ASR-based Features for Emotion Recognition: A Transfer Learning Approach

Noé Tits, Kevin El Haddad, Thierry Dutoit
Numediart Institute,
University of Mons, Belgium
{noe.tits, kevin.elhaddad, thierry.dutoit}@umons.ac.be

Abstract

During the last decade, the applications of signal processing have drastically improved with deep learning. However areas of affecting computing such as emotional speech synthesis or emotion recognition from spoken language remains challenging. In this paper, we investigate the use of a neural Automatic Speech Recognition (ASR) as a feature extractor for emotion recognition. We show that these features outperform the eGeMAPS feature set to predict the valence and arousal emotional dimensions, which means that the audio-to-text mapping learned by the ASR system contains information related to the emotional dimensions in spontaneous speech. We also examine the relationship between first layers (closer to speech) and last layers (closer to text) of the ASR and valence/arousal.

1 Introduction

With the advent of deep learning, areas of signal processing have been drasctically improved. In the field of speech synthesis, Wavenet (Van Den Oord et al., 2016), a deep neural network for generating raw audio waveforms, outperforms all previous approaches in terms of naturalness. One of the remaining challenges in speech synthesis is to control its emotional dimension (happiness, sadness, amusement, etc.). The work described here is part of a larger project to control as accurately as possible, the emotional state of a sentence being synthesized. For this, we present here exploratory work regarding the analysis of the relationship between the emotional states and the modalities used to express them in speech.

Indeed one of the main problems to develop such a system is the amount of good quality data (naturalistic emotional speech of synthesis quality, i.e. containing no noise of any sorts). This is why we are considering solutions such as synthesis by analysis and transfer learning (Pan and Yang, 2010).

Arousal and valence (Russell, 1980) are among the most, if not the most used dimensions for quantizing emotions. Valence represents the positivity of the emotion whereas arousal represents its activation. Since they represent emotional states, these dimensions are linked to several modalities that we use to express emotions (audio, text, facial expressions, etc.).

It has recently been shown that for emotion recognition, deep learning based systems learn features that outperform handcrafted features (Trigeorgis et al., 2016) (Martinez et al., 2013) (Kim et al., 2017a,b). The use of context and different modalities has also been studied with deep learning models. Poria et al. (2017) focus on the contextual information among utterances in a video while Zadeh et al. (2017, 2018) develop specific architectures to fuse information coming from different modalities.

In this work, with the goal to study the relationship between valence/arousal, and different modalities, we propose to use the internal representation of a speech-to-text system. An Automatic Speech Recognition (ASR) system or speech-to-text system, learns a mapping between two modalities: an audio speech signal and its corresponding transcription. We hypothesize that such a system must also be learning representations of emotional expressions since these are contained intrinsically in both speech (variation or the pitch, the energy, etc.) and text (semantic of the words).

In fact, we show here that the activations of certain neurons in an ASR system, are useful to esti-

Proceedings of the First Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML), pages 48–52,
Melbourne, Australia July 20, 2018. ©2018 Association for Computational Linguistics

mate the arousal and valence dimensions of an audio speech signal. In other words, transfer learning is leveraged by using features learned for an automatic speech recognition (ASR) task to estimate valence and arousal. The advantage of our method is that it allows combining the use of large datasets of speech with transcriptions with limited datasets annotated in emotional dimensions.

An example of transfer learning is the work of Radford et al. (2017). They trained a multiplicative LTSM (Krause et al., 2016) to predict next character based on the previous ones to design a text generator system. The dataset used to train their model was the Amazon review dataset presented in McAuley et al. (2015). Then, they used the representation learned by the model to predict sentiment also available in the dataset, and achieved state of the art prediction.

In this paper, we show that the activations of a deep learning-based ASR system trained on a large database can be used as features for the estimation of arousal and valence values. The features would therefore be extracted from both the audio and text modalities which the ASR system learned to map.

2 ASR-based Features for Emotion Prediction Via Regression

Our goal is to study the relationship between valence/arousal, and audio/text modalities thanks to an ASR system. The main idea is that the ASR system that models the mapping between audio and text might learn a representation of emotional expression. So, for our analyses, we use an ASR system as a feature extractor which feeds a linear regression algorithm to estimate the arousal/valence values. This section describes the whole system. First we present the ASR system used as a feature extractor. We then briefly present the data used and present first results on the data analysis.

2.1 ASR system

The ASR system used is implemented in (Namju and Kyubyong, 2016) and pre-trained on the VCTK dataset (Veaux et al., 2017) containing 44 hours of speech uttered by 109 native speakers of English.

Its architecture consists of a dilated convolution of blocks. Each block is a gated constitutional unit (GCU) with a skip (residual) connection. In other words a Wavenet-like architecture (Van Den Oord et al., 2016). There are 15 layers and 128 GCUs in each layer: 1920 GCUs in total.

To lighten the computational cost, the audio signal is compressed in 20 Mel-Frequency Cepstral Coefficients (MFCCs) and then fed into the system.

