UNSUPERVISED LEARNING APPROACH TO FEATURE ANALYSIS FOR AUTOMATIC SPEECH EMOTION RECOGNITION

Sefik Emre Eskimez, Zhiyao Duan, Wendi Heinzelman

University of Rochester
Electrical and Computer Engineering
500 Wilson Blvd, Rochester, NY 14627, USA

ABSTRACT

The scarcity of emotional speech data is a bottleneck of developing automatic speech emotion recognition (ASER) systems. One way to alleviate this issue is to use unsupervised feature learning techniques to learn features from the widely available general speech and use these features to train emotion classifiers. These unsupervised methods, such as denoising autoencoder (DAE), variational autoencoder (VAE), adversarial autoencoder (AAE) and adversarial variational Bayes (AVB), can capture the intrinsic structure of the data distribution in the learned feature representation. In this work, we systematically investigate four kinds of unsupervised feature learning methods for improving speaker-independent ASER. We show that all methods improve the performance regarding unweighted accuracy rating (UAR) and F1-score over methods that use hand-crafted features or that do not perform feature learning on external datasets. We also show that VAE, AAE and AVB methods, which control the distribution of the latent representation, outperform DAE that does not control such distribution. This suggests the benefits of using variational inference methods to learn features from general speech for the speech tasks such as ASER that has very limited labeled data.

Index Terms— Automatic speech emotion classification, unsupervised feature learning, autoencoders, variational inference

1. INTRODUCTION

Emotions are a vital part of social interactions. Designing computational models to recognize emotions is key to an automatic understanding of social interactions. In recent years, researchers have developed automatic emotion recognition systems using different data modalities, including physiological signals [1], facial expressions and body gestures [2], and speech [3]. Among these modalities, speech is more accessible and less intrusive in daily life. Therefore, automatic speech emotion recognition (ASER) has received much attention in this field.

ASER is a challenging task. While automatic systems have been shown to outperform naive human listeners on speech emotion classification [4], unlike speech and image classification tasks, current ASER systems are still not competitive to trained human listeners. One bottleneck for improving ASER is the lack of training data. Recording and annotating emotional speech is a very time-consuming process. Compared to general speech datasets, publicly available speech emotion recognition datasets are much more limited in the number of speakers and utterances, and the coverage of vocabulary and recording conditions [3].

One way to alleviate the data lacking issue is to transfer knowledge learned from unlabeled data or data in other related tasks (source tasks) to the task at hand (target task) [5]. One technique is unsupervised feature learning, which does not utilize the label information but aims to learn robust features that can capture the intrinsic structures of the data. These features are also often discriminative to train better classification models for the target task [6, 7]. For ASER, the most natural and available data sources are general speech. They may not carry strong emotions, but features learned from these data may capture intrinsic structures of speech and be useful for ASER.

Unsupervised feature learning has been rarely explored in ASER beyond autoencoders (AE) [8] and denoising autoencoders (DAE) [6]. AE and DAE aim to learn features that are good for the reconstruction of the input. More advanced techniques, such as variational autoencoders (VAE) [9] and generative adversarial networks (GAN) [10], do not aim to reconstruct the input, but aim to *generate* data that come from the same distribution as the input. This relaxation tends to put more emphasis on the modeling of intrinsic structures of the data during feature learning [7, 9, 10].

In this paper, we design a convolutional neural network (CNN)-based ASER system and make the first systematic exploration of various kinds of unsupervised learning techniques to improve the speaker-independent emotion recognition accuracy. These techniques include the denoising autoencoder (DAE), variational autoencoder (VAE), adversarial autoencoder (AAE) and adversarial variational Bayes (AVB). We compare these systems with two baselines (SVM and CNN) that work on hand-crafted features without unsupervised feature learning. Experiments show that unsupervised feature learning significantly improves the ASER performance, when trained on a large scale general speech dataset, regarding unweighted accuracy rating (UAR) and F1-score. Furthermore, the latent variable models including VAE, AAE, and AVB improve the ASER performance more than the DAE and other baselines. This suggests that unsupervised learning with these latent variable models are useful practices for ASER, where training data is insufficient.

