Learning Paralinguistic Features from Audiobooks through Style Voice Conversion

Zakaria Aldeneh, Matthew Perez, Emily Mower Provost
University of Michigan, Ann Arbor
{aldeneh,mkperez,emilykmp}@umich.edu

Abstract

Paralinguistics, the non-lexical components of speech, play a crucial role in human-human interaction. Models designed to recognize paralinguistic information, particularly speech emotion and style, are difficult to train because of the limited labeled datasets available. In this work, we present a new framework that enables a neural network to learn to extract paralinguistic attributes from speech using data that are not annotated for emotion. We assess the utility of the learned embeddings on the downstream tasks of emotion recognition and speaking style detection, demonstrating significant improvements over surface acoustic features as well as over embeddings extracted from other unsupervised approaches. Our work enables future systems to leverage the learned embedding extractor as a separate component capable of highlighting the paralinguistic components of speech.

1 Introduction

An effective speech-based AI system is capable of not only recognizing and interpreting the linguistic content of speech but also recognizing and interpreting its paralinguistic attributes. While the linguistic elements of speech encode *what* was said (i.e., the content), the paralinguistic elements encode *how* it was said (i.e., emotion, style, etc.) (Schuller and Batliner, 2013). The detection and modeling of paralinguistic attributes have many potential applications; ranging from affect-aware Human-Computer Interaction (HCI) systems (Vinciarelli et al., 2015) to the management of mental health (Karam et al., 2014; Cummins et al., 2015).

One major challenge with building robust paralinguistic models is the limited access to large-scale, labeled datasets that are needed for training the machine learning models. For instance, a typical emotion dataset (e.g., IEMOCAP) that is used for building paralinguistic models contains around 12 hours of speech while a modern dataset used for building speaker recognition models contains around 2000 hours of speech (Nagrani et al., 2017). It is therefore critical that features used in paralinguistic tasks distill relevant information from the original signal to allow the recognizers to effectively detect the target attributes. With this in mind, new methods that can leverage unlabeled data for distilling paralinguistic information from speech should be explored.

In this work, we introduce the Expressive Voice Conversion Autoencoder (EVoCA), an unsupervised framework that distills paralinguistic attributes from speech without relying on explicit emotion or style labels. EVoCA is designed to enable a neural network to learn what it means for speech to be expressive by treating *expressive* speech as a modulation of *neutral* speech. EVoCA is trained using parallel speech inputs: one expressive and one neutral. However, these types of parallel paralinguistic corpora are not available at scale. To address this, we use a large audiobook corpus (i.e., 200 hours) composed of expressive speech and artificially generate the parallel neutral, non-expressive speech using the available transcriptions (see Figure 1).

We train the EVoCA model to convert between non-expressive synthetic speech and expressive real speech, demonstrating how this conversion yields an embedding that captures paralinguistic attributes (see Figure 2). The benefit of the EVoCA framework is that once trained, the component responsible for producing the paralinguistic embeddings can be used as a front-end speech transformer for a variety of downstream applications. We show that these learned paralinguistic embeddings can be used in downstream emotion recognition and speaking style classification tasks.

In summary, the key contributions of this work are the following:

- We present the EVoCA framework for learning speech emotion and style embeddings

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4736–4745
June 6–11, 2021.

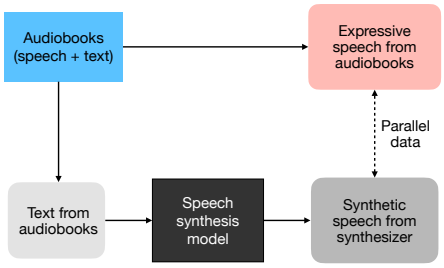


Figure 1: An overview of the parallel data generation process. We use a speech synthesis model to generate a synthetic version of each audio sample in the original audiobook corpus. Synthesized samples lose paralinguistic attributes present in the original samples but retain linguistic information. Our goal is to leverage the resulting real/synthetic sample pairs to learn to extract paralinguistic features.

from audiobooks without relying on manual annotations for those attributes.

- We show that the EVoCA framework learns embeddings that outperform those obtained using other unsupervised and self-supervised speech feature learning methods from the literature.

To the best of our knowledge, ours is the first work to demonstrate how one can learn paralinguistic features by training a neural model to convert between non-expressive synthetic speech and expressive real speech.

