IMPROVING EMOTION CLASSIFICATION THROUGH VARIATIONAL INFERENCE OF LATENT VARIABLES

Srinivas Parthasarathy[†] *Viktor Rozgic*[⋆] *Ming Sun*[⋆] *Chao Wang*[⋆]

[†] University of Texas at Dallas [⋆] Amazon Alexa

ABSTRACT

Conventional models for emotion recognition from speech signal are trained in supervised fashion using speech utterances with emotion labels. In this study we hypothesize that speech signal depends on multiple latent variables including the emotional state, age, gender, and speech content. We propose an Adversarial Autoencoder (AAE) to perform variational inference over the latent variables and reconstruct the input feature representations. Reconstruction of feature representations is used as an auxiliary task to aid the primary emotion recognition task. Experiments on the IEMOCAP dataset demonstrate that the auxiliary learning tasks improve emotion classification accuracy compared to a baseline supervised classifier. Further, we demonstrate that the proposed learning approach can be used for the end-to-end speech emotion recognition, as its applicable for models that operate on frame-level inputs.

Index Terms— Emotion recognition, Autoencoder, Adversarial training

1. INTRODUCTION

Affect sensing plays an important role in many health-care, education, and security related scenarios. Therefore, emotion recognition should be an integral part of modern human computer interaction systems. While the emotion recognition systems can use multi modal inputs (e.g, neoro-physiological, visual, speech) speech remains a primary input due to its prevalence [1]. Typically, the speech emotion recognition is performed in a supervised fashion using short, carefully segmented utterances, with labels that can take two formats - discrete categories such as happiness, sadness, anger and neutral [2], or continuous attributes such as activation (calm versus aroused), valence (negative versus positive) and dominance (weak versus strong) [3]. Prediction of emotional attributes has recently garnered more attention due to its flexibility in describing more complex emotional states [4]. For example, attributes can be used to distinguish various levels within an emotional category such as cold anger versus hot anger. In this study we focus on recognition of emotional attributes from speech.

The performance of a conventional speech emotion recognition system depends on the quantity and quality of labels used for supervision. The perception of speech emotion is a complex process due to various biases which makes the annotation task difficult [5]. Due to the task complexity and cost, annotation is usually done by few trained annotators. Most emotion recognition related public datasets contain a few thousand label utterances with 3-5 annotations per utterance. Lack of labeled data is a bottleneck for training more promissing deep learning models [6]. The recent advancements in generative modeling have enabled generative tasks to be used alongside discriminative classifiers [7]. There are two benefits to this. First, the generative task, such as the speech reconstruction, can be used as an auxiliary task to aid the primary classification task [8]. Second, such frameworks can be extended to semi-supervised scenarios.

In this paper we hypothesize that speech signal is produced via interaction of various latent factors including age, gender, emotional state, and content of speech. We pose the recognition of speech emotion as a latent variable inference problem and solve it using a variational inference procedure. Using Adversarial Autoencoder (AAE) [9], we perform a variational approximation of the true posterior distribution of the latent variables. The latent variables are split into a discrete component corresponding to the speaker's emotional state, and a continuous component capturing other latent factors. The AAE is trained to disentangle the discrete emotion distribution from the continuous component distribution. The input speech representation is then reconstructed from the approximated latent distributions. Besides the primary emotion classification task, the variational inference of latent variables and reconstruction of the input signal representations are used as unsupervised auxiliary tasks. Our experimental evaluations demonstrate that the proposed learning approach performs better than the typical fully supervised training. Further, we demonstrate that the proposed architecture is applicable both to a) Sentence-level input representations obtained by heuristic per-dimension aggregation of frame-level features, and b) Raw frame-level feature inputs. The latter implies applicability of the proposed framework to end-to-end recognition of emotion in speech.

The main contributions of this paper are: a) A novel framework for speech emotion recognition that employs variational inference of latent variables and reconstruction of the speech signal, and b) Demonstration that the proposed frame-

work is applicable to end-to-end speech emotion recognition.

