Investigation of Bottle-Neck Features for Emotion Recognition

Anna Popková[1(✉)], Filip Povolný[2], Pavel Matějka[1,2], Ondřej Glembek[1], František Grézl[1], and Jan "Honza" Černocký[1]

[1] Speech@FIT Group, Brno University of Technology, Brno, Czech Republic
xpopko00@stud.fit.vutbr.cz
[2] Phonexia s.r.o., Brno, Czech Republic

Abstract. This paper describes several systems for emotion recognition developed for the AV+EC 2015 Emotion Recognition Challenge. A complete system, making use of all three modalities (audio, video, and physiological data), was submitted to the evaluation. The focus of our work was, however, on the so called *Bottle-Neck* features used to complement the audio features. For the recognition of arousal, we improved the results of the delivered audio features and combined them favorably with the Bottle-Neck features. For valence, the best results were obtained with video, but a two-output Bottle-Neck structure is not far behind, which is especially appealing for applications where only audio is available.

Keywords: Emotion recognition · Bottle-Neck features · Context · Fusion

1 Introduction

The Speech@FIT group at Brno University of Technology and Phonexia are active and have been successful in multiple aspects of speech data mining. Recently, mainly with the EC-sponsored projects BISON[1] and MixedEmotions[2]

This work has been funded by the European Union's Horizon 2020 programme under grant agreement No. 644632 MixedEmotions and No. 645523 BISON, and by Technology Agency of the Czech Republic project No. TA04011311 "MINT". It was also supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0013. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government. Thanks to Fabien Ringeval for scoring several other systems after the deadline of AVEC 2015 which allowed us to make proper analysis for this paper.

[1] http://bison-project.eu/.
[2] http://mixedemotions-project.eu/.

P. Sojka et al. (Eds.): TSD 2016, LNAI 9924, pp. 426–434, 2016.
DOI: 10.1007/978-3-319-45510-5_49

and with an interest both in academia and industry, emotion recognition has become increasingly important.

This paper presents our systems based on the material provided by Audio–Visual+Emotion Recognition Challenge (AV+EC 2015)[3] [1]. AVEC is an annual challenge held since 2011. Its main purpose is emotion recognition from multimodal data—audio, video, and newly also physiological data. Emotion is understood here as a two-point continuous values on 2D plane according to arousal–valence model [11].

The data comes with three sets of features for audio, video and physiological signals. While the latter two were used as-is (the work concentrated on their post-processing, regressor training and fusion), in audio, we have complemented the provided material by Bottle-Neck (BN) features generated from a narrow hidden layer of a neural network trained toward phonetic targets. BN features were designed for automatic speech recognition [3] and have been included into most top-performing ASR systems including their multi-lingual variants [3]. Recently, BN features (and more general feature extraction schemes based on DNNs) were found very effective in other areas of speech processing, such as language recognition [4,5] and speaker identification [6,10]. Due to their ability to suppress nuisance variability in the speech data, we consider them a promising candidate also for emotion recognition, especially for the AV+EC challenge where very limited amount of labeled data (only 27 speakers) is available.

The rest of the paper provides a description of experiments leading to our submission for the AV+EC challenge and concentrates on BN features used for the audio modality.

2 Provided Material

2.1 Data

The data-set comes from the RECOLA multimodal database [2]. It contains spontaneous interactions in French. Participants were recorded in dyads during a video conference while resolving a collaborative task (winter survival task). Data was collected from 46 participants, but due to consent issues, only 5.5 hours of fully multimodal recordings from 27 participants are usable. The database is gender balanced and the mother tongues of speakers are French, Italian and German. The first 5 min of each recording were annotated by 6 French-speaking emotion annotators in the continuous arousal–valence space, leading to 135 min of data with the emotion ground truth. These recordings are divided into training, development and test sets, where annotations are provided only for training and development ones. The database is freely available[4] and full details are provided in [1,2].

[3] http://sspnet.eu/avec2015/.
[4] https://diuf.unifr.ch/diva/recola/.

2.2 AV+EC Features

Five sets of features were provided by the organizers (please refer to the challenge summary paper [1] for full description and references):

- **Audio** 102-dimensional feature set is extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS). These features are generated from short fixed length segments (3 s) shifted by 40 ms.
- **Video** features include two types of facial descriptors: appearance and geometry based. The former were extracted by Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) leading to total vector size of 84, the latter are facial landmarks leading to vector size of 316. Again, overlapping 3 s segments with 40 ms shift were used. The problem with video features was that for parts of the data, the face was not recognized and no information was provided. For certain recordings, the amounts of unrecognized frames were up to 40 %.
- **Physiological** sets include Electrocardiogram (ECG, 54 parameters) derived features, based on heart rate, its measure of variability, and derived parameters and statistics, and Electrodermal activity (EDA, 60 parameters) including skin conductance response (SCR), skin conductance level (SCL), as well as a number of derived parameters.

