Semi-supervised Model for Emotion Recognition in Speech: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part I

Chapter · September 2018

DOI: 10.1007/978-3-030-01418-6_77

Alexandre M. A. Maciel
Universidade de Pernambuco

Semi-supervised Model for Emotion Recognition in Speech

Ingryd Pereira[1(✉)], Diego Santos[2(✉)], Alexandre Maciel[1(✉)],
and Pablo Barros[3(✉)]

[1] Polytechnic School of Pernambuco, University of Pernambuco, Recife, Brazil
{ivstp,amam}@ecomp.poli.br

[2] Fedreal University of Pernambuco, Recife, Brazil
dgs2@ecomp.poli.br

[3] Knowledge Technology, Department of Informatics, University of Hamburg,
Hamburg, Germany
barros@informatik.uni-hamburg.de

Abstract. To recognize emotional traits on speech is a challenging task which became very popular in the past years, especially due to the recent advances in deep neural networks. Although very successful, these models inherited a common problem from strongly supervised deep neural networks: a large number of strongly labeled samples demands necessary, so the model learns a general emotion representation. This paper proposes a solution for this problem with the development of a semi-supervised neural network which can learn speech representation from unlabeled samples and used them in different emotion recognition in speech scenarios. We provide experiments with different datasets, representing natural and controlled scenarios. Our results show that our model is competitive with state-of-the-art solutions in all these scenarios while sharing the same learned representations, which were learned without the necessity of strong labeled data.

Keywords: Emotion recognition · Semi-supervised learning · GAN
Speech representation · Deep learning

1 Introduction

Recent advances in deep learning provided an increase in popularity and robustness on emotion recognition in speech tasks [14,20,23]. Such models usually make use of a large number of labeled samples to learn general representations for emotion recognition, providing state-of-the-art results in different speech related scenarios [2,8,21].

However, supervised deep learning needs a lot of labeled training data. Another problem with the current supervised deep learning models lies in the nature of emotion description itself. Different persons can express and perceive the same emotion in many ways, which causes a lack of agreement about how to

V. Kůrková et al. (Eds.): ICANN 2018, LNCS 11139, pp. 791–800, 2018.
https://doi.org/10.1007/978-3-030-01418-6_77

annotate samples from different scenarios [7]. One solution for this is the use of an even larger number of labeled samples to represent different emotional states into a general emotion categorization.

The use of unsupervised learning becomes useful to solve this problem since it does not require labeled data to learn general speech representation, which can be transferred to emotions. To work around this problem, recent works like [1,15,16,19] apply semi-supervised training on deep neural networks for image classification in domains where the labeled date is scarcity.

If we train a deep neural model with a dataset from a given domain, the model will specialize in that scenario. But a model specialized on to generate a general representation of the data will be capable of representing the audio in every presented scenario. To be able to be general enough, deep learning models for speech emotion recognition usually rely on a large number of labeled samples. This issue happens because (1) deep neural models need a large number of samples to learn descriptors which are robust enough to generalize the domain where they are applied. (2) Strongly supervised training produces a fast and more focused change on the gradient directions, which usually leads to a better fine-tuning of the descriptors and separation boundaries for classification.

We propose a hybrid neural network, composed of an adversarial autoencoder to learn general speech representations and use it as input to a strongly supervised model to classify emotion expressions in the speech in different scenarios. In the first step, the model learns how to represent the audio through an unsupervised training process. This representation will be the input for the second step where the model learns the separation boundaries and distribution between classes through a supervised learning process. In the unsupervised step, a Generative Adversarial Network (GAN) trains an autoencoder that will be responsible for learning how to represent speech present in the audio. As a GAN has unsupervised training, the model can use unbranded and not emotional data what possibilities that the use of the trained model over different scenarios. After training, the encoder filters can extract prosodic characteristics of the input speech without the necessity of supervised labels. The encoder ends up learning representations based on the data distribution. The second module of the proposed model uses these prosodic characteristics learned by the encoder as low-level feature representations, and, now using a strongly supervised solution, is trained to classify emotion recognition in speech. A set of different filters also composes the classifier. These filters are fine-tuned and learn high-level abstractions of the input signal, which are pertinent to that specific domain.

We make used of an unconstrained and unlabeled corpus to learn general speech representations, which is shared among all our emotion recognition scenario. Our specific classifiers are fine-tuned to specific emotion recognition high-level characteristics. This reduces the training effort and applicability of the model to different emotion recognition scenarios.