2.2 Dataset Used

IEMOCAP Dataset

The ”interactive emotional dyadic motion capture database” (IEMOCAP) dataset (Busso et al., 2008) is used in this paper. It consists of audio-visual recordings of 5 sessions of dialogues between male and female subjects. In total it contains 10 speakers and a total of 12 hours of data. The data is segmented in utterances. Each utterance is transcribed and annotated by category of emotions (Ekman, 1992) and a value for emotional dimensions (Russell, 1980) (valence, arousal and dominance) between 1 and 5 representing the dimension's intensity.

In this work, we only use the audio and text modalities as well as the valence and arousal annotations.

Data Analysis and Neural Features

We investigate the relationship between the activation output of the ASR-based system's GCUs and the valence/arousal values by studying the correlations between them. For every utterance and for each speaker of the IEMOCAP dataset, we compute the mean activation of the GCUs of the ASR. The Pearson correlation coefficient is then calculated between the mean activation outputs and the values of valence/arousal of all utterances of the speaker. In the rest of the paper, we will refer to the mean activation of the GCUs as neural features. As an example, the results concerning the female speaker of session 2 is summarized in a heat map represented in Figure 1

Each row of the heat map corresponds to a layer of GCUs. The color is mapped with the Pearson correlation coefficient value.

One can see that correlations exist for both arousal and valence. This suggests that the ASR-based system learns a certain representation of the emotional dimensions.

2.3 Structure of the system

The system is illustrated in Figure 2. As previously mentioned, the ASR system is used as a fea-

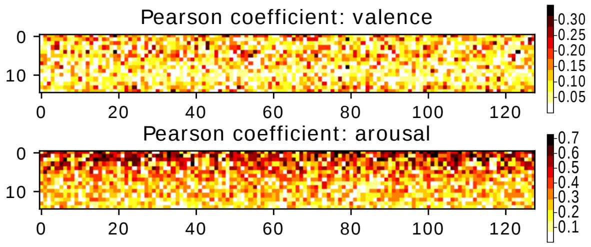

Figure 1: Pearson correlation coefficient between the neural features and valence (up) and arousal (down) - Female speaker of session 2

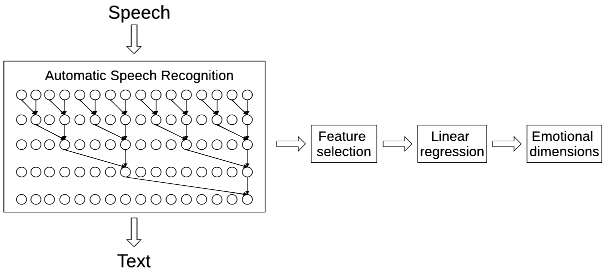

Figure 2: Block diagram of the system

ture extractor. First we compute the 20 MFCCs of the utterances of the IEMOCAP dataset with librosa python library (McFee et al., 2015). These are passed through the ASR to compute the corresponding neural features.

A feature selection is applied on the neural features to keep 100 among the 1920 features for dimensionality reduction purpose. The selection is done using the scikit-learn python library (Pedregosa et al., 2011) with the Fisher score.

Finally a linear regression is trained to estimate the valence/arousal values from the neural features using the IEMOCAP data. The linear regression is done using scikit-learn. The training is done by minimizing the Mean Squared Error (MSE) between predictions and labels.

3 Experiments and Results

In this section, we detail the experiments that we carried out. The first one is the evaluation of the neural features in terms of MSE and its comparison with a linear regression of the eGeMAPS feature set (Eyben et al., 2016). In the second one, we investigate the relationship between the audio and text and modalities and the emotional dimensions.

3.1 First experiment: Linear regression

In this first experiment, we investigate the performance of a linear regression to predict arousal and valence using the neural features. We compare this with a linear regression using the eGeMAPS feature set.

The eGeMAPS feature set is a selection of acoustic features that provide a common baseline for evaluation in researches to avoid differences of feature set and implementations. Indeed, they also provide their implementation with openSMILE toolkit (Eyben et al., 2010) that we used in this work.

The features were selected based on their ability to represent affective physiological nuances in voice production, their proven performance in former research work as well as the possibility to extract them automatically, and their theoretical significance.

The result of this selection is a set of 18 Low-level descriptors (LLDs) related to frequency (pitch, formants etc.), energy (loudness, Harmonics-to-Noise Ratio, etc.) and spectral balance (spectral slopes, ratios between formant energies, etc.). Then several functionals such as standard deviation and mean are applied to these LLDs to have the final features.

The results obtained from the linear regression in terms of MSE are compared to the annotations for each of the arousal and valence values (between 1 and 5) in Table 1.

	Arousal		Valence	
	Mean	Variance	Mean	Variance
Neural features	**0.259**	0.020	**0.660**	0.118
eGeMAPS set	0.267	0.034	0.697	0.135

Table 1: MSE on the prediction of valence and arousal.

We perform a leave-one-speaker-out evaluation scheme with both feature sets for cross-validation. In other words, each validation set in constituted with the utterances corresponding to one speaker and the corresponding training set with the other speakers. We train a model with each training set and evaluate it on the validation set in terms of MSE. The table contains the mean and standard deviation of the MSEs.