The rest of the paper is organized as follows: Related work on ASER and its usage of unsupervised learning are presented in Section 2. Section 3 describes the proposed ASER system and the explored unsupervised learning approaches. Section 4 presents experimental results, and Section 5 concludes the paper.

2. RELATED WORK

Traditional ASER systems that utilize Gaussian mixture models (GMMs) [11–13], hidden Markov models (HMMs) [14, 15], and support vector machines (SVMs) [16–18], rely on well-established hand-crafted speech features. These features usually include spec-

ZD acknowledges National Science Foundation Grant No. 1617107.

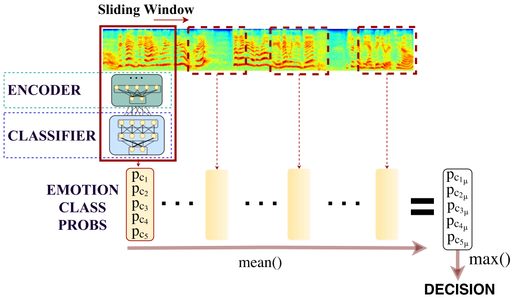

Fig. 1. Proposed ASER system overview. The dashed red windows represent the sliding window with 50% overlap. From each window, emotion class probabilities (p_1, p_2, p_3, p_4 and p_5) are predicted and the average of these vectors is calculated over all windows is calculated for each utterance. The emotion that has the highest probability is predicted as the emotion of the utterance.

tral, cepstral, pitch, and energy features of the speech signal at the frame level. Statistical functionals of these features are then applied across multiple frames to obtain an utterance-level feature vector.

Some researchers explored deep learning methods to find robust features for the ASER task. Xia et al. [19] proposed a modified DAE that maps input speech to two hidden representations, a neutral representation learned by reconstructing neutral speech beforehand, and an emotional representation learned by reconstructing emotional speech with the neutral representation fixed. During testing, the emotional representation of a test speech sample is fed to an SVM classifier for emotion classification. In their follow-up work [20], Xia et al. incorporated the speaker gender information which resulted in further improvements.

Ghosh et al. [21, 22] trained stacked DAEs and a bidirectional long short-term memory (BLSTM) AE to obtain a latent representation of the input spectrogram extracted from the speech and the glottal flow waveform. These latent representations were then fed to a multilayer perceptron (MLP) with a softmax output for 4-class emotion classification.

Deng et al. [23] proposed a single-layer sparse autoencoder (SAE) for feature transfer learning between different emotion corpora. One SAE was trained for each emotion class in the source domain using hand-crafted features as the input. Then each training sample in the target domain was reconstructed by the SAE of the corresponding class. Finally, an SVM model was trained on the reconstructed data to classify the original test samples without going through the SAEs. Deng et al. [24] obtained further improvements by replacing the SAEs with denoising autoencoders (DAEs).

Although these studies have demonstrated the benefits of unsupervised feature learning using DAEs, more advanced latent variable methods such as VAE, AAE, and AVB have not been explored for ASER. These methods attempt to model the distribution of data and are likely to learn more meaningful, controllable and discriminative features, leading to better classification performance, especially when the amount of labeled data is small [7].

3. METHOD

We propose to adopt a convolutional neural network (CNN)-based architecture (shown in Fig. 1) for ASER and to investigate the effects of different unsupervised learning techniques. Specifically, the network contains a pre-trained encoder network to extract features from the log-Mel spectrogram of the input speech, and a fully connected (FC) network to classify their emotions. The encoder includes three convolutional layers with a leaky rectified linear unit (LReLU) activation and an FC layer with a linear activation as shown in Table 1. The encoder gradually reduces the dimension of the input into the latent dimension. During classification, the encoder network weights are frozen. The classifier consists of three fully connected layers with LReLU activations except for the last activation, which uses softmax to represent probabilities of each emotion class. There are two dropout layers with 0.25 drop rate between FC layers. The categorical cross-entropy loss is used during the training of the FC.

