Speaker Awareness for Speech Emotion Recognition

https://doi.org/10.3991/ijoe.v16i04.11870

Gustavo Assunção (✉), Paulo Menezes
Institute of Systems and Robotics, Coimbra, Portugal
gustavo.assuncao@isr.uc.pt

Fernando Perdigão
Instituto de Telecomunicações, Coimbra, Portugal

Abstract—The idea of recognizing human emotion through speech (SER) has recently received considerable attention from the research community, mostly due to the current machine learning trend. Nevertheless, even the most successful methods are still rather lacking in terms of adaptation to specific speakers and scenarios, evidently reducing their performance when compared to humans. In this paper, we evaluate a largescale machine learning model for classification of emotional states. This model has been trained for speaker identification but is instead used here as a front-end for extracting robust features from emotional speech. We aim to verify that SER improves when some speaker's emotional prosody cues are considered. Experiments using various state-of-the-art classifiers are carried out, using the Weka software, so as to evaluate the robustness of the extracted features. Considerable improvement is observed when comparing our results with other SER state-of-the-art techniques.

Keywords—Speech emotion recognition, machine learning, CNN, VGG

1 Introduction

Emotion and its expression undoubtedly govern many aspects of human interaction. It is self-evident that the emotional phenomena experienced by a person should tend to mold their behavior and conversational register in the social settings they engage with. Effectively, by adulthood most humans will have developed a large set of highly nature/nurture dependent, distinct behavioral responses to the multiple emotional states they experience throughout their personal lives. An emulated corroboration of this notion can be inferred by the specific articulation patterns observed in professional actors simulating emotion [1]. Further support is given by the widely accepted effects of recurrent stress in an individual's emotional state [2], [3] naturally affecting their prosody.

The role of someone's personality, and consequently their way of communicating, are often overlooked when it comes to emotion recognition from speech. In fact, even though the state-of-the-art machine learning models are exceptionally competent at evaluating data with an unspecified set of relevant features, most of the designs tend to focus directly and solely on overt emotional cues. As of now, this should be con-

sidered an erroneous approach given the observed higher performances of systems with deeper adaptation levels, such as domain-based [4], [5] or context-based [6], [7]. Hence, more information channels should be considered when analyzing emotion in speech, one of which is potential speaker dependency of human emotional prosody.

Empirical evidence of prosody variations for identical emotional sates in different people is of relevance to interactive systems and other social robotic applications. Provided a machine can identify an intervening speaker, the ability to further adapt itself to not only said speaker but also to their emotional conveyance mannerisms, can certainly boost the quality of the system's behavior and response suitability to the situation at hand. This concept is not unlike how pet social companions are able to perceive how their owners feel and modify their behavior to conform with the respective emotional state. Emulation of this in machines would be highly useful.

We introduce our approach to speech emotion recognition (SER), based on the use of the speaker recognition CNN model VGGVox [8], for feature extraction from 6 standard and established emotional speech databases, with minimal preprocessing. The application of the features extracted using our technique, in state-of-the-art classifiers, confirmed that speaker specific features extracted from speech are robust enough to allow for clear classification of emotional states. Moreover, our technique's performance was shown to surpass that of other state-of-the-art methods.

This paper is divided in the following manner: section II provides an overview of recent related work while section III outlines the methodology of our approach. This is followed by section IV where detail is given about the experiments carried out and the obtained supporting results are discussed, and finally section V where a conclusion and overview of future work are presented.

2 Related Work

Given the necessity of evaluating a panoply of informational cues embedded in speech, added to the already complex task of considering as many vocal features as possible, most classical recognition systems based on speech have seen their performance greatly surpassed by machine leaning models. As such, these architectures have been used as baselines in the performance evaluation of new emotion recognition techniques which use representations learned from other paralinguistic tasks.