2 Related Work

Speech emotion recognition applications rely on an extensive set of acoustic features that have evolved over the years (Schuller et al., 2009, 2010, 2011, 2013; Eyben et al., 2015). Spectral features are a crucial component of any emotion feature set and are included in the widely used ComParE and eGeMAPs feature sets (Schuller et al., 2013; Eyben et al., 2015). Common surface features that are derived from the speech spectrum include Mel-frequency cepstral coefficients (MFCCs) and Mel-filterbanks (MFBs). In this work, we propose a framework for learning an MFB transformation that highlights the paralinguistic content of an utterance; we demonstrate the effectiveness of the learned transformation over surface MFB features on emotion and speaking style classification tasks.

Our work also explores the utility of using both synthetic and real speech to learn paralinguistic information. Lotfian and Busso have previously demonstrated how speech synthesizers can be used to remove emotion from a speech utterance to provide trained emotion recognizers with a neutral reference to aid in the recognition of expressive speech (Lotfian and Busso, 2015). One limitation with their approach, however, is that it relied on having access to a real-time speech synthesizer to generate a neutral version of the input utterance for use by the emotion recognizer. In contrast, we use the speech synthesizer only during the data preparation process (Figure 1) and not during test time.

Our approach is related to works that focused on unsupervised and self-supervised speech representation learning. Chung et al. introduced two auto-regressive methods for learning MFB transformations for speech applications without relying on explicit labels (Chung et al., 2019). Both of the proposed models were trained to predict future frames of the input speech sequence in order to learn global structures represented in the speech signal. They showed that the resulting transformation improved performance over surface features on speaker verification and phone recognition tasks. Hsu et al. devised a variational autoencoder that is capable of learning hierarchical information present in speech data (Hsu et al., 2017). Their approach disentangled frame-level features from utterance-level features in order to provide robust embeddings for both speaker recognition and automatic speech recognition tasks. Although several unsupervised learning strategies exist for learning speech transformations, ours is the only approach that is targeted at learning transformations that highlight expressive characteristics in speech.

Recent works in voice conversion have also inspired our proposed approach. The goal of voice conversion is to convert an utterance from one speaker so that it sounds as if it was spoken by another speaker (Mohammadi and Kain, 2017). In other words, a voice converter retains all linguistic content and only modulates the paralinguistics of speech. Previous works demonstrated that voice conversion techniques can be used to convert between emotional states (Gao et al., 2019; Shankar et al., 2019a,b). In this work we primarily focus on the use of parallel voice conversion methods and future work will explore the trade-offs between parallel and non-parallel approaches. However, to the best of our knowledge, our work is the first to show that the voice conversion task can be adapted

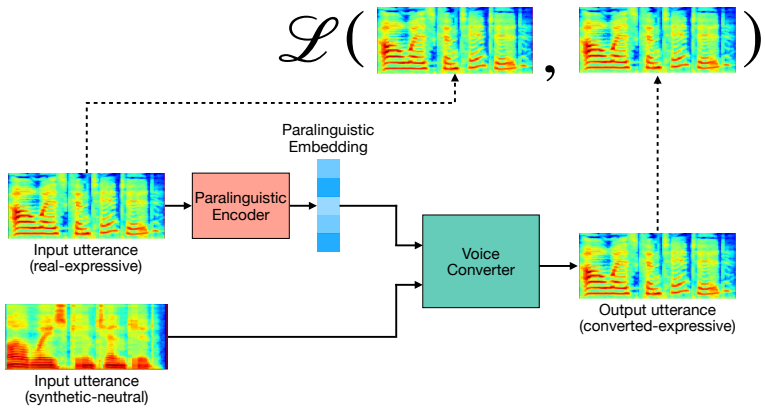


Figure 2: An overview of the proposed Expressive Voice Conversion Autoencoder (EVoCA). The model takes two inputs, the expressive and synthetic speech samples, and outputs the reconstructed expressive speech sample. The paralinguistic encoder extracts an embedding from the expressive speech sample such that it can be used by the Voice Converter to insert paralinguistics into the synthetic speech input sample. The network is trained with an $L2$ loss between the generated expressive sample and the original expressive sample. Once the full model is trained, the paralinguistic encoder is disconnected and used as a general purpose paralinguistic feature extractor.

and incorporated into a framework that enables a neural network to learn compact embeddings that capture speech expressiveness.

