Learning emotions in speech using Variational Auto Encoders

CSE 674 - Advance Machine Learning
Final report

Sri Sadhan Jujjavarapu	Chaitanya Thammineni	Mira Moukheiber
srisadha@buffalo.edu	sthammin@buffalo.edu	miramouk@buffalo.edu

1 Motivation

Speech emotion recognition (SER) is the process of extracting the emotional state of a speaker from his or her voice signal. Emotions are universal. The six universally recognized emotions are happiness, sadness, fear, anger, surprise and disgust. The SER model is a very important classification model. That is because humans communicate primarily through speech and emotion. Thus being able to extract emotions from speech is very valuable. SER is widely used in the field of human computer interaction. It has been used in a variety of applications such in social robots, computer games and E-learning. For examples, SER has been used to help teachers better manage the emotional state of students in classrooms Kerkeni et al. [2018].

2 Introduction

SER involves two parts: feature extraction and emotion classification. The features extraction is a challenging task due to the ambiguity and complexity of speech in differentiating between pitch, energy, quality etc. Moreover, the emotion is affected by factors such as age Mill et al. [2009] and gender Vogt and André [2006]. Another major challenge is the lack of correctly labelled data.

Recently, Deep learning (DL) techniques have shown promising results in the speech recognition. One of the popular DL techniques is to use Convolutional neural networks (CNN) on the spectrograms of segmented speech. These have observed to be more robust against the noise compared to other DL techniques Palaz et al. [2015]. DL has also been proven to improve the performance of emotional recognition Stuhlsatz et al. [2011], Cibau et al. [2013], Li et al. [2013]. Multiple papers in the literature , have depicted that using Long Short-Term Memory models (LSTM) in combination with CNNs boosts speech emotion recognition Trigeorgis et al. [2016]. Another paper on attentive CNN compared different features for use in SER, showed that the log Mel filter-banks (logMel) and Mel-frequency cepstral coefficients (MFCC) preformed better compared to the other types of features in terms of accuracy Mill et al. [2009].

Autoencoders have been widely used in SER Deng et al. [2013], Cibau et al. [2013]. However, Variational Autoencoders (VAE's) have been mainly used to model natural images and their use in SER is sparse. In terms of speech, VAE's have been mainly used to model the process of speech generation and to learn the latent representations of speech such as the phonetic content and Hsu et al. [2017]. One recent study used VAE's for speech emotion classification Latif et al. [2017]. The authors of this study showed that the features generated by VAE produced very promising results for speech emotion classification.