Gideon et al. [9] assessed the effectiveness of progressive neural networks (ProgNets) at freezing the weights of a model's initial layers, tuned for speaker recognition and gender detection from speech, and using these transitional representations as input to the posterior layers, trained for emotion recognition. Performance rates were somewhat higher than those of standard DNN or simple pre-training and fine-tuning (PT/FT) networks. On a separate note, Sidorov et al. [10] explored the effects of adding speaker specific and gender information as features in the vectors used to train emotion recognition models, essentially further detailing the datafiles in one experiment. Parallelly, the group predicted speaker and gender information with ANN-based recognizers, adding the obtained hypotheses to the feature sets fed into the used emotion recognizer. This plain method of extending the feature vector with

additional speaker specific information was found to improve emotion recognition performance on both experiments. It was also found that including more specific speaker information besides gender into the feature vectors yielded better results.

Our work is relevant in the sense that it does very minimal preprocessing on the raw data fed to the network. Plus, instead of merely relying on the participation of actors, the transferred learning related to speaker recognition comes in the form of feature matrices generated directly by a large-scale model trained with utterances from hundreds of persons with different ethnicities, accents, professions and ages.

3 Methodology

In this section we provide an outline of the speech corpora used. Following that, detail is given on the applied *VGGVox* model, and on how feature matrices were extracted from it when fed the data from the emotional speech databases.

3.1 Emotion speech databases

For this work, a set of 6 emotional speech databases was gathered, totaling over 9000 utterances of varying duration in 8 different languages, and portraying 9 different emotional states, to be applied in a speaker recognition model for feature extraction. The set of clips from the databases was reduced to only include clips corresponding to anger, disgust, fear, happiness, sadness, surprise and the neutral state, which were common to all databases. The list of databases is shown in Table 1.

Table 1. Catalog of emotional Speech Databases. Label code: A=Anger, H=Happiness, Sd=Sadness, F=Fear, D=Disgust, Sr=Surprise, B=Boredom, C=Calmness, N=Neutral

Database	Languages	Number of Utterances	Emotional States	Access
EMODB [11]	De	535	A, H, Sd, F, D, B, N	Public
EMOVO [12]	It	588	A, H, Sd, F, D, Sr, N	Public
SAVEE [13]	En	480	A, H, Sd, F, D, Sr, N	Public
RAVDESS [14]	En	1440	A, H, Sd, F, D, Sr, C, N	Public
RML [15]	En, Man, Ur, Pa, Fa, It	720	A, H, Sd, F, D, Sr	Private
S0329 [16]	Es	6041	A, H, Sd, F, D, Sr, N	Private

All files were converted to the WAV format, at a sampling rate of 16 kHz, as this value has been proved to be more than enough to capture all information embedded in a speech signal. In accordance with the *VGGVox* model's implementation, and in order to take full advantage of all the provided audio, files were adapted to be 1 to 10 seconds in length as well. Therefore, a small number of clips below the 1 second mark were disregarded, as these would hardly provide any emotional information, and clips above the 10 second mark were divided into equally long audio segments.

3.2 The VGGVox model

This model developed by Nagrani et al. [10] and based on a VGG-M architecture and composed of 12 layers, is fed raw data which undergoes minimal processing. With that in mind, narrowband magnitude spectrograms are generated using a sliding hamming window of width 25ms and step 10ms, meaning an n-second input will provide a 100n frames spectrum. Normalization is also performed on mean and variance, at every frequency bin of the spectrum, as it was observed that such a step produced an increase of 10% in classification accuracy. Yet, no other operations are performed on the input data, and the CNN is fed essentially raw spectrograms.

Variable length inputs are also efficiently dealt with by varying the support filter dimension of the apool6 layer. As such, the implementation is adaptable to an audio clip's duration, provided it is between 1 and 10 seconds in length, according to Table 2. The dimension values are conforming with the stride and padding methods used by the model, for each duration value. It should be noted that the model does handle clips longer than 10 seconds, by considering only the central 10-second segment of the clip, in spite of losing all the other potentially relevant surrounding information.

In terms of purpose, the model was directed towards speaker classification, and trained using the VoxCeleb1 dataset [10] also developed by Nagrani and her team. This dataset is of large scale, including over 100,000 utterances by 7000+ speakers of varied backgrounds, resulting in more than 2000 hours of audio. Consequently, the model is an ideal candidate for capturing copious amounts of speaker specific cues and prosody mannerisms from any type of human speech, emotional included. Training iterations also included batch normalization [17] and used the default hyper parameter values of the used MatConvNet toolbox [18].