2. RELATED WORK

Few recent studies used a variational approximation for emotion recognition. Sahu et al. [10] considered an adversarial autoencoder to extract features for emotion recognition. They matched the encoded distribution of the autoencoder to a 4 component GMM corresponding to 4 emotional categories. The encoded representation was used as a syntethic feature representation along with the original features. Eskimez et al. [11] similarly extracted features comparing the performance of various autoencoders. These studies propose a two-step process: 1) Learning general representations for speech reconstruction, and 2) training emotion recognizers on the obtained representations. We propose a one-step approach treating the emotion state as a part of latent representation used for reconstruction of speech signal.

Convolutional neural networks have recently gained popularity for emotion recognition tasks. Trigeorgis et al. [12] proposed a CNN architecture to replicate mel-frequency cepstrum coefficients *MFCCs* from raw speech waveform inputs. Most recent works have considered spectrogram features as inputs to the CNNs. Cummins et al. [13] built CNNs treating spectrograms as images. Yang and Hirschberg [14] predicted arousal and valence from spectrogram features. Few other studies have considered mel filter bank energies (MFBE) as inputs to CNN models. Aldeneh and Provost [15] used CNNs on 40 MFBEs as inputs to capture regional information emotion recognition and compared it with utterance level statistics. In this study we demonstrate that AAE modeling is applicable to CNN architectures that operate on frame-level feature (MFBE) inputs.

3. RESOURCES

3.1. Data

We test the proposed learning approach on the IEMOCAP corpus [16] that consists of dyadic interactions between 5 distinct pairs of actors who improvise a scripted conversation. The corpus contains 10,039 utterances from 10 speakers/actors. All utterances are annotated for activation, valence and dominance emotional attributes by 2 or 3 raters on a 5-likert scale. Similar to previous studies, we divide attribute dimensions into 3 classes with $\{1, 2\}$, $\{3\}$, and $\{4, 5\}$ representing low, neutral and high class respectively.

3.2. Features

We employ feature set used in the Interspeech 2013 paralinguistic challenge (IS2013_ComParE) [17]. For the experiments with frame-level inputs we use 65 low-level descriptors (LLD) extracted over 20 ms frames. A number of high level statistics are then computed on the LLDs and their deltas, over the entire utterance, resulting in 6,373 high-level functionals (HLF) which we used in experiments with utterance-level inputs.

4. METHODOLOGY

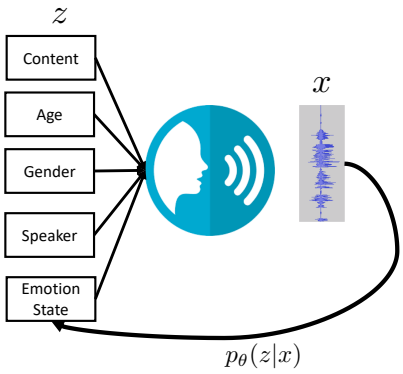

Fig. 1. Figure illustrates our hypothesis for speech production from latent variables and the inference of the latent variable as a posterior distribution on the observed variable

Latent speech production factors include speaker's characteristics (age, gender, accent, and speaker traits), speaker's emotional state, and content of speech.The emotion detection problem involves the inference of these latent variables. Figure 1 illustrates the overall framework. Given the observed variable x corresponding to the speech signal and the hidden variables z, then the inference of z given x is is based on the distribution $p_\theta(z \mid x)$ where θ models the data. The true posterior distribution $p_\theta(z \mid x)$ is computationally intractable and is approximated by the *variational* model $q_\phi(z \mid x)$. The goal is to optimize ϕ such that $q_\phi(z \mid x) \approx p_\theta(z \mid x)$. The optimization can be conducted by minimizing the KL divergence between $q_\phi(z \mid x)$ and $p_\theta(z \mid x)$. This leads to the formulation of the variational lower bound (Equation 1).