3 Method

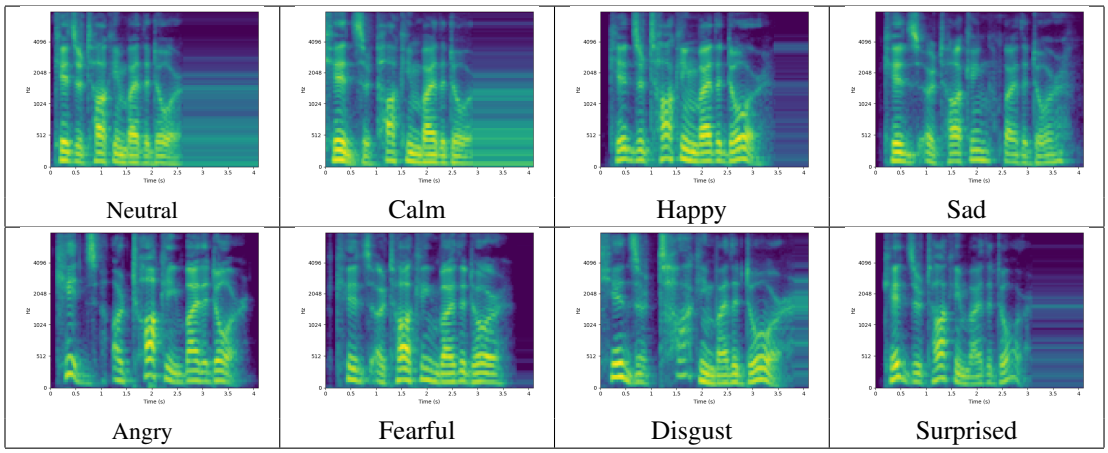
3.1 Dataset Description

An open source Ryerson Audio Visual Database of Emotional Speech and Song (RAVDESS) Livingstone and Russo [2018] is used for the emotion detection. The database has 7356 files of 24 actors (12 male, 12 female) saying two lexically matched sentences(" Kids are talking by the door." and "Dogs are sitting by the door.") in a neutral North American Accent. Speech entails calm, happy, sad, angry, fearful, disgust and surprise expressions and the Song entails calm, happy, sad, angry and fearful emotion. Each expression was generated at normal and strong emotional intensities along with an additional neutral expression. In this project, we have considered the audio speech signal available in the ".wav" file format containing **8 emotions** with **4 repetitions** each (we have considered intensities as repetitions). The dataset can be found at: https://smartlaboratory.org/ravdess/

3.2 Audio signal processing

In the RAVDESS dataset, the audio signal is sampled at 48000 Hz. Each of the raw audio signal is resampled to 16000 Hz and the voice activity is detected using the WebRTC Voice Activity Detector (VAD). The obtained voice activity is then converted into a Mel-Spectrogram (spectrogram with 80 Mels) using the *LibROSA package* McFee et al. [2015] popularly used for music and audio analysis. The obtained power of the spectrogram is then converted into decibels and transformed into a segment of constant length (128) as shown in Table. 1. Here, the speech segment is padded to obtain constant length samples. Therefore, the size of all the log-MelSpectrograms are 80×128. The x-axis (128) corresponds to the time period of the speech and the y-axis to the frequency in Mel scale (80) i.e 80 frequency banks. The examples of the audio clips of the constant length samples are attached in the folder **input_samples**.

Table 1: Log-Mel spectrograms of the audio clips. Short audio-clips are padded to match the image size of 80×128.



4 Variational Auto Encoders

One of the motivations of this project is to find disentangled representations of the given dataset of observations $X = x^i$, where $i \in (1, N)$ in a supervised or unsupervised manner. Simply stated disentangled representations are the independent underlying factors of variation in the data . They are explained in more detail in the section that follows, but lets setup the way we intend to find these hidden factors of variation $Z \in R^d$ in an unsupervised way.

We approach this problem with a certain method called Variational Auto Encoders(VAE) as introduced in Kingma and Welling [2013]. If we consider our dataset X given by observations $X = x^i$, where $i \in (1, N)$ and the hidden factors of variation (generation) in the real data as Z, we can assume the process of generating X from Z be given by $p_\theta(x|z)$. Our intention here is to find the opposite i.e, given the dataset X, we need to find the hidden factors Z, given by $q_\phi(z|x)$.

The objective to achieve $p_\theta(x|z)$ is the max log-likelihood of generating x given z as

$$max_\theta \sum_i \log p_\theta(x^i) = \sum_i \log \sum_z p_Z(z) p_\theta(x^i|z) \tag{1}$$

Using importance sampling and a proposal distribution based on $q_\phi(z|x)$ the objective in Eq.1 can be rewritten as

$$\sum_i \log \frac{1}{K} \sum_{k=1}^{K} \frac{p_Z(z_k^i)}{q(z_k^i)} p_\theta(x^i|z_k^i) \tag{2}$$

with z_k^i sampled from $q(z_k^i)$.

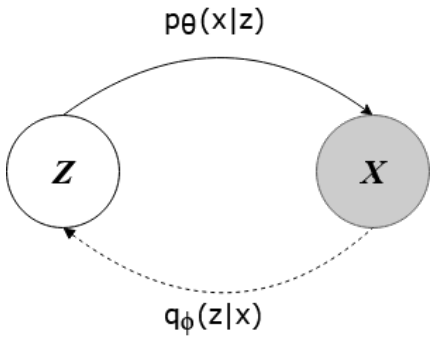


Figure 1: VAE as a graphical model. Dashed lines denote variational approximation.