It is clear from this table that the neural features outperform the eGeMAPS in this experiment. This confirms the fact that the ASR system learns representations of emotional dimensions in spontaneous speech.

3.2 Second experiment: Influence of modalities

During the data exploration, we noticed that, for some speakers, the layers closer to the speech input were more correlated to arousal and the ones closer to the text output to valence. An example is shown in Figure 3. We present, in this section, preliminary studies regarding this matter.

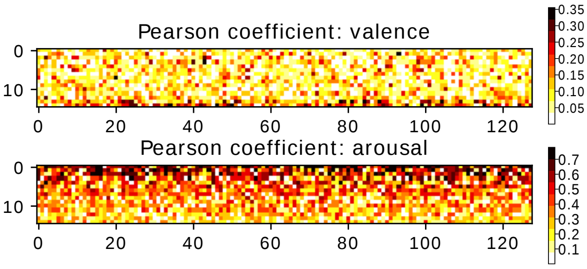

Figure 3: Pearson correlation coefficient between the neural features and valence (up) and arousal (down) - Female speaker of session 1

In order to analyze this phenomenon as precisely as possible, we only considered the utterances from the IEMOCAP database for which the valence/arousal annotators were consistent with each other, leaving us with 7532 utterances in total instead of 10039.

Then we performed linear regression with 4 different sets of feature to study their influence. For the first set, we select the 100 best features among the 3 first layers of the neural ASR in terms of Fisher score using scikit-learn. For the second set, we apply the same selection to the 3 last layers. The third set selection is applied among all neural features. The last set is the eGeMAPS feature set.

The results are summarized in Figure 2. As expected, the results show, that for the speakers considered, the layers closer to the audio modality outperform the ones closer to the text modality in the ASR architecture for arousal prediction and vice versa for the valence prediction. On this we build a hypothesis that the arousal-related features learned are more related to the audio modality than the text and vice versa for the valence-related features. This hypothesis will be further explored in future work.

4 Conclusions and Future work

In this paper, we show that features learned by a deep learning-based system trained for the Automatic Speech Recognition task can be used for emotion recognition and outperform the eGeMAPS feature set, the state of the art hand-crafted features for emotion recognition. Then we investigate the correlation of the emotional dimensions arousal and valence with the modalities of audio and text of the speech. We show that for some speakers, arousal is more correlated to neural features extracted from layers closer to the speech modality and valence to the ones closer to the text modality.

	Arousal		Valence	
	Mean	Variance	Mean	Variance
First layers	0.325	0.069	0.714	0.114
Last layers	0.357	0.038	0.661	0.089
All	0.296	0.044	0.621	0.099
eGeMAPS set	0.328	0.064	0.683	0.124

Table 2: Means and variances of the MSE on the prediction of valence and arousal.

To improve the system, we plan to perform an end-to-end training including the average operation. Another avenue to explore is to replace the average over time by a max-pooling over time which according to Aldeneh and Provost (2017) select the frames that are emotionally salient.

Then an analysis of the underlying activation evolutions could be done to see if it is possible to extract a frame-by-frame description of valence and arousal without having to annotate a database frame-by-frame.

Concerning the second experiment, we intend to investigate why these correlation patterns are only visible for some speakers and not others and the relationship between the arousal/valence and audio/text. We thereby hope to better understand the way multidimensional representations of emotions can be used to control the expressiveness in synthesized speech.

Acknowledgments

Noé Tits is funded through a PhD grant from the Fonds pour la Formation à la Recherche dans l'Industrie et l'Agriculture (FRIA), Belgium.

References

Zakaria Aldeneh and Emily Mower Provost. 2017. Using regional saliency for speech emotion recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 2741–2745. IEEE.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2016. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462. ACM.

Jaebok Kim, Gwenn Englebienne, Khiet P. Truong, and Vanessa Evers. 2017a. Deep temporal models using identity skip-connections for speech emotion recognition. In *Proceedings of the 2017 ACM on Multimedia Conference*, MM '17, pages 1006–1013, New York, NY, USA. ACM.

Jaebok Kim, Gwenn Englebienne, Khiet P. Truong, and Vanessa Evers. 2017b. Towards speech emotion recognition "in the wild" using aggregated corpora and deep multi-task learning. In *INTERSPEECH*.

Ben Krause, Liang Lu, Iain Murray, and Steve Renals. 2016. Multiplicative lstm for sequence modelling. *arXiv preprint arXiv:1609.07959*.

Hector P Martinez, Yoshua Bengio, and Georgios N Yannakakis. 2013. Learning deep physiological models of affect. *IEEE Computational Intelligence Magazine*, 8(2):20–33.

Julian McAuley, Rahul Pandey, and Jure Leskovec. 2015. Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, pages 18–25.

Kim Namju and Park Kyubyong. 2016. Speech-to-text-wavenet. https://github.com/buriburisuri/speech-to-text-wavenet.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 873–883.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

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

Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. 2017. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. *arXiv preprint arXiv:1802.00927*.