Table 2. Average Pooling layer's k-th dimension adaptation to clip's n-second Duration

Frames	100	200	300	400	500	600	700	800	900	1000
Dimension	2	5	8	11	14	17	20	23	27	30

3.3 Feature extraction

Feature arrays were obtained from the output of the *apool6* layer of the *VGGVox* model, corresponding to its bottleneck. This was done considering an ideal middle point of speaker adaptation, meaning the extracted features would not suffer from either under-specialization or over-specialization issues. A simple application of the model to the audio clips without any form of processing other than the already specified was performed in order to obtain these feature arrays, which given their origin, had the dimension of 1x1x4096.

4 Experimental Results

Several experiments were carried out in order to evaluate the robustness and efficacy of the extracted feature arrays in terms of emotion recognition. The Weka software

[19] was employed so as to apply the feature arrays on the following state of the art classifiers: Naive Bayes [20], kNN [21], Random Forest [22], Logistic Model Tree (LMT) [23] and Support Vector Machine (SVM) [24]. A neural network-based approach was not followed during the classification stage given the fact that the amount of data available was not enough to credibly train a machine learning model. In this section, we provide more detail on the carried-out experiments and the obtained results, as well as a discussion and comparison of these to other state-of-the-art techniques.

4.1 Classifier performance

The Naive Bayes classifier was used as a mere baseline for evaluation against the rest of classifiers in the Weka software, when fed the provided feature arrays for emotion recognition. Performance results in terms of accuracy were obtained using 5-fold cross validation, on each database individually. Furthermore, the k-statistic [25] was also calculated to further support the validity of the obtained results against random chance, in parallel with unweighted average recall (UAR), a favored metric in emotion recognition systems which attributes the same significance to all possible classes [26]. All these results are shown in Table 3, with the highlighted cells corresponding to the best performance results used for comparison later on.

Table 3. State-of-the-art classifier performance on standalone emotional database 1x1x4096 feature arrays. Code: A=classifier accuracy (percentage), B=k-statistic and C=UAR.

	Naive Bayes			**k-Nearest Neighbors**			**Random Forest**			**Logistic Model Tree**			**Support Vector Machine**		
	A	***B***	***C***	***A***	***B***	***C***	***A***	***B***	***C***	***A***	***B***	***C***	***A***	***B***	***C***
EMODB	72.9	0.67	0.73	74.9	0.69	0.73	76.0	0.70	0.72	80.4	0.76	0.80	74.9	0.68	0.72
EMOVO	52.8	0.45	0.53	66.2	0.61	0.66	65.3	0.60	0.65	68.5	0.63	0.69	57.9	0.51	0.58
RAVDESS	51.8	0.44	0.53	65.5	0.60	0.65	61.9	0.55	0.60	71.6	0.67	0.71	55.0	0.47	0.60
RML	70.4	0.65	0.70	73.5	0.68	0.73	75.3	0.70	0.75	79.9	0.76	0.80	71.3	0.66	0.71
S0329	74.7	0.70	0.74	86.2	0.83	0.84	87.4	0.85	0.85	92.4	0.91	0.91	91.0	0.89	0.90
SAVEE	61.7	0.54	0.58	65.2	0.59	0.61	64.8	0.57	0.60	70.4	0.65	0.68	55.6	0.45	0.49

4.2 Discussion

Results are clearly varying from database to database. This suggests that, even though database size must also be taken into consideration, emotional prosody is affected differently in each population due to language and cultural diversity. As such, adaptation to cultural background is likely an additional approach worthwhile researching in order to improve SER systems.

Altogether, the obtained results always surpassed the proposed baseline having LMT given the best results (see highlighted column). As such, this results column was used for comparison against other state-of-the-art techniques, whose performances are

shown in Table 4. Here it is possible to verify that our technique did almost always produce better results on the databases. As such, the efficacy of our proposed method for speech emotion recognition is affirmed, having surpassed other state-of-the-art techniques. Finally, our observations certainly support the existence of relevant emotional information in speaker specific speech features. As such, speaker adaptation should be performed in systems aiming for successful SER.