So, in emotion recognition task, the use of general speech representation, training in an unsupervised manner, improve the application performance and also build an adaptive model for others scenarios, once the speech representation doesn't be stuck in the scenario of the dataset evaluated. The main contribution of this proposition is the general speech recognition model. This model can fit in different emotional recognition scenarios and different datasets without retraining. In other emotional recognition works the audio representation ends up stuck in the scene obtained from the training dataset. In our proposition the audio representation is more robust, being able to represent different domains, situations, and languages.

We evaluate the performance of our model in three different scenarios: indoor, outdoor and cross-language and compare it with state-of-the-art solutions. We prove that our model learned a general speech representation which is shared among all these scenarios, and the different specific filters learn high-level abstractions which are unique for each of these scenarios. For that, we use three different datasets: the Surrey Audio-Visual Expressed Emotion Dataset (SAVEE) [13] which represents a controlled environment, usually found in indoor scenarios or simple interactions, the OMG Emotion Dataset [3], which represents an in-the-wild, outdoor, unrestricted scenario and finally the Berlin Database of Emotional Speech (EmoDB) [6] which evaluates how well the learned representations learned with speech signals in one language can be transferred to other for emotion recognition. This way, we can prove the universal aspect of emotion recognition, and that our fine-tuning step learns to correlate the emotional aspects of the general speech representation, ignoring the information which is not necessary for this task.

2 Proposed Model

In this work, we propose a semi-supervised model for emotion recognition. The model contains two modules: the first one is the general speech representation and the second one is the classifier model. Figure 1 presents the model illustration.

The training of the first module of our network happens in an unsupervised way. The first model is composed of an autoencoder trained by a GAN. We use the encoder present in the autoencoder model for learning the general speech representation. The GAN was chosen because have an autoencoder in its structure and allows an adversary training with a large amount of unlabeled data. The speech representation generated by the model will be the input for the second part of the model.

The second part is responsible for the distribution between the classes in an emotion classification or for prediction from the values in a dimensional model. The training of this module is in a supervised way. We adapt the output of the classifier accordingly to the task: or we use binary classification for categorical emotions (e.g., anger, fear, happiness, etc.) or we use a double-head one unit structure, for arousal/valence regression.

2.1 Adversarial Autoencoder

The Generative Adversarial Network (GAN) [12] has had a significant impact on data generation, mainly of images, but also in audio applications, for example

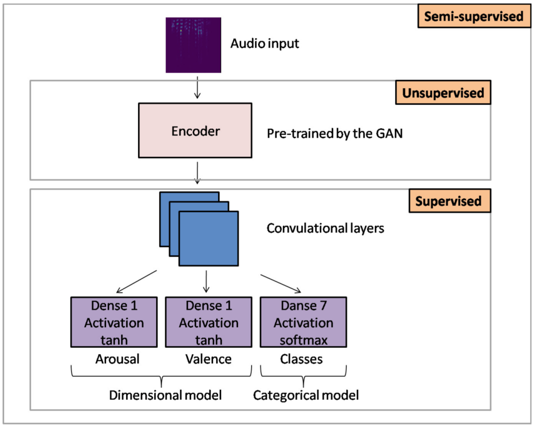

Fig. 1. Abstraction of the classifier and prediction models

for melody generation [22] and noise cleaning [18]. The basic idea of a GAN is to conduct unsupervised adversarial training in two artificial neural networks, a discriminator model (D) and a generator model (G). The training process occurs similarly to a minimax two-player game, in which G captures the data distribution, and D estimates the likelihood of an example coming from G to be real. The G training procedure is to maximize the probability of D making a mistake.

The Boundary Equilibrium Generative Adversarial Networks (BEGAN) [5] is a GAN variation, and have a differential are the use of an autoencoder as a discriminator. Others particularities of BEGAN is the loss derived from Wasserstein's distance; the addition of a γ variable to balance GAN training; and the addition of a new metric called *m global*.

The training of the basic GAN, proposed by Goodfellow [12], G e o D is trained in an The training of the basic GAN, proposed by Goodfellow [12], G and the D are trained in an adversarial way. Figure 2 presents the training representation of a GAN. In this figure, x represents the real samples, $\ddot{x}$ represents the generated samples by the generator and z is the generated noise that is the generator inputs. The training of D has two different moments: first one where the inputs are real samples and the expected output is the real class that example belongs to (i.e., class 1). The second one where the inputs are samples generated by the generator from the noise and the expected output is a fake classification (i.e., class 0). For the generator, the flow is: the generator module receives as input a noisy, and a fake sample generated as an input of the discriminator. The objective is to make this sample be confused with a real sample, so the expected output is a real classification by the discriminator. In this step the discriminator training is frozen, and only training the generator.