$$KL(q_\phi(z \mid x)||p_\theta(z \mid x)) = \log p_\theta(x) - E_{q_\phi(z|x)}[\log p_\theta(x \mid z)] + KL(q_\phi(z \mid x)||p_\theta(z)) \quad (1)$$

where $p_\theta(z)$ is a prior distribution over the latent variable z. Minimization of the divergence between the true and the approximate posterior also decreases the difference between the marginal log-likelihood $\log p_\theta(x)$ and the variational lower bound $E_{q_\phi(z|x)}[\log p_\theta(x \mid z)] - KL(q_\phi(z \mid x)||p_\theta(z))$. The optimization process resembles an autoencoder (ϕ parameterizes the encoder and θ parameterizes the decoder) training where the objective is to maximize likelihood of generated data given the observed data x.

$$\max_{\phi,\theta} E_{p_{D(x)}} E_{q_\phi(z|x)}[\log p_\theta(x \mid z)] - KL(q_\phi(z \mid x)||p_\theta(z)) \quad (2)$$

The second term in Equation 2 acts as a regularizer that forces the approximate posterior to match a prior distribution. Given enough capacity the encoder ϕ is able to produce a distribution that matches the true posterior and the decoder is able to generate data likely to be seen in the dataset.

4.1. Adversarial Autoencoder

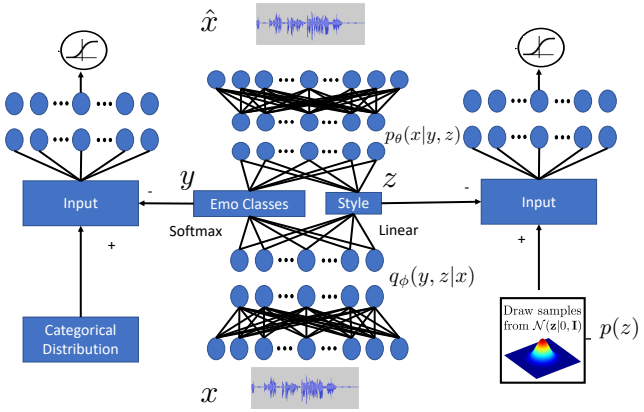

Fig. 2. Figure illustrates the proposed AAE model

We employ AAE to optimize Equation 2. The first term in the equation represents the reconstruction loss of the autoencoder. The second term is optimized through adversarial training. The adversary learns to discriminate between 'true' samples sampled from $p_\theta(z)$ and 'fake' samples sampled from an aggregated posterior $q_\phi(z)$. This adversarial training has shown to produce a posterior distribution that better matches the prior distribution [9]. Moreover, with adversarial training we only need to sample from the two competing distributions, whereas directly minimizing the KL divergence requires a prior knowledge of the functional form for the loss. This allows AAE the flexibility of choosing any distribution for the prior $p_\theta(z)$. Figure 2 illustrates the AAE architecture used in this study. The encoder maps the speech signal x to the latent space which is divided into two components - a discrete component y that captures the emotion variability in speech and a continuous component z that captures all the other speech variabilities. The posterior distribution is denoted $q_\phi(y, z \mid x)$. These aggregated posteriors $q_\phi(y)$, $q_\phi(z)$ are matched to a categorical and a Gaussian prior. Note that the categorical posterior $q_\phi(y \mid x)$ also represents the emotion classifier. The decoder $p_\theta(x \mid y, z)$ then maps the joint latent space back to the data distribution. The training of the autoencoder is done in 3 phases. First, in the reconstruction phase, the autoencoder is trained from top to bottom by minimizing a reconstruction loss between the reconstructed signal $\hat{x}$ and the true signal x. The second phase is a regularization phase where both posterior distributions are matched to their respective priors by training adversarial networks. In the third phase, the categorical encoder $q_\phi(y \mid x)$ is trained using a cross entropy loss between the predicted labels y and the true emotion labels. This framework can be used in a semi-supervised setting when limited labeled data is available. The reconstruction and regularization phases are learned without the emotion labels and therefore are unsupervised with respect to the primary task.