2.3 Evaluation and Baselines

The results were evaluated using the concordance correlation coefficient (CCC) to measure the correlation between the prediction and the reference. CCC combines the Pearson correlation coefficient of two time series ρ with mean square error:

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2\sigma_y^2 + (\mu_x - \mu_x)^2}. \quad (1)$$

CCC produces values from -1 to 1. The value of 1 means that the two variables are identical, -1 means that they are opposite, and 0 means that they are totally uncorrelated.

The organizers experimented with several emotion recognition schemes and provided the best obtained values in [1]. These serve as baselines for our work and are mentioned in the tables.

3 Bottle-Neck Features

We used Stacked Bottle-Neck (SBN) features as our additional feature set. The architecture for this kind of feature extraction consists of two NNs trained towards phonetic targets. The output of the first network is stacked in time, defining context-dependent input features for the second NN, hence the term Stacked Bottleneck Features [4].

The NN input features are filter-bank energies concatenated with fundamental frequency (F0) features produced by four different estimators: BUT F0

detector produces 2 coefficients (F0 and probability of voicing), Snack F0 gives a single F0, and Kaldi F0 estimator outputs 3 coefficients (Normalized F0 across sliding window, probability of voicing and F0 delta). Fundamental frequency variation (FFV) estimator [7] produces a 7-dimensional vector. Therefore, the whole feature vector has $24 + 2 + 1 + 3 + 7 = 37$ coefficients [8].

The conversation-side based mean subtraction is applied on the whole feature vector. 11 frames of log filter bank outputs and fundamental frequency features are stacked together. Hamming window followed by DCT consisting of 0^{th} to 5^{th} base are applied on the time trajectory of each parameter resulting in $(24+13) \times 6 = 222$ coefficients on the first-stage NN input [5].

The first-stage NN has four hidden layers with 1500 units each except the BN layer. BN layer's size is 80 neurons and it is the third hidden layer. Its outputs are stacked over 21 frames and downsampled (every 5^{th} is taken) before they enter the second-stage NN, which has the same structure as the first-stage NN. The outputs from 80 neurons in BN layer are the final BN features for the recognition system [8].

For training the neural networks, the IARPA Babel Program data[5] were used. 11 languages were used to train the multilingual SBN feature extractor [3]: Cantonese, Pashto, Turkish, Tagalog, Vietnamese, Assamese, Bengali, Haitian, Lao, Tamil, Zulu. Details about the characteristics of the languages can be found in [9]. The training speech was force-aligned using our BABEL ASR system [8].

4 Systems and Experiments

The general scheme of our system is shown in Fig. 1. The following subsections deal with individual building blocks and results of experiments therewith in detail. The regressor producing arousal and/or valence values is linear, except for indicated cases, where a neural network is used.

First, a number of single systems was built: for each feature set (5 supplied ones + bottlenecks) and each dimension of emotion (arousal vs. valence), making up 12 systems in total. Each of them was investigated for optimum preprocessing, regressor training, and post-processing. All systems were trained on the training set and evaluated on the development set.

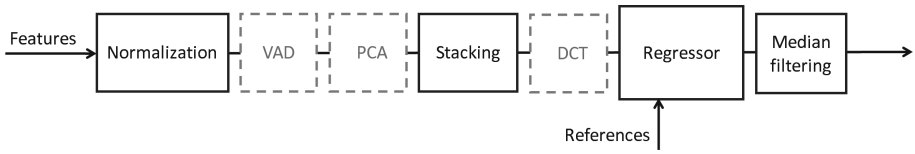

Fig. 1. Emotion recognition system scheme

[5] Collected by Appen, http://www.appenbutlerhill.com.

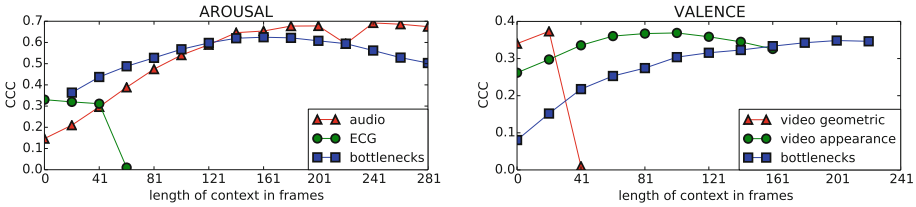

Fig. 2. Dependency of CCC on the number of stacked frames.

4.1 Pre-processing

The data-set from RECOLA includes 27 recordings from 27 different subjects. To prepare the data for the following regression, the whole set was globally mean- and variance-normalized.

We tried to use Voice Activity Detection (VAD) on the training data: frames with detected silence were dropped, and the system was trained only on the remaining speech frames. However, in video, there is always an indication of emotion even in case of silence, therefore, we tested on whole test recordings without silence removal. It is also necessary to note that the recordings are dialogues and the result of the emotion recognition of the observed person could be disturbed by speech of the second person, whose emotions we do not want to recognize. For these two reasons, the VAD does not help to improve our results, and was not used in our final systems.