Net	Layers	Activ.	F. No	F. Size	Strides	Output Shape
encoder (q_θ)	Input (x)	-	-	–	–	$64 \times 64 \times 1$
	Conv2D	LReLU	32	9×9	2×2	$32 \times 32 \times 32$
	Conv2D	LReLU	64	7×7	2×2	$16 \times 16 \times 64$
	Conv2D	LReLU	128	5×5	2×2	$8 \times 8 \times 128$
	Flatten	-	-	–	–	8192
	FC	Linear	-	–	–	256
decoder (p_ϕ)	Input (z)	-	-	–	–	256
	FC	LReLU	-	–	–	8192
	Reshape	-	-	–	–	$8 \times 8 \times 128$
	Conv2DT	LReLU	128	5×5	2×2	$16 \times 16 \times 128$
	Conv2DT	LReLU	64	7×7	2×2	$32 \times 32 \times 64$
	Conv2DT	LReLU	32	9×9	2×2	$64 \times 64 \times 32$
	Conv2D	Sigmoid	1	1×1	1×1	$64 \times 64 \times 1$
AAE discriminator	Input (z)	-	-	–	–	256
	FC	LReLU	-	–	–	2048
	FC	LReLU	-	–	–	2048
	FC	LReLU	-	–	–	2048
	FC	Sigmoid	-	–	–	1
AVB discriminator	Input (z)	-	-	–	–	256
	FC	LReLU	-	–	–	4096
	Reshape	-	-	–	–	$64 \times 64 \times 1$
	Input (x)	-	-	–	–	$64 \times 64 \times 1$
	Concat	-	-	–	–	$64 \times 64 \times 2$
	Conv2D	LReLU	32	9×9	2×2	$32 \times 32 \times 32$
	Conv2D	LReLU	64	7×7	2×2	$16 \times 16 \times 64$
	Conv2D	LReLU	128	5×5	2×2	$8 \times 8 \times 128$
	Flatten	-	-	–	–	8192
	FC	LReLU	-	–	–	256
	FC	Sigmoid	-	–	–	1
classifier	Input (z)	-	-	–	–	256
	FC	LReLU	-	–	–	1024
	Dropout	-	-	–	–	1024
	FC	LReLU	-	–	–	1024
	Dropout	-	-	–	–	1024
	FC	Softmax	-	–	–	5

Table 1. The architecture of the encoder, decoder, discriminator and emotion classifier networks. AEs share the encoder and decoder structures, except AVB where we modify the encoder to accept external noise input similar to AVB discriminator architecture. *Conv2D* is a 2-d convolution layer, where *Conv2DT* is a transposed 2-d convolution (or deconvolution) layer. *Concat* is the concatenation layer. *F. No* is the number of filters, where *F. Size* is the filter size.

The proposed network processes each utterance by segments that are 1 second long. During training, we randomly choose patches to form training batches from each utterance and use the utterance-level label as the label for the segment. During testing, we segment each utterance into 1-second long segments with a 0.5-second overlap. We predict the emotion probabilities in each segment and then average the probabilities across all segments. We finally choose the emotion category, which has the highest mean probability, as the utterance-level emotion classification result.

In the following, we describe different architectures, and inference models for the encoder explored in this paper, including denoising autoencoder (DAE), variational autoencoder (VAE), adversarial autoencoder (AAE) and Adversarial Variational Bayes (AVB).