We also need to add to ensure proposal distribution is valid, leading to an additional objective:

$$min_\phi \sum_i KL q_\phi(z|x)||p_\theta(x|z) \tag{3}$$

Combining Eq.1, Eq.2 and Eq.3 the final training objective is maximizing

$$\mathbb{E}_{\epsilon\sim\mathcal{N}} \log p_\theta(x) - KL q_\phi(z|x)||p_\theta(x|z) \tag{4}$$

We assume $p(\theta)$, the prior distribution and $p_\theta(x|z)$, the likelihood to be from a parametric family of distributions. This gives us

$$\begin{aligned} q_\phi(z|x) &= \mathcal{N}(\mu_\phi(x), \sigma_\phi^2(x)) \\ z &= \mu_\phi(x) + \epsilon\sigma_\phi(x),\ \epsilon \sim (0, I) \end{aligned} \tag{5}$$

The objective in Eq.4 can also be arrived at by using the Variational Lower Bound(VLB) method as described in Kingma and Welling [2013]. The model is trained by maximizing the VLB or ELBO(Evidence Lower BOund). Maximizing this ELBO is equal to maximizing the lower bound of the objective in Eq.4. The latent space embeddings, z are calculated from the obtained μ_z and ${\sigma^2}_z$ values by applying the *reparameterization trick* as discussed in Kingma and Welling [2013]

4.1 Network architecture

We use 1D convolutions throughout the encoder. All the convolution operations are followed by ReLU activations. The Kernel size, stride and padding across all convolutions throughout the model are 4,2,1 respectively. The output of the encoder after passing through a fully connected layer that yields 16 values that correspond to the μ_z and $\log {\sigma^2}_z$ values of the distribution $p_Z(z)$ for the 8 emotions. The architecture is summarized in Table 2 and Fig 2.

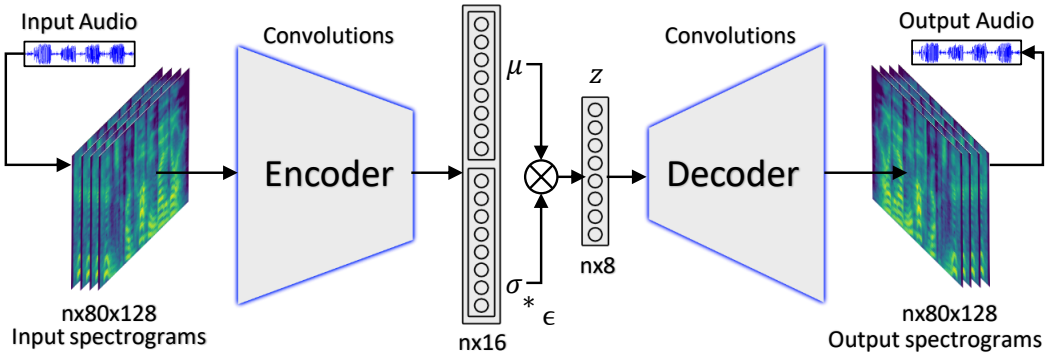


Figure 2: The Variational Auto Encoder architecture used in this project. More details about the convolution operations in the encoder and decoder are provided in Table. 2

Table 2: 1d Convolution operations on input spectrogram

operation	input size	output size	kernel size	stride	padding
Convolution	80	64	4	2	1
Convolution	64	32	4	2	1
Convolution	32	16	4	2	1
Convolution	16	16	4	2	1
Convolution	16	128	4	2	1
Convolution	128	256	4	2	1
Linear	256	16	-	-	-

5 Results and Discussion

The VAE was first trained on the emotion data from a single actor and the latent space representation were analyzed.