Table 4. Comparative results between our proposed method and other state-of-the-art techniques on the same standalone databases.

	Kerkeni et al. [27]	**Jannat et al. [28]**	**Latif et al. [29]**	**Avots et al. [30]**	**Sidorov et al. [10]**	**Proposed Method**
EMODB	69.6 %	-	72.4 %	-	74.6 %	80.4 %
EMOVO	-	-	76.2 %	-	-	68.5 %
SAVEE	-	-	56.8 %	77.4 %	63.8 %	70.4 %
RAVDESS	-	66.4 %	-	-	-	71.6 %
RML	-	-	-	69.3 %	-	79.9 %
S0329	90.1 %	-	-	-	-	92.4 %

5 Conclusion

In this paper, we examined the robustness of speech features extracted using a large-scale speaker recognition model, for emotion recognition. We determined that, regardless of language, there is valuable emotional information embedded within speaker specific features. Acceptable but varying performance ratios were obtained on standalone databases of different languages. This suggests varying degrees of emotional prosody mannerism for different cultural backgrounds. Finally, and based on a general observation of the results, we can conclude that an initial step of speaker adaptation is of paramount importance and should be performed in any SER system, in order to achieve higher accuracy rates.

In the future, we intend to assess the efficacy of dimension reduction techniques such as PCA or LDA, and delve deeper into adaptable emotion recognition, by considering additional speaker information, such as cultural background, and incorporating facial expression analysis into a multi-modal emotion recognition system.

6 Acknowledgement

The authors would like to thank the respective database curators for providing access to their emotional speech datasets. This work has been partially supported by OE - national funds of FCT/MCTES (PIDDAC) under project UID/EEA/00048/2019.

7 References

[1] R. Jürgens et al. "Effect of Acting Experience on Emotion Expression and Recognition in Voice: Non-Actors Provide Better Stimuli than Expected" Journal of nonverbal behavior vol. 39, 3 (2015): 195-214. https://doi.org/10.1007/s10919-015-0209-5

[2] Paulmann S, Furnes D, Bøkenes AM, Cozzolino PJ. "How Psychological Stress Affects Emotional Prosody", (2016). Plos One 11(11): e0165022. https://doi.org/10.1371/journal.pone.0165022

[3] M. Spada, A. Nikcevic, G. Moneta and A. Wells. "Metacognition, perceived stress, and negative emotion". Science Direct. Personality and Individual Differences 44. (2008) 1172–1181. https://doi.org/10.1016/j.paid.2007.11.010

[4] J. Deng, X. Xu, Z. Zhang, S. Frühholz, B. Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition", IEEE Signal Processing Letters 2017. https://doi.org/10.1109/LSP.2017.2672753

[5] J. Deng, Z. Zhang, F. Eyben and B. Schuller, "Autoencoder-based Unsupervised Domain Adaptation for Speech Emotion Recognition," in IEEE Signal Processing Letters, vol. 21, no. 9, pp. 1068-1072, Sept. 2014. https://doi.org/10.1109/LSP.2014.2324759

[6] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016. https://doi.org/10.1109/ICASSP.2016.7472669

[7] W. Lim, D. Jang, and T. Lee. Speech emotion recognition using convolutional and recurrent neural networks. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016. https://doi.org/10.1109/APSIPA.2016.7820699

[8] A. Nagrani, J. S. Chung and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset", INTERSPEECH, 2017. https://doi.org/10.21437/Interspeech.2017-950

[9] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis and E. M. Provost, "Progressive Neural Networks for Transfer Learning in Emotion Recognition", INTERSPEECH, (2017) https://doi.org/10.21437/Interspeech.2017-1637

[10] Sidorov, S. Ultes and A. Schmitt. "Emotions Are A Personal Thing: Towards Speaker-Adaptive Emotion Recognition", ICASSP, (2014). https://doi.org/10.1109/ICASSP.2014.6854514 https://doi.org/10.1109/ICASSP.2014.6854514

[11] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss. "A database of german emotional speech". INTERSPEECH, 2005.