3.1. Denoising Autoencoder (DAE)

Denoising autoencoders (DAEs) [6] aim to extract robust features by reconstructing clean data from their corrupted versions. They have been applied to ASER systems [19–22] and yielded performance increase. The model can be expressed as:

$$z \sim q_\theta(z|\tilde{x}), \tag{1}$$

$$\hat{x} \sim p_\phi(x|z), \tag{2}$$

where $z, x, \tilde{x}$ and $\hat{x}$ are the latent representation, clean data, corrupted data and reconstructed clean data, respectively. q_θ and p_ϕ are the probabilistic notation of the encoder and decoder networks, where θ and ϕ are the trainable parameters of the networks. When cross-entropy is used to measure the reconstruction error, the loss function is defined as:

$$\min_{\theta,\phi} -\mathbb{E}_{z\sim q_\theta(z|\tilde{x})}[\log p_\phi(x|z)]. \tag{3}$$

As we do not have an estimation nor control of the distribution of the latent representation, it is difficult to generate new but realistic data using the decoder of DAEs.

We train a DAE using the same encoder-decoder architecture as shown in Table 1. The encoder and decoder networks are symmetrical except for the last layer of the decoder network.

3.2. Variational Autoencoder (VAE)

VAE [9] is another version of AE that performs variational inference by constraining the latent representation to match an explicit distribution such as a normal distribution. The latent representation is defined as follows:

$$(z_\mu, z_\sigma) \sim q_\theta(z_\mu, z_\sigma|x), \tag{4}$$

$$z = z_\mu + z_\sigma \odot \mathcal{N}(0, I), \tag{5}$$

where z_μ, z_σ are the mean and standard deviation obtained from the encoder network, and $\mathcal{N}(0, I)$ is the Gaussian distribution with zero mean and unit standard deviation. The loss function is defined as

$$\min_{\theta,\phi} KL\left(q_\theta(z|x) \| p(z)\right) - \mathbb{E}_{q_\theta(z|x)}[\log p_\phi(x|z)], \tag{6}$$

where $p(z) = \mathcal{N}(z; 0, I)$ is the prior multivariate Gaussian distribution that we want latent representation to match and KL is the *Kullback-Leibler* (KL) divergence respectively. The first term regularizes the output latent distribution of the encoder and the second term is the reconstruction loss of AE. Since the latent representation distribution is controlled, new but realistic samples can be easily generated by feeding to the decoder the randomly drawn latent representations according to the normal distribution.

We train a VAE using the same architecture as the encoder-decoder shown in Table 1 except that we modify the encoder network by replacing the last layer with two fully connected layers, which output z_μ and z_σ. We calculate the latent representation z using Eq. (4), and feed it to the decoder network.

3.3. Adversarial Autoencoder (AAE)

Generative adversarial networks (GANs) have achieved remarkable success in generating realistic data [10]. GANs are zero-sum two player game where the players are the counterfeiter and the police. The counterfeiter forges a fake sample and presents it to the police, and the police try to distinguish between real and fake samples. In neural network terminology, the counterfeiter is called the generator network and the police is called the discriminator network.

Adversarial autoencoders (AAEs) [25] are a type of AE that performs variational inference by constraining the latent distribution to match a specified distribution $p(z)$ through adversarial training. In GAN terms, the encoder $q_\theta(z|x)$ tries to fool the discriminator by generating latent codes that mimic $p(z)$. The min-max game can be expressed as:

$$\begin{aligned} \min_{\theta,\phi}\max_{\psi} \mathbb{E}_{z\sim p(z)}[\log D_\psi(z)]+ \\ \mathbb{E}_{x\sim p_{data}}[\log(1 - D_\psi(q_\theta(z|x)))]- \\ \mathbb{E}_{z\sim q_\theta(z|x)}[\log p_\phi(x|z)], \end{aligned} \tag{7}$$

where $D_\psi(\cdot)$ is the discriminator, and ψ is its parameter. The first two terms are the GAN loss involving the encoder and the discriminator, while the third term is the reconstruction loss involving the encoder and the decoder. AAEs rely on reconstruction loss to capture the data distribution where adversarial loss acts as a regularization term over latent distribution to match the prior distribution.