5.1 Training on single actor

The dataset of a single actor has 2 intensities and 2 repetition for each emotion. Therefore, a total of 32 spectrograms ($8 \times 2 \times 2$) should be available for each subject. Emotion 1 does not have 2^{nd} intensity, therefore, the total number of samples available are 30. These 30 samples are padded for constant length and the log-Mel spectrograms are extracted. The obtained spectrograms are used to train a VAE for 25000 epochs. While training, the reconstruction loss and the KL divergence were obtained and plotted against the training epochs. Figure. 3 illustrates the loss on two separate plots, we observe that the loss saturate around 16000 epochs. During training, a random spectrogram is picked from the final reconstruction at every 100 epochs and converted into audio samples. Some of the audio samples obtained during training are attached in the folder **training_samples** to demonstrate the variation in the audio reconstructions over the training period.

5.2 Decoder as a Generative model

While training the variational auto encoder, the decoder learns the representation of the speech along with the speaker's emotion, however the representations depends on the extent of disentaglement. In this section, we show how the decoder can be used as a generative model to generate new samples by sampling the latent variables, z from a Normal distribution ($\mathcal{N}(0, 1)$) with zero mean and unit standard deviation, Doersch [2016]. Table. 3 displays one set of generated audio samples by randomly sampling one latent variable and assigning zeros to the rest. The samples obtained by converting these log-Mel spectrograms back to the audio are present in the folder **latent_samples**.

It is observed that the images in the Table. 3 are a little smooth compared to the actual audio samples presented in Table. 1. The reason for such behavior is that the VAE's produce blurred/smoothened outputs due to the gaussian prior. This is one of the disadvantages of the VAE when compared to the Generative Adversarial Networks that can provide sharp output images. Moreover, we observed

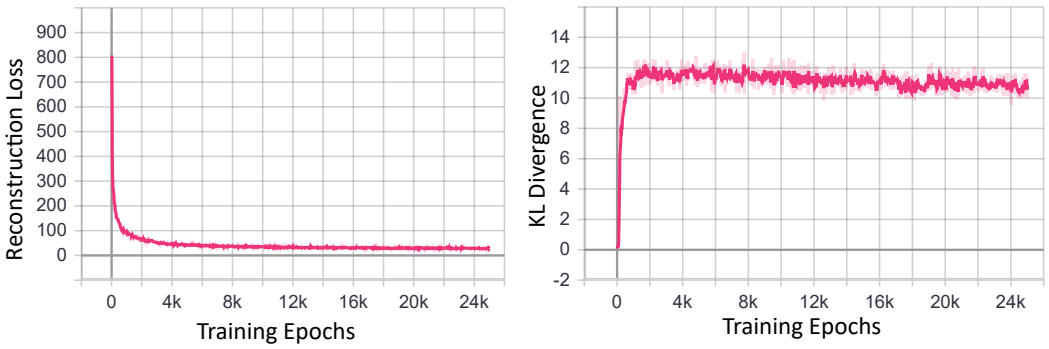
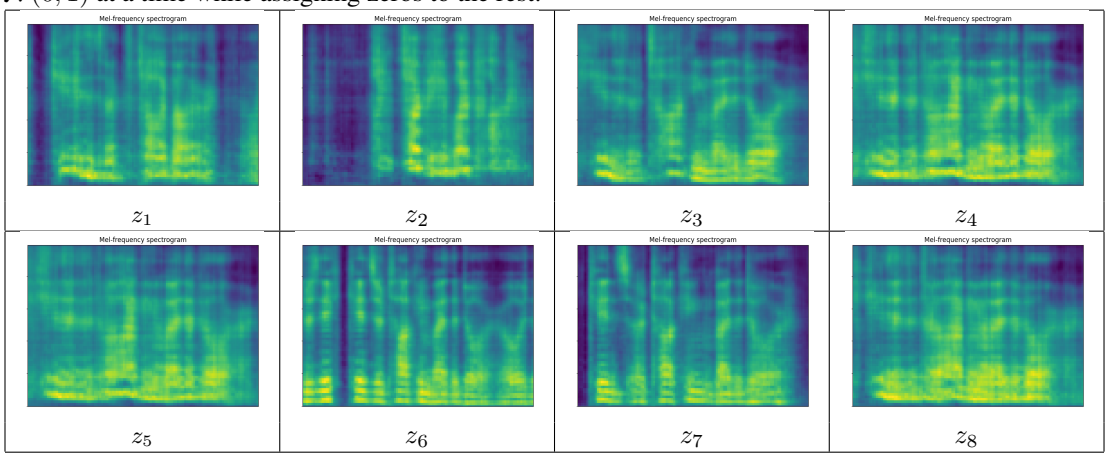