A characteristic BEGAN is the application of a balance paired with a loss derived from Wasserstein's distance to the autoencoder training [5]. In the

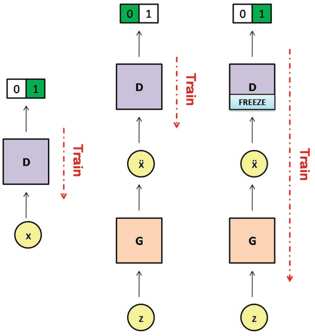

Fig. 2. Training representation of discriminator (D) and generator (G)

training step, the BEGAN has a balancing factor, defined by a variable γ, with a range of 0 to 1. This variable penalizes D training, slowing it down. Since G training is more difficult and slower than D, this penalty balances the algorithms, thus increasing the performance of GAN [5].

GANs were used recently on semi-supervised learning for image classification tasks [1,15,16,19] and was shown to be more effective than strongly supervised classification. That happens because the use of unsupervised training makes possible to the model to learn general representations of the domain, while the supervised fine-tuning specializes in the model to solve the specific tasks. We choose to use a variation of an adversarial autoencoder, the BEGAN [5] because it presented better results than common GANs results on learning general representations.

2.2 Supervised Classifiers

The supervised module of the proposed model varies according to the emotion recognition scenario. But the basic structure is: it receives as input the speech representation obtained by the unsupervised module, then it applies convolutional layers and a softmax classifier which is adapted depending on the scenario.

We optimize the hyper-parameters of the supervised module for each task. For that, we use the Hyperas [4] framework, where is specialized in optimizing search spaces with real, discrete and conditional dimensions.

2.3 Semi-supervised Learning

The adversarial autoencoder will be pre-trained with a database with a larger number of data. Once trained the autoencoder, don't need to retrain this model and this same autoencoder can be reuse in others applications, also without the need of retraining.

The supervised model training happens during the semi-supervised model training process. In this process, we freeze the encoder layers trained previously and train only the supervised module. The layers of the autoencoder are freeze because if it is trained too with the evaluated dataset, lose the nature of general speech representation, specializing in the dataset scenario.

3 Experimental Methodology

3.1 Datasets

We use one dataset to train the unsupervised part of our model, and three to evaluate the whole model in different scenarios. The LibreSpeech [17] dataset is one of the largest audio datasets available, and we use it to train the unsupervised module. We use this dataset because its amount of data and variability of speakers and scenario is interesting to generate a general representation of speech. LibriSpeech is a dataset with approximately 1000 h of English speech.

We use three others datasets, and these datasets are emotional, multimodal and multispeaker. Each one has different characteristics and scenarios, which possibility different analysis. Therefore, we evaluated our model in an indoor, outdoor and cross-language scenarios, witch SAVEE [13], OMG Emotion [3] and EmoDB [6] datasets, respectively.

SAVEE. We used the Surrey Audio-Visual Expressed Emotion Dataset (SAVEE) [13] in our experiments. SAVEE is an emotional audiovisual dataset, with consists of recordings of four male actors speaking phrases in 7 different emotion intonations based on the Universal Emotions [11] with the addition of the neutral emotion, where the speaker not present any of the six universal emotions.

This work uses only the auditive module of the dataset, and has 480 statements in total. The SAVEE database is balanced, recorded in a controlled and noise-free environment and only has male voices. Therefore is considered a simple base and applied as the starting point of the experiments.

OMG Emotion Dataset. The One-Minute Gradual-Emotional Behavior dataset (OMG-Emotion) [3] is the database from the One-Minute Gradual-Emotion Behavior Challenge, which takes place at IJCNN 2018. The dataset contains 567 unique videos totaling 7371 clips each clip consisting of a single utterance. Each video has a different utterance number with an average duration of 8 s by utterance and total average video duration next to 1 min.

The dataset has dimensional and categorical labels, being seven different emotions, based on the Universal Emotions [11] with the addition of the neutral emotion. The dataset also has continuous dimensional label being arousal and valence with values in a range between -1 and 1. OMG emotion dataset is a complex, given its variability of speakers, scenarios, dialogs and videos duration. The dataset labels are either categorical and dimensional, what makes possible to verify the proposed model performance in different emotional recognition tasks.