We use the same architecture that is used for the other AEs for the encoder and decoder networks. We add a discriminator network shown in Table 1 to distinguish between real and fake latent codes.

3.4. Adversarial Variational Bayes (AVB)

AVB is a training technique for VAEs that replaces the KL term with an adversarial loss [26]. The discriminator inputs are pairs of (x, z) where x is sampled from the real data distribution and z is either sampled from the prior distribution or obtained from the inference model. The discriminator tries to distinguish whether the pairs are sampled from the prior distribution or the inference model.

The encoder-decoder model parameters are updated with Eq. (8) where the discriminator parameters are updated with Eq. (9).

$$\begin{aligned} \min_{\phi,\theta} \mathbb{E}_{x\sim p_{data}}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}[D_\psi(x, q_\theta(z|x,\epsilon))]- \\ \mathbb{E}_{z\sim q_\theta(z|x,\epsilon)}[\log p_\phi(x|z)], \end{aligned} \tag{8}$$

$$\begin{aligned} \max_{\psi} \mathbb{E}_{x\sim p_{data}}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}[\log D_\psi(x, q_\theta(z|x,\epsilon))]+ \\ \mathbb{E}_{x\sim p_{data}}\mathbb{E}_{z\sim p(z)}[\log(1 - D_\psi(x, z))], \end{aligned} \tag{9}$$

We modify the discriminator to accept both the data and latent code. The latent code dimensionality is increased by an FC layer than added to the data as a second channel. The architecture is shown in Table 1. We modify the encoder network to accept external noise $\epsilon \sim \mathcal{N}(0, I)$; we follow the same steps described for the discriminator network to merge ϵ into the data as a second channel.

4. EXPERIMENTS

4.1. The Data

In our experiments we use USC-IEMOCAP audio-visual dataset [27] that contains scripted and improvised interactions between actors, we only use the audio files. There are five sessions totaling about 12 hours of data, where each session includes interactions between a female and a male. There are three annotators, where annotations include both categorical and real-valued. Categorical emotions include anger, disgust, excitement, fear, frustration, happiness, neutral, sadness and surprise. We only considered categorical

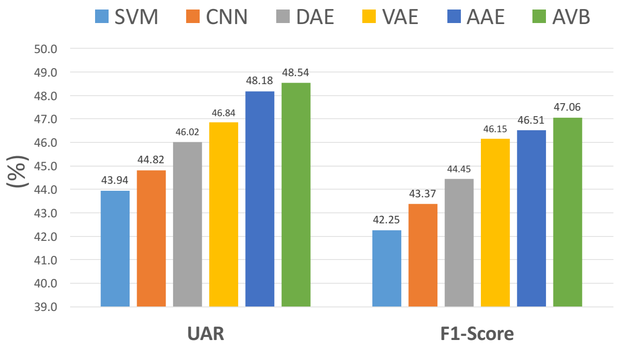

Fig. 2. The unweighted accuracy rating (UAR) and F1-score results for the baseline systems and the proposed systems. F1-score is calculated for each class, and their unweighted mean is presented.

annotations that are agreed by at least two annotators. This database is commonly used in the ASER literature [13, 20, 22].

While most existing work on this database considered only four emotion categories, we consider five, which are anger (972 samples), excited (948), frustration (1670), neutral (1507) and sadness (1039). In all of our experiments, we apply leave-one-session-out cross-validation, where for each rotation we train on four sessions (from eight speakers) and test on the other session (from the other two speakers). This assures that the evaluation is speaker-independent. To tune hyperparameters and decide early stopping, we reserve 20% of training data as the validation set for each rotation.