Figure 3: The Reconstruction loss and KL-divergence obtained while training the VAE on a single actor

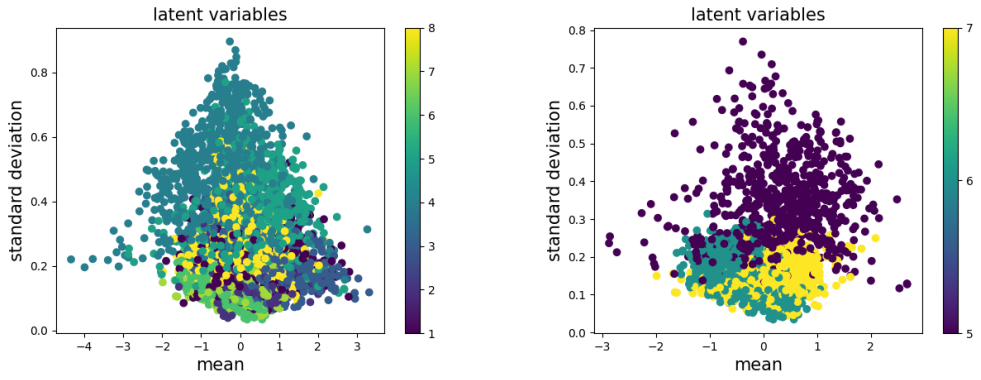
Table 3: Log-Mel spectrograms of the audio clips generated by sampling one latent variable, z_i from $\mathcal{N}(0, 1)$ at a time while assigning zeros to the rest.



that the obtained audio clips as a result of the sampled latent variables could not disentangle all the emotions. We will study more about the disentangled representations in Section 5.3.

5.3 Disentanglement

It is mentioned by Kim and Mnih [2018] that even in a simplest setting, a robust disentanglement of the latent space embeddings is not completely guaranteed. In this project, we explore the disentanglement of the latent space by plotting the mean and standard deviation values obtained after passing the all the audio samples through the encoder. There are a total of 8 embeddings in the latent space where each embedding should correspond to an emotion provided that the VAE can perfectly disentagle all the 8 emotions from the speech signal. To understand how well the VAE has disentangled the representations, we plotted the standard deviation σ vs μ values of all the samples on a 2d plot. The log-Mel spectrograms of the complete dataset (24 actors) were passed through the encoder part of the previously trained VAE (VAE trained on single actor) and the resultant σ and μ are plotted for all the emotions. The samples were color coded according to the true labels of the emotions. The resultant plots of all the embeddings are illustrated in Fig. 4. Figure. 4a) represents the plot of all 8 embeddings and Fig. 4b) represents only 3 out the 8 latent space variables that are well disentangled i.e. embedding 4, 5 and 6. The reason for such behavior could be that there might be some similarities in the considered 8 emotions by nature. For example, angry and disgust can be similar when analyzed through the spectrograms (Table .1). On the other hand, neutral and calm might not contain much differences to be disentangled.



(a) (b)

Figure 4: μ and σ of the Latent space distribution obtained by passing the complete dataset (24 actors) through the encoder of the VAE (trained on the emotions of 1 actor) are plotted for the complete dataset (24 actors): a) μ_i and σ_i where $i \in (1, 8)$, a) μ_i and σ_i where $i \in (4, 6)$ (well learnt representations)

5.4 Training on 12 actors of same gender

When the VAE is trained on the complete dataset, a representation of the statement "Kids are talking by the door" is embedded in all the latent variables whereas the emotion data could not be learnt i.e. the losses were not converging. The behavior is because the complete dataset consists of 12 male and 12 female actors where it is observed that for a constant emotion and the spectrograms differ across male and female voices due to the variation in pitch and frequencies (which can also vary across the same gender with age groups). Therefore, the dataset is segregated into male and female speakers and the trained individually. Figure. 5 demonstrates the loss while training the VAE with 12 actors of same gender. Due to the computational expense, the training was only carried for 6000 epochs. Some of the reconstructed audio samples are attached in the folder, **training_samples_12_actors**.