3 Approach

3.1 Creating Parallel Data using Speech Synthesis

A sketch of our data generation setup is shown in Figure 1. Given an audiobook corpus, where both speech and text modalities are available, we use the text to create synthetic speech samples using a speech synthesizer. The created synthetic speech should lack expressiveness. This provides our system with the opportunity to learn how to characterize expressiveness and imbue the non-expressive speech with expressive characteristics. We use the open-source Festival toolkit[1], as previous research has demonstrated its utility for generating neutral, non-expressive speech (Lotfian and Busso, 2017). Once the speech synthesis process finishes, our data now contain pairs of real (expressive) speech and synthetic (neutral non-expressive) speech. Our EVoCA model then leverages the resulting parallel data to learn an embedding transformation that facilitates the conversion from synthetic to real speech without relying on any manual emotion or style labels.

[1] http://festvox.org/festival/

3.2 Expressive Voice Conversion Autoencoder Setup

A sketch of EVoCA is shown in Figure 2. The EVoCA model converts neutral speech to expressive speech. In the process, the paralinguistic encoder learns a compact embedding that encodes paralinguistic elements, including expressiveness. The paralinguistic embedding and the paired synthetic speech sample are fed into the voice converter, which produces expressive speech. A reconstruction loss ($L2$) between the generated expressive speech and the original expressive speech is computed and used to train the style autoencoder in an end-to-end fashion. Once trained, the paralinguistic encoder can be used as a speech transformer to create features that highlight the expressive components of input speech.

4 Datasets, Features, and Metrics

4.1 Datasets

We use four datasets in this work: Blizzard2013, IEMOCAP, MSP-IMPROV, and VESUS. Blizzard2013 is used to train the EVoCA model while the other three datasets are used to test the effectiveness of the learned embeddings on speech emotion recognition and speaking style detection.

Blizzard2013. The Blizzard2013 dataset contains around 200 hours from 55 American English audiobooks read by Catherine Byers. Although other audiobook-based datasets are publicly available, we choose the Blizzard2013 corpus due to its highly expressive and animated nature. This corpus

was used in previous research to model style and prosody in speech synthesis applications (Wang et al., 2018; Zhang et al., 2019c). We use a segmented version of the corpus, which we obtained from the 2013 Blizzard Challenge website.[2]

IEMOCAP. The IEMOCAP dataset was created to explore emotion expression in dyadic interactions (Busso et al., 2008). Pairs of actors, one male and one female, were recorded while interacting in scripted and improvised roles that were designed to elicit emotional expressions. The dataset contains 12 hours of speech from 10 individuals. The recordings from each interaction were manually segmented into utterances based on speaker turns in the dialogue. The resulting utterances were manually labeled by five annotators for both categorical and continuous emotion labels. We only consider utterances that had majority agreement among the annotators and focus on four basic categorical emotions: *happy* (merged with *excited*), *angry*, *neutral*, and *sad*. In addition to emotion labels, the IEMOCAP dataset provides spontaneity labels (acted vs. spontaneous), which we use in our speaking style detection experiments.

MSP-IMPROV. The MSP-IMPROV dataset was created to capture naturalistic expressions from improvised scenarios while partially controlling for variations in the lexical modality (Busso et al., 2016). Similar to IEMOCAP, pairs of actors, one male and one female, were recorded while interacting in improvised scenarios, which included prespecified target sentences that actors were asked to incorporate into their dialogue. The dataset is nine hours in duration from 12 speakers. The resulting utterances were manually labeled for emotion using crowd-sourced annotators. We only consider utterances whose labels had a majority agreement among the annotators and focus on four basic emotion labels: *happy*, *angry*, *neutral*, and *sad*.

VESUS. The VESUS dataset provides around seven hours of lexically-controlled emotional data (Sager et al., 2019). In contrast to IEMOCAP and MSP-IMPROV where emotion elicitation and expression happen in improvised scenarios, actors in the VESUS dataset were asked to read the same set of 250 semantically neutral phrases in five different emotions: *happy*, *angry*, *neutral*, *sad*, and *fearful*. The dataset contains around seven hours of speech from 10 speakers, five males and five females. The resulting utterances were labeled for emotional content by 10 crowd-sourced annotators. In our experiments, we focus on utterances that achieved at least 50% agreement among the crowd-sourced annotators with respect to the actor's intended emotion.