4.2. The Baseline Models

We use the SVM based ASER system described in [4] as one of the baseline models. We extract frame-level features that include 13 Mel-frequency cepstral coefficients (MFCCs), first four formant frequencies and bandwidths, zero-crossing rate (ZCR), fundamental frequency (F_0), root-mean-square (RMS) energy and their first and second-time derivatives, totaling 72 features per frame. We apply mean, std, min, max, and range functionals to frame-level features to obtain utterance-level features, which have a dimensionality of $72 \times 5 = 360$. We normalize each dimension of the utterance-level features of the entire training samples to the range between 0 and 1; we normalize the test data using the same scaling factor. We then train a one-against-all binary SVM for each emotion category, with a radial-basis function kernel. During testing, we calculate the probabilities for each class and select the maximum one as the final emotion class for each test sample.

We design another CNN-FC network as our second baseline system. It takes the same hand-crafted features used in the SVM baseline with a temporal length of 64 (approximately 1 second) as inputs to the CNN encoder network. The CNN output is then fed to an FC network for classification. The architectures for the CNN encoder and the classifier are shown in Table 1. Note that the input dimension of the encoder is different, which is $64 \times 72 \times 1$. We train this network with Adam optimizer and 0.0002 learning rate. We adopted early stopping criteria, where the training stops if the validation loss is not improved for four epochs.

For the third baseline, we construct another CNN-FC network to take the log-Mel spectrogram directly as input, the same as the proposed four networks. This is to directly test the benefit of the adopted four unsupervised feature learning methods. For this purpose, we use the CNN encoder and FC classifier shown in Table 1 and train them from scratch. The resulting system, however, yielded very poor results, close to the chance performance. Therefore, we do not include it in Figure 2. We believe that the poor results were due to the scarcity of the training data (only 6136 samples) and the complexity of the CNN network taking log-Mel spectrogram inputs.

4.3. Proposed Models

The AEs presented in Section 3 are trained by an Adam optimizer with a learning rate of 0.0002. As for the training dataset, we select the Librispeech automatic speech recognition (ASR) corpus [28], which contains read speech that is often emotionally neutral. We calculate a 64-bin log-Mel spectrogram for each utterance with a 32 ms window size and a 16 ms hop size. We normalize the spectrogram values between 0 and 1 per utterance. We form training batches with a size of 256, by selecting random segments with a temporal length of 64 (approximately 1 seconds) from the utterances. The AEs are trained for 200 epochs.

The proposed ASER systems described in Section 3 are trained with the four pre-trained inference models (encoders), whose parameters are frozen, by an Adam optimizer with a learning rate of 0.001. We adopt an early stopping criterion, where training ends if the validation loss is not improved for four epochs. The emotion models are trained up to 50 epochs. The number of samples in each training batch is set to 256.

4.4. Results

We report the unweighted accuracy ratings (UARs) and F1-score in Figure 2 for the SVM and CNN baselines and the proposed systems. Several interesting observations are made. First, the CNN baseline yields slightly better UAR and F1-score than the SVM method. This suggests that deep models, taking the same hand-crafted features as inputs, outperform shallow models. Second, for both metrics, we are able to verify that the DAE-based unsupervised feature learning method using an external emotion-neutral dataset improves the ASER performance over SVM and CNN baselines that do not have the unsupervised feature learning module. This suggests that the learned features from the external emotion-neutral dataset are better than hand-crafted features (SVM baseline) and deep features learned only on the emotion dataset (CNN baseline). Third, the latent variable models VAE, AAE, and AVB outperform the DAE model in terms of both metrics, although they learn features from the same external dataset. This suggests that the latent variable models capture the more discriminative inherent structures of speech data than the reconstruction models such as the DAE. Fourth, adversarial models AAE and AVB achieve the best result, showing the importance of GAN loss on feature learning. In particular, AVB, which defines the GAN loss on input-code pairs, behaves the best.

5. CONCLUSION

In this work, we systematically explored the unsupervised methods in the context of ASER. We utilize unsupervised methods namely, DAE, VAE, AAE, AVB and trained on general speech, and use the learned features for ASER task. We show that these methods yield UAR and F1-score increase over the SVM and CNN baselines. Furthermore, we demonstrated that the inference models VAE, AAE, and AVB, outperform the reconstruction model DAE for unsupervised feature learning for ASER.

