14] M. Bancroft, R. Lotfian, J. Hansen, and C. Busso, "Exploring the intersection between speaker verification and emotion recognition," in *International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 2019.

[15] J. H. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal processing magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[16] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[17] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker diarization with LSTM," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[18] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[19] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[20] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "x-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[21] J.-A. Bachorowski and M. J. Owren, "Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context," *Psychological science*, vol. 6, no. 4, pp. 219–224, 1995.

[22] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech communication*, vol. 40, no. 1-2, pp. 227–256, 2003.

[23] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. v. Son, F. Weninger, F. Eyben, T. Bocklet *et al.*, "The interspeech 2012 speaker trait challenge," in *Interspeech*, 2012.

[24] Z. Aldeneh and E. M. Provost, "Using regional saliency for speech emotion recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[25] S. Mariooryad and C. Busso, "Compensating for speaker or lexical variabilities in speech for emotion recognition," *Speech Communication*, vol. 57, pp. 1–12, 2014.

[26] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification." in *Interspeech*, 2017.

[27] Z. Meng, J. Li, Z. Chen, Y. Zhao, V. Mazalov, Y. Gang, and B.-H. Juang, "Speaker-invariant training via adversarial learning," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[28] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Learning representations of affect from speech," in *ICLR (Workshop Track)*, 2015.

[29] M. Sarma, P. Ghahremani, D. Povey, N. K. Goel, K. K. Sarma, and N. Dehak, "Emotion identification from raw speech signals using dnns." in *Interspeech*, 2018.

[30] B. Zhang, Y. Kong, G. Essl, and E. M. Provost, "f-similarity preservation loss for soft labels: A demonstration on cross-corpus speech emotion recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[31] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Workshop on Automatic Speech Recognition and Understanding*, 2013.

[32] D. Raj, D. Snyder, D. Povey, and S. Khudanpur, "Probing the information encoded in x-vectors," in *Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019.

[33] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[34] A. McCree, G. Sell, and D. Garcia-Romero, "Speaker diarization using leave-one-out gaussian PLDA clustering of dnn embeddings." in *Interspeech*, 2019.

[35] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011.

[36] J. Williams and S. King, "Disentangling style factors from speaker representations," in *Interspeech*, 2019.

[37] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[38] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interac-

tions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2016.
[39] B. Zhang, S. Khorram, and E. M. Provost, "Exploiting acoustic and lexical properties of phonemes to recognize valence from speech," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.
[40] J. Sager, R. Shankar, J. Reinhold, and A. Venkataraman, "Vesus: A crowd-annotated database to study emotion production and perception in spoken english." in *Interspeech*, 2019.
[41] E. Marchi, F. Vesperini, F. Eyben, S. Squartini, and B. Schuller, "A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional lstm neural networks," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.
[42] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, "Deep autoencoding gaussian mixture model for unsupervised anomaly detection," in *ICLR*, 2018.
[43] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
[44] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.
[45] D. Bates, M. Maechler, B. Bolker, S. Walker, R. H. B. Christensen, H. Singmann, B. Dai, G. Grothendieck, P. Green, and M. B. Bolker, "Package âĂŸlme4âĂŹ," *Convergence*, vol. 12, no. 1, pp. 470–474, 2015.
[46] R. C. Team *et al.*, "R: A language and environment for statistical computing," 2013.
[47] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily, "Random effects structure for confirmatory hypothesis testing: Keep it maximal," *Journal of memory and language*, vol. 68, no. 3, pp. 255–278, 2013.
[48] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning." in *Interspeech*, 2017, pp. 1103–1107.
[49] M. Abdelwahab and C. Busso, "Study of dense network approaches for speech emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5084–5088.
[50] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020.
[51] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct modelling of speech emotion from raw speech," in *Interspeech*, 2019.
[52] C.-W. Huang and S. S. Narayanan, "Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition," in *International Conference on Multimedia and Expo (ICME)*, 2017.
[53] H. Li, M. Tu, J. Huang, S. Narayanan, and P. Georgiou, "Speaker-invariant affective representation learning via adversarial training," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.
[54] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. Schuller, "Snore sound classification using image-based deep spectrum features," in *Interspeech*, 2017.
[55] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.
[56] M. Schmitt, F. Ringeval, and B. W. Schuller, "At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech." in *Interspeech*, 2016.

Zakaria Aldeneh is a Ph.D. candidate working with Professor Emily Mower Provost in Computer Science and Engineering at the University of Michigan, Ann Arbor. He received his B.S. in Electrical Engineering from the University of Cincinnati in 2014, and his M.S. in Computer Science and Engineering from the University of Michigan in 2016. He was selected to receive the IBM Ph.D. Fellowship in 2018. He was also an intern with the Apple Siri speech team for the summers of 2018 and 2019. Zakaria is currently interested in affective computing using acoustic signal processing and machine learning methods.

Emily Mower Provost is an Associate Professor in Computer Science and Engineering at the University of Michigan. She received her Ph.D. in Electrical Engineering from the University of Southern California (USC), Los Angeles, CA in 2010. She is a Toyota Faculty Scholar (2020) and has been awarded a National Science Foundation CAREER Award (2017), the Oscar Stern Award for Depression Research (2015), a National Science Foundation Graduate Research Fellowship (2004-2007). She is a co-author on the paper, "Say Cheese vs. Smile: Reducing Speech-Related Variability for Facial Emotion Recognition," winner of Best Student Paper at ACM Multimedia, 2014, and a co-author of the winner of the Classifier Sub-Challenge event at the Interspeech 2009 emotion challenge. Her research interests are in human-centered speech and video processing, multimodal interfaces design, and speech-based assistive technology. The goals of her research are motivated by the complexities of the perception and expression of human behavior.