[2] http://www.cstr.ed.ac.uk/projects/blizzard/

4.2 Features

We first pre-process speech samples from all datasets such that they have a sampling rate of 16 kHz and then extract 80-dimensional MFB features using the `Librosa` toolkit (McFee et al., 2015) with a 50 ms Hanning window and a step size of 12.5 ms, consistent with previous research in voice conversion (Zhang et al., 2019a). We z-normalize the frequency bins per utterance for the voice converter and mean-normalize the bins per-utterance for the paralinguistic encoder; consistent with normalization methods used in previous works (Snyder et al., 2018; Zhang et al., 2019c). Normalization ensures that the features are robust to variations that could arise from having different recording conditions (Benesty et al., 2007).

4.3 Tasks

Voice conversion is a regression task where the goal is to output the MFB features of an expressive speech utterance given the MFB features of the synthesized speech utterance. Emotion recognition is posed as a multi-class classification task where the goal is to recognize the target emotion. Lastly, speaking style detection is posed as a binary classification task where the goal is to recognize whether the target data are acted or spontaneous.

4.4 Metrics

We use Mel-cepstral distortion (MCD) and root mean square error (RMSE) of F0 for evaluating the quality of the converted speech (Zhang et al., 2019a) when training the end-to-end model. MCD and F0 RMSE cannot be directly extracted from the MFB acoustic features used by our conversion model. Thus, we use `Librosa` to invert the MFB features to audio by first approximating the Short-time Fourier transform (STFT) magnitude and then using the Griffin-Lim algorithm to reconstruct the phase. We extract the F0 and 24-dimensional mel cepstral coefficients from the waveform using the WORLD vocoder (Morise et al., 2016) following (Zhang et al., 2019a,c).

We use unweighted average recall (UAR) and accuracy for evaluating the performance on the

emotion recognition and speaking style detection tasks, respectively. The UAR metric is used to account for the class imbalance that is inherent in the emotion data (Rosenberg, 2012).

5 Experiments

5.1 Experimental Questions

We design our experiments to address the following four questions regarding the proposed framework shown in Figure 2:

1. Is the proposed framework capable of inserting expressiveness into synthetic speech?

2. Can the learned paralinguistic embeddings be used for emotion and style classification?

3. How do changes to the structure of the proposed framework affect both the quality of the converted speech and the effectiveness of the extracted embeddings for emotion and speaking style detection tasks?

4. How does the performance of paralinguistic embeddings in emotion and speaking style detection tasks compare to those of feature transformations learned using other unsupervised methods?

5.2 Expressive Voice Conversion Autoencoder (EVoCA)

The proposed EVoCA consists of two components: the voice converter and the paralinguistic encoder. The voice converter consists of a stack of four Bidirectional Long Short-Term Memory (BLSTM) layers, each with a hidden size of 256, followed by a 1D convolution layer with 80 channels and a kernel size of one. The paralinguistic encoder we use consists of a stack of two BLSTM layers, each with a hidden size of 256. The fixed-size embeddings from the paralinguistic encoder are induced by taking the mean of the hidden representations from the last BLSTM layer and then passing the outputs through a linear layer, which reduces the size by half. The reasoning for this linear layer is to counteract the bidirectional property of BLSTM which outputs hidden representations that are twice the size of the hidden layer. Our voice converter is inspired by the one used in (Zhang et al., 2019b). However, in this work we utilize a basic version of the model that does not include a two-layer fully connected PreNet, a five-layer 1D convolution PostNet, nor an attention module. We opt to use a simple implementation for voice conversion since our problem does not follow the sequence-to-sequence learning paradigm as our input features are prealigned using dynamic time warping (DTW) (Mohammadi and Kain, 2017). Our final style autoencoder model has approximately 2.2 million parameters.

We investigate how changes to the structure of the proposed EVoCA affect not only the quality of the converted speech, but also the quality of the extracted embeddings. We study the impact that the paralinguistic embedding and synthetic speech have on the voice converter by comparing the voice conversion performance when only one component is present. We also investigate the effect of reducing the capacity (i.e., the number of hidden units) of the paralinguistic encoder and the voice converter on the converted speech as well as on the extracted embeddings for downstream classification tasks. Specifically, we keep the voice converter fixed and reduce the hidden size of the BLSTM paralinguistic encoder gradually from 256 units to 32 units (reducing the number of parameters from 2.2 million to 1.5 million), noting performance changes on the two tasks. Then, we keep the paralinguistic encoder fixed and reduce the hidden size of the BLSTM voice converter from 256 units to 32 units (reducing the number of parameters from 2.2 million to 0.7 million), again noting performance changes on the two tasks. Note that these hyperparameters are not and should not be tuned based on the performance of the downstream task as the goal of this experiment is to analyze how these parameters affect the qualities of the transformed features and the converted speech.