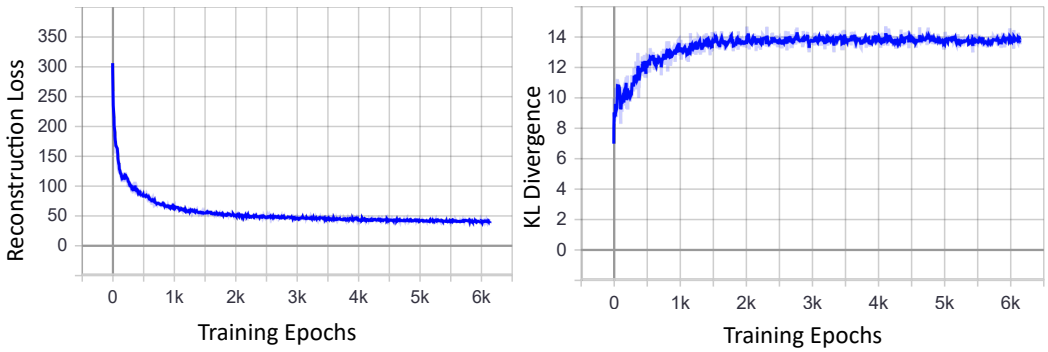
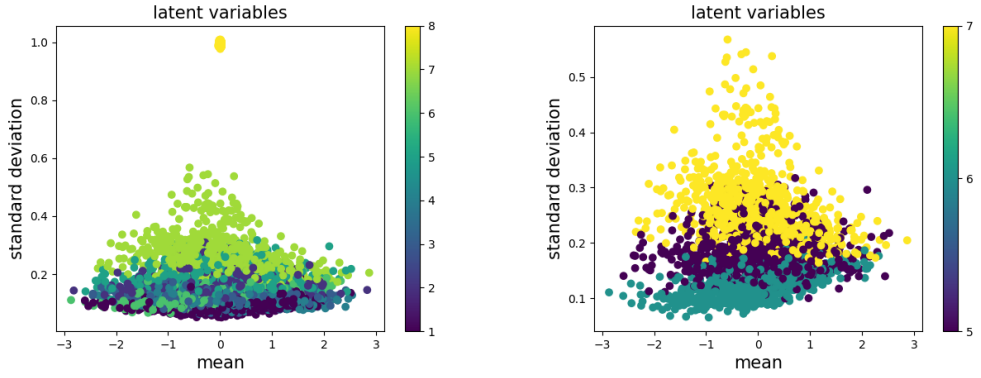


Figure 5: The Reconstruction loss and KL-divergence obtained while training the VAE on 12 actors of same gender

Similar to the previous latent embeddings plot (4), the complete dataset is passed through the encoder of the trained model but this time the model is trained on 12 actors of same gender. The resultant plots of all the embeddings are illustrated in Fig. 6. Here, it is observed that the latent space embeddings that were previously separable (4, 5, 6) are no longer that well separated when the VAE trained on 12 actors is used. Even with the dataset of same gender, differences in the latent space are observed. These differences are implied by the differences in the pitch and energy across the varying age group of individuals. Such variations makes it harder to detect the emotions of all the individuals with same accuracy.