5. EXPERIMENTAL EVALUATION

5.1. Baseline

We first implement a completely supervised architecture as a baseline for the proposed AAE model. The baseline architecture is a 2-layer fully connected neural network with 256 nodes in each layer. A ReLU activation was used at the hidden layer and a softmax activation was used at the output layer for classification. To regularize the model a dropout of 0.5 was used between the hidden layers. The model was optimized with a Nadam optimizer and a learning rate of 1e-4. The model was trained for 100 epochs

We evaluate our models using speaker independent tests. The IEMOCAP dataset is divided into 10 speaker independent folds. For each test, data belonging to 8 speakers from 4 sessions of the dataset are used to train the models. Data from Speakers in the remaining session are used to mutually validate and test the models i.e one speaker's data is used to validate the model and the other is used to test the model and vice versa. We evaluate the models based on the unweighted accuracy over the 3 emotion classes. We train the models for 100 epochs and use the model parameters that perform best on the validation set for the test set evaluation. While we do this for each fold independently, we report the unweighted accuracy over the entire dataset by accumulating the predictions over all folds. Experiments are run 10 times with different initializations, to avoid the effect of random seeds in our evaluation. The first row in Table 1 reports the baseline accuracy comparable to previous related studies (Section 2). As shown in previous studies, recognizing valence from acoustic features remains harder task then recognizing activation and dominance.

5.2. Proposed AAE architecture

We select parameters of the AAE based on the validation set performance. We use a 2 layer neural network with 256 nodes per layer to parameterize the encoder and the decoder. The ReLU function is used as activation function and 0.5 dropout is used between the encoder's hidden layers. Note that the encoder complexity (i.e., number of parameters) is comparable to the baseline architecture. The latent space is divided into two components, y representing the categorical emotion distribution and z the continuous style distribution. The dimension of y, constrained by the emotion class number, equals 3. The dimension of z was fixed at 100. For the prior distributions to match the latent space, we used a categorical distribution with equal probabilities of choosing each class

for $p_\theta(y)$ and for the continuous prior $p_\theta(z)$, we use standard Gaussian distribution, i.e. $p_\theta(z) \sim \mathcal{N}(0, I)$. Both discriminators for adversarial training consist of two 256-node layers. A sigmoid activation is used at the discrimnators' outputs. The emotion classifier $q_\phi(y)$ is trained using the cross entropy cost function. The model was trained for 100 epochs. The best parameters on the validation set were used to for testing. The performance of the vanilla AAE is shown in the second row of the Table 1.

The vanilla AAE performs worse than the utterance level baseline. We identified several factors that were important for success of the AAE training:

- Batch Normalization: We normalized all layers including the softmax and linear layer at the output of the encoder and the linear layer at the end of the decoder.
- Discriminator regularization: A dropout of 0.5 in the discriminator layers improved results.
- "Strengthened" generator: The discriminators of the adversarial components were trained with 0.1*generator learning rate.

We conducted additional experiments to evaluate impact of the different auxiliary tasks. We observed that the extra regularization phase, where the latent space is matched to the prior distribution, was important in improving performance compared to using only the auxiliary reconstruction loss. The third row in Table 1 contains performance of the optimized AAE models. We confirmed significance of the performance improvement for all emotion dimensions using a one-tailed t-test with a p-value 0.05.

5.3. Emotion recognition with frame-level features

The experiments described in the previous session rely on hand engineered utterance level representations. There has been a push in the scientific community to move towards a data driven representation learning. Some of the recent works on speech emotion recognition focused on learning representations from log mel frequency band energies, spectrograms or even the audio waveforms (Section 2).

In this section we address speech emotion recognition using the frame-level features as modeling input. Starting with the 65 frame level LLDs from the Interspeech 2013 paralinguistic feature set (Section 3.2) as inputs, we build convolutional neural networks (CNN) to learn feature representations necessary for the discriminative task. This method is a step towards end-to-end learning where we avoid the brute force computation of multiple high level statistics over the utterance and, instead, enable the model to learn necessary abstractions. Similar to the experiments in Section 5.1 and Section 4.1 we train both the baseline and the AAE models. For the baseline model, we use multiple convolutional and max pooling layers followed by dense layers and a softmax classifier. The encoder of the AAE uses the same architecture. The decoder employs dense transpose, convolutional transpose layers and unpooling layers. The architecture specifications are summarized in Table 2.