We split the Blizzard2013 data into training, validation, and test partitions following a random 90%-5%-5% split rule. We train our style autoencoder on the training partition and use the validation partition for loss monitoring and early stopping. Conversion performance is reported on the test partition of the data. We construct the network in `PyTorch` and train it from scratch with batches of size 128 using the ADAM optimizer for a total of 80 epochs. We use an initial learning rate of 10^{-4} and decrease it exponentially using a decay factor of 0.95 after each epoch starting from epoch 30. We monitor the validation loss after each epoch and perform early stopping if the validation loss does not improve for

15 consecutive epochs.

5.3 Unsupervised Baselines

The first unsupervised baseline that we consider is a convolutional autoencoder that is applied to fixed-length MFB segments of 128 frames. The autoencoder is similar to the one used in (Eskimez et al., 2018). The encoder consists of three 2D convolution layers, of shape: $[32 \times 9 \times 9]$, $[64 \times 7 \times 7]$, and $[128 \times 5 \times 5]$, followed by a linear layer with 256 units. A $[2 \times 2]$ max pooling operation is applied after each layer to reduce the dimensionality of the input by two. The decoder consists of a linear layer with 256 units followed by four 2D convolution layers of shape: $[128 \times 5 \times 5]$, $[64 \times 7 \times 7]$, $[32 \times 9 \times 9]$, and $[1 \times 1 \times 1]$. A $[2 \times 2]$ nearest neighbor up-sampling operation is applied after each layer to get back the original size of the input. Both the encoder and the decoder use Leaky ReLU activation units and the autoencoder has approximately 3.9 million parameters.

The second unsupervised baseline that we consider is the Autoregressive Predictive Coding (APC) model that was introduced in (Chung et al., 2019). Given an input of MFB features, the APC model is trained to predict the features n time-steps in the future. The APC model that we use is similar to the one used by Chung et al. and it consists of three LSTM layers, each with a width of 512. We run our experiments with three values for n: 5, 10, and 20. Once trained, the outputs from the last LSTM layer are averaged to obtain fixed-size features for downstream tasks. The APC model that we use has approximately 5.5 million parameters.

We train both the autoencoder and the APC baselines on the Blizzard2013 dataset. We use the same protocol we use for training EVoCA when training the autoencoder baseline. However, we train the APC baselines for 100 epochs following the authors' recommendation.

5.4 Emotion and Speaking Style Recognition

We test the utility of the learned paralinguistic encoder for transforming MFB features to highlight their paralinguistic attributes in emotion recognition and speaking style detection tasks. First, we assess if transforming MFB features provides any advantage over using surface MFB features on the two tasks. Then, we compare the learned feature transformation to those obtained using the unsupervised and supervised baselines.

Table 1: Objective performance measures for the style voice conversion task with different setups. The base EVoCA consists of a 256-dimensional paralinguistic encoder and a 256-dimensional voice converter. Reference numbers are computed using the synthetic speech and ground-truth expressive speech. All other numbers are computed using converted speech and ground-truth expressive speech.

Setup	**MCD (dB)**	F_0 **RMSE (Hz)**
Reference	24.01	146.20
Base EVoCA	10.71	64.36
w/o synth. ref.	+8.33	+106.23
w/o para. enc.	+1.90	+79.50
w/ 128-dim para. enc.	+0.31	+6.14
w/ 64-dim para. enc.	+0.69	+19.41
w/ 32-dim para. enc.	+0.97	+31.06
w/ 128-dim converter	+1.03	+15.60
w/ 64-dim converter	+1.77	+31.73
w/ 32-dim converter	+2.61	+61.82

We follow a leave-one-speaker-out evaluation scheme and report the average performance across all test speakers on all four downstream tasks. For each test speaker, we pick the model that gives the best performance on a held-out validation set. The hyper-parameters that we optimize on the validation set include the number of hidden layers $\{1, 2, 3\}$, the width of each hidden layer $\{64, 128, 256\}$, and the activation unit {Tanh, ReLU}. We construct the networks in `PyTorch` and train them with batches of size 32 using the ADAM optimizer with learning rate of 10^{-4} and a cross-entropy loss function. We train each model for a maximum of 100 epochs and apply early stopping if the validation loss does not improve for five consecutive epochs. We repeat each experiment with 30 different random seeds and report the average and standard deviation to account for performance fluctuation due to random initialization and training.

6 Results

In this section, we provide the results of our four experiments (Section 5.1).