(a) (b)

Figure 6: μ and σ of the Latent space distribution obtained by passing the complete dataset (24 actors) through the encoder of the VAE (trained on the emotions of 12 actors) are plotted for the complete dataset (24 actors): a) μ_i and σ_i where $i \in (1, 8)$, a) μ_i and σ_i where $i \in (4, 6)$ (well learnt representations)

6 Conclusion

In this project, we used a Variational Auto Encoder for learning the representations of emotions from the speech data. The speech data is downloaded from an open source Ryerson Audio Visual Database of Emotional Speech and Song (RAVDESS). There are a total of 24 actors who produced 4 repetitions of 8 emotions (calm, happy, sad, angry, fearful, disgust and surprise). The speech data from each individual is converted into a spectrogram of constant length (80×128) in time-frequency domain (log-Mel Spectrogram). Here, 80 corresponds to the total number of Mel scale frequency bands and 128 corresponds to the length of speech (so that the complete statement is encoded into the spectrogram). Shorter spectrograms are padded with values to maintain constant size of the image. These images are then used to train a VAE with encoder and decoder consisting of 1d convolution layers and ReLU activations. The output of the convolution layer is passed through a fully connected layer to get a vector of 16 dimensions. The first 8 are considered as mean and rest as the log-variance of the latent space distributions. With the help of a reparameterization trick, these values are converted into latent space embeddings z. The obtained embeddings are passed through a decoder to obtain the output spectrogram. The objective here is to get a well disentagled latent space along with a good reconstruction of the passed input.

We observed that the model performed well in disentangling some of the latent space embeddings while trained on a single actor's dataset. This performance degraded with the inclusion of data from multiple actors for training due to the differences in the voice across age-groups and gender. We also demonstrated the disentanglements by plotting the mean and standard deviation of the latent space embeddings. Additionally, the use of VAE as a data generation model is demonstrated by sampling the latent space from a normal distribution and generating audio samples using the pre-trained decoder (one trained on 1 actor and the other on 12 actors).

The main challenges experienced here is the the generalization of the VAE. The decoder embeds the speech along with the emotion. Other ways should be explored to train the VAE independent of the speech. In future, we will explore recurrent neural networks (RNN) such as long-short term memory (LSTM) cells. They will be used to process a speech signal of variable length without the need for padding. The drawback with using such models is the computational expense in training the models.

References

Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, and Mohamed Ali Mahjoub. Speech emotion recognition: Methods and cases study. In *ICAART (2)*, pages 175–182, 2018.

Aire Mill, Jüri Allik, Anu Realo, and Raivo Valk. Age-related differences in emotion recognition ability: A cross-sectional study. *Emotion*, 9(5):619, 2009.

Thurid Vogt and Elisabeth André. Improving automatic emotion recognition from speech via gender differentiaion. In *LREC*, pages 1123–1126, 2006.

Dimitri Palaz, Ronan Collobert, et al. Analysis of cnn-based speech recognition system using raw speech as input. Technical report, Idiap, 2015.

André Stuhlsatz, Christine Meyer, Florian Eyben, Thomas Zielke, Günter Meier, and Björn Schuller. Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5688–5691. IEEE, 2011.

Neri E Cibau, Enrique M Albornoz, and Hugo L Rufiner. Speech emotion recognition using a deep autoencoder. *Anales de la XV Reunion de Procesamiento de la Informacion y Control*, 16:934–939, 2013.

Longfei Li, Yong Zhao, Dongmei Jiang, Yanning Zhang, Fengna Wang, Isabel Gonzalez, Enescu Valentin, and Hichem Sahli. Hybrid deep neural network–hidden markov model (dnn-hmm) based speech emotion recognition. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 312–317. IEEE, 2013.

George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5200–5204. IEEE, 2016.

Jun Deng, Zixing Zhang, Erik Marchi, and Björn Schuller. Sparse autoencoder-based feature transfer learning for speech emotion recognition. In *2013 humaine association conference on affective computing and intelligent interaction*, pages 511–516. IEEE, 2013.

Wei-Ning Hsu, Yu Zhang, and James Glass. Learning latent representations for speech generation and transformation. *arXiv preprint arXiv:1704.04222*, 2017.

Siddique Latif, Rajib Rana, Junaid Qadir, and Julien Epps. Variational autoencoders for learning latent representations of speech emotion: A preliminary study. *arXiv preprint arXiv:1712.08708*, 2017.

Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5), 2018.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.