EmoDB. The Berlin Database of Emotional Speech (EmoDB) [6] is an emotional speech database recorded in German. It contains about 500 utterances spoken by the actors in a happy, angry, anxious, fearful, bored and disgusted manner, as well as in a neutral version. It has statements from 10 different actors and ten different texts. We used the EmoDB dataset to verify it the proposed model can also generalize emotional characteristics from other languages.

3.2 Preprocessing

Our first preprocessing step was to change the audio frequency to 16 kH. Then each audio track was decomposed into 1-s chunks without overlapping. After that, the raw audio was converted to a spectrogram via Short Time Fourier Transform, with an FFT of size 1024 and a length of 512.

3.3 Experiments Setup

To evaluate our model on the SAVEE dataset, we train the BEGAN with part of the LibreSpeech dataset but evaluating the emotion classification model with SAVEE dataset. To be possible to compare with another work, this experiment follows the same protocol of Ashwin work et al. [2], where perform the job of classifying emotions present in audio and video proposing a novel hybrid SVM-RBM classifier. We compare just with the audio module. Ashwin et al. perform the experiment called dependent speaker, which uses each speaker sets for training, and for each evaluated test of each speaker (DC, JE, JK, KL). The division of the base is approximately 60% for training and 40% for testing.

Experiments were also carried out with the OMG Emotion dataset, with categorical and dimensional labels, which allows the evaluation of two emotion recognition tasks: the classification of static emotion and prediction of dimensional values arousal and valence. For all experiments with OMG Emotion dataset, the training process of BEGAN uses part of the LibreSpeech dataset, and the division of the training and testing process follows the same distribution made available in the database itself.

The experiments performed on EmoDB dataset follow the Leave One Speaker Out protocol (LOSO) to be possible perform the comparative with other works that follow the same protocol. In the experiment, we train the BEGAN with part of the LibreSpeech dataset recorded in English and the model evaluated on EmoDB dataset which is one German language recorded database.

We train the algorithms in each experiment with 100 epochs, with a batch size of 16. The discriminator and the generator of the BEGAN used the Adam optimizer with a learning rate of 0.00005. The BEGAN also has a *gamma* value that balances the generator and the discriminator with a value of 0.7.

4 Results and Discussion

Table 1 shows the accuracy averages achieved with ten executions of the model and the best results obtained in Ashwin's work [2]. The results obtained with this proposal are bigger than the related work. The standard deviation is small, so this means that the model proves to be stable.

Table 1. Comparison between the accuracy (%) averages

	DC	JE	JK	KL
Ashwin et al. [2]	79	78	76	80
This work	80.69 (±2.96)	80.96 (±3.41)	80.15 (±1.85)	82.46 (±2.70)

Table 2 presents the summary of the executions of the model when tested with the OMG Emotion Dataset, the baseline results [3], and the best result obtained in the challenge in audio modality[1]. The table has the F-score of the classifier and also has the CCC of the arousal and valence values predicted. The result F-score obtained with the classifier model was higher than the baseline, and has the advantage that a general speech representation was used and that it can be reused without the need of re-training in other datasets. The CCC obtained in our experiments is smaller than the result obtained in the challenge, but is better than baseline work. We obtained this result with the same model of the classification experiments, without specific treatment for this task and still the result is better than the baseline.

Table 2. Results with the OMG emotion dataset

	F-score	Arousal CCC	Valence CCC
Barros et al. [3]	0.39	0.07	0.04
OMG emotion challenge	-	0.29	0.36
This work	0.73	0.17	0.16

Table 3 presents the results from the executions with the EmoDB dataset and the comparison with other works that use the same experimentation protocol. As can be seen, our proposal is above of the related works. But considering that our model learns how to represent the emotional data in another language, the results can still be relevant for being next of the related works.

The BEGAN trained with LibreSpeech database used in our experiments perform the training process only once. After saving the model, it can execute different experiments without the need for retraining. The no reed of retraining is one of the principal advantages of the proposed approach since once trained the model; we can use it for different databases and several tasks without the need for retraining.

[1] https://www2.informatik.uni-hamburg.de/wtm/OMG-EmotionChallenge/.

Table 3. Results with EmoDB dataset

	Accuracy
Deb and Dandapat [10]	83.80%
Deb and Dandapat [9]	85.10%
This work	72%

5 Conclusion

The work proposed is the development a new semi-supervised model for emotion recognition tasks. The use of this algorithm can help overcome one of the common challenges of emotion recognition field, which is the speech representation.

We propose a general speech representation model, which is constructed with a GAN and trained in an unsupervised way and then incorporated into the models, thus building the semi-supervised model. From a set of experiments, with different datasets in the same algorithm, it was possible to verify that the use of GAN can help in the training of an emotion recognizer, that besides needing a smaller amount of training data in the supervised part, also achieves superior performance and provides a more stable algorithm.