Is the proposed framework capable of inserting expressiveness into synthetic speech? Table 1 shows that we obtain an MCD of 24.01 and an

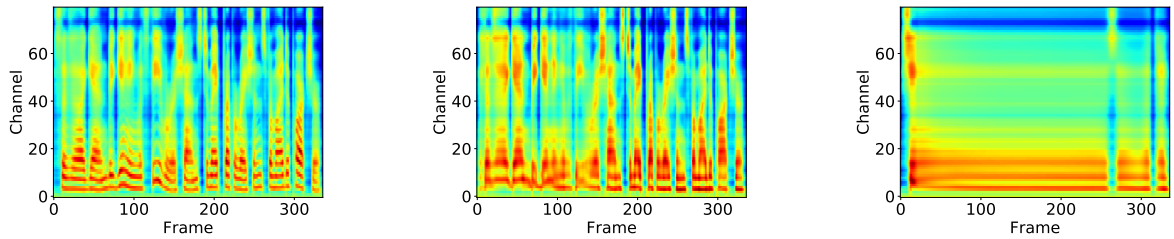
Table 2: Performance obtained using different features for emotion recognition and speaking style classification. Emotion recognition performance is measured using the unweighted average recall (UAR) while speaking style detection performance is measured using accuracy. Performance is evaluated using a leave-one-speaker-out scheme and the numbers reported are averages (± 1 standard deviation) from 30 runs to account for randomness in initialization and training. $*$ indicates that the marked performance is significantly higher than MFBs. † indicates that the marked performance is significantly higher than best APC model. Significance is assessed at $p < 0.05$ using the Tukey's honest test on the ANOVA statistics.

Features	**Emotion (UAR)** **IEMOCAP**	**MSP-IMPROV**	**VESUS**	**Style (Accuracy)** **IEMOCAP**
Baseline – Surface Features				
Chance	25.0	25.0	20.0	52.3
MFBs	53.0 ± 0.6	43.6 ± 1.2	36.0 ± 1.4	67.0 ± 0.7
Baseline – Unsupervised				
Autoencoder	50.6 ± 0.9	38.7 ± 1.0	33.6 ± 1.1	64.2 ± 0.6
APC (5-steps)	51.7 ± 0.8	42.2 ± 0.8	33.5 ± 1.2	68.3 ± 0.6
APC (10-steps)	53.9 ± 0.9	44.6 ± 0.9	35.5 ± 1.6	69.7 ± 0.6
APC (20-steps)	54.3 ± 0.9	44.1 ± 0.9	36.1 ± 1.5	69.7 ± 0.6
Paralinguistic Embeddings (ours)				
Base EVoCA	$56.4 \pm 0.6^{*\dagger}$	$46.0 \pm 0.6^{*\dagger}$	$\mathbf{44.2 \pm 0.9}^{*\dagger}$	$\mathbf{71.7 \pm 0.5}^{*\dagger}$
w/ 128-dim para. enc.	$55.4 \pm 0.8^{*\dagger}$	$45.3 \pm 0.9^{*}$	$42.6 \pm 1.4^{*\dagger}$	$69.6 \pm 0.5^{*}$
w/ 64-dim para. enc.	53.0 ± 0.6	42.9 ± 0.8	$38.2 \pm 0.9^{*\dagger}$	67.2 ± 0.5
w/ 32-dim para. enc.	51.7 ± 0.6	41.0 ± 0.4	36.0 ± 1.3	65.7 ± 0.5
w/ 128-dim converter	$\mathbf{57.1 \pm 0.5}^{*\dagger}$	$\mathbf{46.3 \pm 0.9}^{*\dagger}$	$43.5 \pm 1.3^{*\dagger}$	$70.4 \pm 0.5^{*\dagger}$
w/ 64-dim converter	$57.0 \pm 0.7^{*\dagger}$	$44.9 \pm 0.9^{*}$	$41.0 \pm 0.9^{*\dagger}$	$69.6 \pm 0.6^{*}$
w/ 32-dim converter	$54.9 \pm 0.6^{*}$	$44.6 \pm 0.7^{*}$	$38.1 \pm 1.0^{*\dagger}$	$68.8 \pm 0.5^{*}$

F0 RMSE of 146.20 when computing the performance using the synthetic reference speech and ground-truth expressive speech. In comparison, we obtain an MCD of 10.71 and an F0 RMSE of 64.36 when computing the performance using the converted speech and the ground-truth expressive speech. This suggests that the proposed EVoCA framework converts the synthetic speech so that its closer to the expressive speech. We note that it is possible to obtain better conversion performance if we increase the capacity of the model and utilize a more sophisticated vocoder. However, as the results for question 3 will suggest, increasing the capacity of the voice converter might not necessarily yield better embeddings for downstream classification tasks.