[12] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco. "Emovo corpus: an Italian emotional speech database". LREC (2014).

[13] S. Haq, P.J.B. Jackson, and J.D. Edge. "Audio-Visual Feature Selection and Reduction for Emotion Classification". In Proc. Int'l Conf. on Auditory-Visual Speech Processing, pages 185-190, 2008.

[14] S. R. Livingstone and F. 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, 2018. https://doi.org/10.1371/journal.pone.0196391

[15] Z. Xie and L. Guan. "Multimodal information fusion of audiovisual emotion recognition using novel information theoretic tools". IEEE International Conference on Multimedia and Expo (ICME), 2013. https://doi.org/10.4018/ijmdem.2013100101

[16] European Language Resources Association (ELRA). Database Elra-S0329.

[17] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv preprint arXiv: 1502.03167, 2015.

[18] A. Vedaldi and K. Lenc, "Matconvnet – convolutional neural networks for MATLAB," CoRR, vol. abs/1412.4564, 2014. https://doi.org/10.1145/2733373.2807412
[19] E. Frank, M. A. Hall and I. H. Witten, The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.
[20] G. H. John, P. Langley. "Estimating Continuous Distributions in Bayesian Classifiers.", In: Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, 338-345, 1995.
[21] D. Aha, D. Kibler."Instance-based learning algorithms.",Machine Learning (1991). 6:37-66. https://doi.org/10.1007/BF00153759
[22] L. Breiman. "Random Forests." Machine Learning (2001). 45(1):5-32. https://doi.org/10.1023/A:1010933404324
[23] N. Landwehr, M. Hall, E. Frank. "Logistic Model Trees." Machine Learning (2005). 95(1-2):161-205. https://doi.org/10.1007/s10994-005-0466-3
[24] CC Chang, and CJ Lin. "LIBSVM: A library for support vector machines". ACM Transactions on Intelligent Systems and Technology, 2011, Vol. 2(3), pp 27:1–27:27. https://doi.org/10.1145/1961189.1961199
[25] J. Cohen, "A coefficient of agreement for nominal scales". Educational and Psychological Measurement (1960). 20 (1): 37–46. https://doi.org/10.1177/001316446002000104
[26] B. W. Schuller, S. Steidl, A. Batliner et al., "The interspeech 2009 emotion challenge." in Interspeech, vol. 2009, 2009, pp. 312–315
[27] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. Ali Mahjoub and C. Cleder. "Automatic Speech Emotion Recognition Using Machine Learning". Social Media and Machine Learning. (2019) https://doi.org/10.5772/intechopen.84856
[28] Jannat, Sk Rahatul & Tynes, Iyonna & La Lime, Lott & Adorno, Juan & Canavan, Shaun. (2018). Ubiquitous Emotion Recognition Using Audio and Video Data. 956-959. https://doi.org/10.1145/3267305.3267689
[29] S. Latif, R. Rana, S.Younis, J.Qadir, Julien Epps."Transfer Learning for Improving Speech Emotion Classification Accuracy". arXiv:1801.06353. (2018). https://doi.org/10.21437/Interspeech.2018-1625
[30] Avots, E., Sapiński, T., Bachmann, M. et al. "Audiovisual emotion recognition in wild". Machine Vision and Applications (2018). https://doi.org/10.1007/s00138-018-0960-9

8 Authors

Gustavo Assunção is a PhD student at the University of Coimbra, and a researcher of the Institute of Systems and Robotics in Coimbra, Portugal. gustavo.assuncao@isr.uc.pt

Paulo Menezes is a tenured professor at the University of Coimbra and senior researcher of the Institute of Systems and Robotics in Coimbra, Portugal.

Fernando Perdigão is a Professor at the University of Coimbra and a senior researcher of the Instituto de Telecomunicações, in Portugal.

Article submitted 2019-10-15. Resubmitted 2019-12-15. Final acceptance 2019-12-17. Final version published as submitted by the authors.