6. REFERENCES

[1] Johannes Wagner, Jonghwa Kim, and Elisabeth André, "From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification," in *International Conference on Multimedia and Expo (ICME)*. IEEE, 2005, pp. 940–943.

[2] Hatice Gunes and Massimo Piccardi, "Bi-modal emotion recognition from expressive face and body gestures," *Journal of Network and Computer Applications*, vol. 30, no. 4, pp. 1334–1345, 2007.

[3] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[4] Sefik Emre Eskimez, Kenneth Imade, Na Yang, Melissa Sturge-Apple, Zhiyao Duan, and Wendi Heinzelman, "Emotion classification: how does an automated system compare to naive human coders?," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2274–2278.

[5] Sinno Jialin Pan and Qiang Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[6] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 1096–1103.

[7] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling, "Semi-supervised learning with deep generative models," in *Advances in Neural Information Processing Systems*, 2014, pp. 3581–3589.

[8] Geoffrey E Hinton and Ruslan R Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

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

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[11] Daniel Neiberg, Kjell Elenius, and Kornel Laskowski, "Emotion recognition in spontaneous speech using gmms," in *Ninth International Conference on Spoken Language Processing*, 2006.

[12] Kiran K Rachuri, Mirco Musolesi, Cecilia Mascolo, Peter J Rentfrow, Chris Longworth, and Andrius Aucinas, "Emotionsense: a mobile phones based adaptive platform for experimental social psychology research," in *Proceedings of the 12th ACM international conference on Ubiquitous computing*. ACM, 2010, pp. 281–290.

[13] Kalani Wataraka Gamage, Vidhyasaharan Sethu, Phu Ngoc Le, and Eliathamby Ambikairajah, "An i-vector gplda system for speech based emotion recognition," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific*. IEEE, 2015, pp. 289–292.

[14] Tin Lay Nwe, Say Wei Foo, and Liyanage C De Silva, "Speech emotion recognition using hidden markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.

[15] Björn Schuller, Gerhard Rigoll, and Manfred Lang, "Hidden markov model-based speech emotion recognition," in *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*. IEEE, 2003, vol. 1, pp. I–401.

[16] Björn Schuller, Dejan Arsic, Frank Wallhoff, Gerhard Rigoll, et al., "Emotion recognition in the noise applying large acoustic feature sets," *Speech Prosody, Dresden*, pp. 276–289, 2006.

[17] Dmitri Bitouk, Ragini Verma, and Ani Nenkova, "Class-level spectral features for emotion recognition," *Speech communication*, vol. 52, no. 7, pp. 613–625, 2010.

[18] Na Yang, Jianbo Yuan, Yun Zhou, Ilker Demirkol, Zhiyao Duan, Wendi Heinzelman, and Melissa Sturge-Apple, "Enhanced multiclass svm with thresholding fusion for speech-based emotion classification," *International Journal of Speech Technology*, vol. 20, no. 1, pp. 27–41, 2017.

[19] Rui Xia and Yang Liu, "Using denoising autoencoder for emotion recognition.," in *Interspeech*, 2013, pp. 2886–2889.

[20] Rui Xia, Jun Deng, Bjorn Schuller, and Yang Liu, "Modeling gender information for emotion recognition using denoising autoencoder," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 990–994.

[21] Sayan Ghosh, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer, "Learning representations of affect from speech," *arXiv preprint arXiv:1511.04747*, 2015.

[22] Sayan Ghosh, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer, "Representation learning for speech emotion recognition.," in *INTERSPEECH*, 2016, pp. 3603–3607.

[23] Jun Deng, Zixing Zhang, Erik Marchi, and Bjorn Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2013, pp. 511–516.

[24] Jun Deng, Zixing Zhang, Florian Eyben, and Bjorn Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1068–1072, 2014.

[25] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[26] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger, "Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks," *arXiv preprint arXiv:1701.04722*, 2017.

[27] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.

[28] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.