Can the learned paralinguistic embeddings be used for emotion and style classification? Table 2 shows that our paralinguistic embeddings significantly outperform MFB surface features on both the emotion recognition and the speaking style detection tasks. This suggests that the paralinguistic encoder learns a feature transformation that highlights latent paralinguistic attributes in surface acoustic features.

How do changes to EVoCA's structure affect the converted speech quality as well as the quality of the extracted embeddings for downstream tasks? Figure 3 visually demonstrates the effect of each input on the quality of a converted utterance. Figure 3a shows that the converted speech has higher quality when the paralinguistic embedding is provided as an input compared to Figure 3b. Specifically, the harmonic structure in Figure 3a is well defined and dynamic while that in Figure 3b is relatively static and not well separated. Figure 3c shows that the model is unable to generate speech solely from paralinguistic embeddings. We hypothesize that this is due to the embeddings' limited capacity to encode both linguistic and paralinguistic information present in the original signal to allow for accurate reconstruction.



(a) all components (b) w/o paralinguistic embedding (c) w/o synthetic reference

Figure 3: Sample converted test utterance with three model setups.

Additionally, we believe paralinguistic embeddings struggle to model time-varying phenomena like rhythm and speech activity because they are computed using a global average over LSTM outputs.

Table 1 quantitatively shows the effect of each of these two inputs on the conversion performance. We find that the synthesized reference input is more important to the conversion task than the paralinguistic embedding is. This is highlighted by the larger impact that reducing the capacity of the voice converter has on the converted speech quality compared to the impact of reducing the capacity of the of the paralinguistic encoder. This can be due to the fact that the paralinguistic embeddings do not have enough capacity to encode the linguistic attributes in speech that are necessary for obtaining good voice conversion performance.

Tables 1 and 2 show the results obtained on the voice conversion task and the downstream classification tasks, respectively. We find that while a high capacity voice converter improves the quality of the converted speech, it can also degrade the quality of the extracted embeddings as measured on the classification tasks. For instance, we find that reducing the capacity of the voice converter from 256 to 128 decreases the conversion performance on the voice conversion task but improves the classification performance on two out of the four downstream tasks. The results suggest that using a high-capacity voice converter can reduce EVoCA's reliance on the paralinguistic encoder for providing style and emotion information, causing the encoder to perform poorly when used to transform features for downstream applications.

How does the performance of paralinguistic embeddings compare to the embeddings learned from other unsupervised methods? Table 2 shows that paralinguistic embeddings encode information that is more suited to paralinguistic tasks than those extracted from other unsupervised methods, namely APC and a traditional autoencoder. The APC model provides improvements over surface features on all four downstream tasks when using the 20-step setup and shows improvements over surface features on three downstream tasks when using the 10-step setup. In contrast, a standard autoencoder fails to provide any improvements over surface features on all tasks. We believe that the success of the extracted embeddings from EVoCA demonstrate the importance of targeted unsupervised tasks.

7 Concluding Remarks

We proposed EVoCA, a framework for learning a surface feature transformation that highlights paralinguistic content needed for detecting emotion and speaking style. We first showed that speech synthesizers can be used to strip away paralinguistic attributes from speech while retaining linguistic information. We demonstrated how a neural voice conversion model can be adapted to facilitate the extraction of paralinguistic features by converting synthetic neutral speech to real expressive speech. Finally, we showed that these extracted embeddings improve performance over surface features and can outperform other embeddings extracted from existing unsupervised methods on emotion recognition and speaking style detection tasks. Future work will consider how the choice of the synthesis model, the number of speakers in the training set, and the architecture used for the encoder affect the quality of the extracted embeddings.

8 Broader Impact

Potential Benefits. A variety of applications can benefit from the automatic detection of paralinguistic attributes (e.g., emotion) from speech; some of these applications include: human-robot interaction, medical applications, and speaker verification to name a few. The framework that we introduce

can impact these applications by enabling the utilization of data that are not labeled for paralinguistic attributes when building the detection models for these domains.