Type	Act	Val	Dom
UTT + Base	77.74 ± 0.35	61.92 ± 0.22	63.94 ± 0.48
UTT + AAE	75.30 ± 0.53	60.94 ± 0.53	63.53 ± 0.36
UTT + AAE Optimized	78.08 ± 0.23	63.66 ± 0.42	65.32 ± 0.33
FR + Base	78.19 ± 0.50	62.14 ± 0.64	64.64 ± 0.72
FR + AAE	78.42 ± 0.38	64.45 ± 0.37	65.45 ± 0.24

Table 1. Unweighted accuracy for utterance & frame models. *Act* - Activation, *Val* - Valence and *Dom* - Dominance

Type	Description
CNN Baseline & AAE encoder	Conv1D {4, 128, Stride 1, Max Pool 4}x2 Conv1D {4, 256, Stride 1, Max Pool 4}x2 Conv1D 4, 256 Dense {256 }x2 Softmax 3
AAE Decoder	Dense {256 }x2 Conv1D Trans 4, 256 Conv1D Trans {4, 256, Stride 1, Unpool 4}x2 Conv1D Trans {4, 128, Stride 1, Unpool 4}x2

Table 2. Convolutional architectures

Rows 4 and 5 of Table 1 show the results for the baseline and proposed architectures with frame level feature inputs. We see that the performance with utterance level features can be replicated using frame-level features. Further, the AAE with frame-level features outperforms the frame level baseline achieving the best performance amongst all trained models.

6. CONCLUSION

In this paper we examined hypothesis that speech is produced by multiple latent factors, including emotional state of the person. The latent variables were separated into emotional component and a style component corresponding to all other factors. The emotion recognition was performed through the variational inference paradigm. We used the adversarial autoencoder to approximate the posterior distribution of the latent variables. The latent variables were matched to prior distributions using an adversarial network. We demonstrated that the proposed approach is applicable to fully-connected network models operation on utterance-level features and convolutional neural network models operating on frame-level features. For both model types the proposed learning approach outperformed fully supervised training. Our future work will include experimentation on larger partially annotated datasets.

7. REFERENCES

[1] Björn W Schuller, "Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends,"

Communications of the ACM, vol. 61, no. 5, pp. 90–99, 2018.

[2] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards more reality in the recognition of emotional speech," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, Honolulu, HI, USA, April 2007, vol. 4, pp. 941–944.

[3] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10-11, pp. 787–800, October-November 2007.

[4] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1103–1107.

[5] A. Metallinou and S.S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE 2013)*, Shanghai, China, April 2013.

[6] M. Abdelwahab and C. Busso, "Study of dense network approaches for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 5084–5088.

[7] Jost Tobias Springenberg, "Unsupervised and semi-supervised learning with categorical generative adversarial networks," *arXiv preprint arXiv:1511.06390*, 2015.

[8] S. Parthasarathy and C. Busso, "Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes," *ArXiv e-prints*, pp. 1–5, April 2018.

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

[10] Saurabh Sahu, Rahul Gupta, Ganesh Sivaraman, Wael AbdAlmageed, and Carol Espy-Wilson, "Adversarial auto-encoders for speech based emotion recognition," *arXiv preprint arXiv:1806.02146*, 2018.

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

[12] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5200–5204.

[13] Nicholas Cummins, Shahin Amiriparian, Gerhard Hagerer, Anton Batliner, Stefan Steidl, and Björn W Schuller, "An image-based deep spectrum feature representation for the recognition of emotional speech," in *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 2017, pp. 478–484.

[14] Zixiaofan Yang and Julia Hirschberg, "Predicting arousal and valence from waveforms and spectrograms using deep neural networks," *Proc. Interspeech 2018*, pp. 3092–3096, 2018.

[15] Z. Aldeneh and E. Mower Provost, "Using regional saliency for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 2741–2745.

[16] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, December 2008.

[17] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Interspeech 2013*, Lyon, France, August 2013, pp. 148–152.