In this work, experiments were performed with the SAVEE dataset, which is a simple dataset, and also with the OMG Emotion Dataset, which is a complex database, given its speakers and scenarios variability, and has categorical and dimensional labels. In the experiments, it was possible to verify that the proposed model is superior to the baseline, and also the benefit of using a speech representation model that can be reused in other models and other databases.

Experiment with a dataset of other language was performed. The speech representation module was trained with one dataset of the English language and was performed the emotion classification in a Germany dataset. The results were similar to related works used how baseline. This experiment proves that unsupervised model represents the speech emotional characteristics independent of the language.

As a continuation of this work will be carried out sets of experiments where BEGAN will be trained with different datasets, and the semi-supervised learning model will be evaluated with other datasets with different domains (e.g., a dataset with only children's voices, a dataset in other languages, etc.).

References

1. Adiwardana, D.D.F., Matsukawa, A., Whang, J.: Using generative models for semi-supervised learning
2. Ashwin, T., Saran, S., Reddy, G.R.M.: Video affective content analysis based on multimodal features using a novel hybrid SVM-RBM classifier. In: 2016 IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics Engineering (UPCON), pp. 416–421. IEEE (2016)

3. Barros, P., Churamani, N., Lakomkin, E., Siqueira, H., Sutherland, A., Wermter, S.: The OMG-emotion behavior dataset. arXiv preprint arXiv:1803.05434 (2018)
4. Bergstra, J., Yamins, D., Cox, D.D.: Hyperopt: a python library for optimizing the hyperparameters of machine learning algorithms. In: Proceedings of the 12th Python in Science Conference, pp. 13–20. Citeseer (2013)
5. Berthelot, D., Schumm, T., Metz, L.: Began: boundary equilibrium generative adversarial networks. arXiv preprint arXiv:1703.10717 (2017)
6. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B.: A database of German emotional speech. In: Ninth European Conference on Speech Communication and Technology (2005)
7. Cabanac, M.: What is emotion? Behav. Process. **60**(2), 69–83 (2002)
8. Chang, J., Scherer, S.: Learning representations of emotional speech with deep convolutional generative adversarial networks. arXiv preprint arXiv:1705.02394 (2017)
9. Deb, S., Dandapat, S.: Emotion classification using segmentation of vowel-like and non-vowel-like regions. IEEE Trans. Affect. Comput. (2017)
10. Deb, S., Dandapat, S.: Multiscale amplitude feature and significance of enhanced vocal tract information for emotion classification. IEEE Trans. Cybern. (2018)
11. Ekman, P.: An argument for basic emotions. Cogn. Emot. **6**(3–4), 169–200 (1992)
12. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
13. Haq, S., Jackson, P.J.: Multimodal emotion recognition. In: Machine Audition: Principles, Algorithms and Systems, pp. 398–423 (2010)
14. Huang, Z., Dong, M., Mao, Q., Zhan, Y.: Speech emotion recognition using CNN. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 801–804. ACM (2014)
15. Kingma, D.P., Mohamed, S., Rezende, D.J., Welling, M.: Semi-supervised learning with deep generative models. In: Advances in Neural Information Processing Systems, pp. 3581–3589 (2014)
16. Odena, A.: Semi-supervised learning with generative adversarial networks. arXiv preprint arXiv:1606.01583 (2016)
17. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: LibriSpeech: an ASR corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5206–5210. IEEE (2015)
18. Pascual, S., Bonafonte, A., Serrà, J.: SEGAN: speech enhancement generative adversarial network. arXiv preprint arXiv:1703.09452 (2017)
19. Springenberg, J.T.: Unsupervised and semi-supervised learning with categorical generative adversarial networks. arXiv preprint arXiv:1511.06390 (2015)
20. Trigeorgis, G., et al.: 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), pp. 5200–5204. IEEE (2016)
21. Weißkirchen, N., Bock, R., Wendemuth, A.: Recognition of emotional speech with convolutional neural networks by means of spectral estimates. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), pp. 50–55. IEEE (2017)
22. Yang, L.C., Chou, S.Y., Yang, Y.H.: MidiNet: a convolutional generative adversarial network for symbolic-domain music generation. In: Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017), Suzhou, China (2017)
23. Zheng, W., Yu, J., Zou, Y.: An experimental study of speech emotion recognition based on deep convolutional neural networks. In: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 827–831. IEEE (2015)