Potential Risks. The behavior and performance of all data-driven models heavily depend on the data that are used for building them. Thus, the decisions that these models make will reflect any biases that exist in the data. Some attributes that can bias speech data include: age, gender, dialect, accent, language, recording conditions, and environment. We encourage the deployment of our framework with full consideration of these biases and their consequences on the target application.

References

Jacob Benesty, M Mohan Sondhi, and Yiteng Huang. 2007. *Springer Handbook of Speech Processing*. Springer.

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*, 42(4):335.

Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. 2016. MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80.

Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James R Glass. 2019. An unsupervised autoregressive model for speech representation learning. In *Proc. Interspeech 2019*, pages 146–150.

Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49.

Sefik Emre Eskimez, Zhiyao Duan, and Wendi Heinzelman. 2018. Unsupervised learning approach to feature analysis for automatic speech emotion recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5099–5103.

Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

Jian Gao, Deep Chakraborty, Hamidou Tembine, and Olaitan Olaleye. 2019. Nonparallel Emotional Speech Conversion. In *Proc. Interspeech 2019*, pages 2858–2862.

Wei-Ning Hsu, Yu Zhang, and James Glass. 2017. Unsupervised learning of disentangled and interpretable representations from sequential data. In *Advances In Neural Information Processing Systems*, pages 1878–1889.

Zahi N Karam, Emily Mower Provost, Satinder Singh, Jennifer Montgomery, Christopher Archer, Gloria Harrington, and Melvin G Mcinnis. 2014. Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4858–4862.

Reza Lotfian and Carlos Busso. 2015. Emotion recognition using synthetic speech as neutral reference. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4759–4763.

Reza Lotfian and Carlos Busso. 2017. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. Librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*, volume 8, pages 18–25.

Seyed Hamidreza Mohammadi and Alexander Kain. 2017. An overview of voice conversion systems. *Speech Communication*, 88:65–82.

Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. 2016. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, 99(7):1877–1884.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. VoxCeleb: A large-scale speaker identification dataset. *Proc. Interspeech 2017*, pages 2616–2620.

Andrew Rosenberg. 2012. Classifying skewed data: Importance weighting to optimize average recall. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Jacob Sager, Ravi Shankar, Jacob Reinhold, and Archana Venkataraman. 2019. VESUS: A crowd-annotated database to study emotion production and perception in spoken english. In *Proc. Interspeech 2019*, pages 316–320.

Björn Schuller and Anton Batliner. 2013. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*, 1st edition. Wiley Publishing.

Björn Schuller, Stefan Steidl, and Anton Batliner. 2009. The Interspeech 2009 emotion challenge. In *Tenth Annual Conference of the International Speech Communication Association*.

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth S Narayanan. 2010. The Interspeech 2010 paralinguistic challenge. In *Eleventh Annual Conference of the International Speech Communication Association*.

Björn Schuller, Stefan Steidl, Anton Batliner, Florian Schiel, and Jarek Krajewski. 2011. The interspeech 2011 speaker state challenge. In *Twelfth Annual Conference of the International Speech Communication Association*.

Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. 2013. The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proc. Interspeech 2013*.

Ravi Shankar, Hsi-Wei Hsieh, Nicolas Charon, and Archana Venkataraman. 2019a. Automated emotion morphing in speech based on diffeomorphic curve registration and highway networks. In *Proc. Interspeech 2019*, pages 4499–4503.

Ravi Shankar, Jacob Sager, and Archana Venkataraman. 2019b. A multi-speaker emotion morphing model using highway networks and maximum likelihood objective. In *Proc. Interspeech 2019*, pages 2848–2852.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333.

Alessandro Vinciarelli, Anna Esposito, Elisabeth André, Francesca Bonin, Mohamed Chetouani, Jeffrey F Cohn, Marco Cristani, Ferdinand Fuhrmann, Elmer Gilmartin, Zakia Hammal, et al. 2015. Open challenges in modelling, analysis and synthesis of human behaviour in human–human and human–machine interactions. *Cognitive Computation*, 7(4):397–413.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International Conference on Machine Learning*, pages 5180–5189. PMLR.

Jing-Xuan Zhang, Zhen-Hua Ling, and Li-Rong Dai. 2019a. Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:540–552.

Jing-Xuan Zhang, Zhen-Hua Ling, Li-Juan Liu, Yuan Jiang, and Li-Rong Dai. 2019b. Sequence-to-sequence acoustic modeling for voice conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(3):631–644.

Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling. 2019c. Learning latent representations for style control and transfer in end-to-end speech synthesis. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6945–6949.