Principal Component Analysis (PCA) was tested for dimensionality reduction of the feature sets, and had good results in experiments with supplied audio features for arousal. Reduction from the baseline dimensionality of 102 to 13 dimensions performed the best. On contrary, no or only very little reduction helps for both video features, which are used mainly for valence.

In our experiments, we train regression models for valence and arousal values for each frame (every 40 ms). In many other classification and recognition tasks, we have seen the need of adding larger temporal context to make a good prediction. The results with changing context size are shown in Fig. 2. The context of 141 stacked frames (70 to the left and 70 to the right of the current frame + current frame) was found optimal for arousal recognition from audio. Shorter context is necessary for valence while it is recognizing from video.

A further dimensionality reduction can follow the stacking of context frames. A standard technique is to project the temporal trajectories of features to the discrete cosine transform (DCT) bases. We observed, that for systems using bottleneck features, it is beneficial to perform DCT reduction to the first 7 coefficients for arousal, and the first 30 coefficients for valence. For recognition of valence from video appearance features, the first 3 DCT coefficients were found optimal on down-sampled feature trajectories (only every second frame was retained). For all DCT projections, Hamming windowing of the trajectories was applied first.

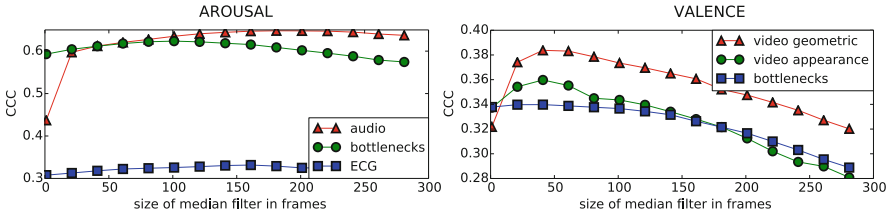

Fig. 3. Influence of the length of median filter applied after the regression.

4.2 Regression Model and Its Training

Linear regression is used on all single systems for arousal and all single systems for valence except for processing video geometric features, where neural network with one hidden layer is used (topology: 948–474–3). The regression can be trained in different ways, as six different labels from 6 different annotators plus another label as the gold standard (normalized and averaged as described in [1]) are available. After experimenting, we empirically found that for training arousal recognizer, data from annotators one and three give the best performance. Those annotations seem diverse and with bigger variation. For valence, we choose mostly annotators one, two and three, whose annotations seem more consistent. Our single systems produce estimates of two values for arousal, and three values for valence (trained to match the best annotators). A weighted average of these is then taken to produce one single value for arousal and one for valence. The weights were determined experimentally.

4.3 Post-processing

The outputs from our initial regression models seemed very noisy with respect to the reference. Median filter was used for smoothing and its optimum length was investigated, see Fig. 3. For arousal, a longer filter (over approx. 7 s) than for valence (over approx. 2 s) is needed.

4.4 Results and Comments

The final results of all investigated systems are summarized in Table 1 along with the baselines in brackets. It is evident that in most cases, our results are outperforming the baseline. They confirm our expectations—the recognition of arousal is better from audio, while for valence, video features perform better. Linear regression was used in all cases except for system processing video geometric features. We trained a neural network with one hidden layer with topology 945–474–3 in this case.

The major improvements are listed below. Using long context—in particular systems as long as 6–7 s, applying the median filter on the output and training on the data from the particular annotators instead of training on the Gold Standard.

Table 1. Comparison of single systems of different modalities, AV+EC 2015 baseline results are in brackets.

	Development		Test	
CCC	Arousal	Valence	Arousal	Valence
Audio	0.704 (0.287)	0.190 (0.069)	**0.595** (0.228)	0.160 (0.068)
Video geometric	0.054 (0.231)	0.403 (0.325)	0.151 (0.162)	**0.302** (0.292)
Video appearance	0.126 (0.103)	0.346 (0.273)	0.110 (0.114)	**0.334** (0.234)
ECG	0.305 (0.275)	0.231 (0.183)	-* (0.192)	-* (0.139)
EDA	0.117 (0.078)	0.235 (0.204)	0.118 (0.079)	0.226 (0.195)
Bottlenecks	0.625	0.344	**0.525**	0.176

*Numeric error prevented us from finishing this evaluation, unfortunately we have no longer access to the references of test data to re-evaluate them.

5 Bottle-Neck System Investigation

Inspired by the positive results of BN features on emotion recognition from audio, we created another system using only bottleneck features for simultaneous recognition of both arousal and valence (multi-task). In this system, long—up to 7 s (181 frames)—but downsampled context is used (only every 4^{th} frame was taken), then DCT is applied and 30 first bases are retained. This system is also based on a simple linear regression. As a post-processing, median filter over 183 frames (7 s) for arousal and over 145 frames (6 s) for valence is used. The results in Table 2 indicate, that multi-task training is more efficient than having two single-task systems especially for the arousal prediction.

Table 2. Comparison of single- and multi-task systems based on Bottle-Neck features.

	Development		Test	
CCC	Arousal	Valence	Arousal	Valence
Single-task*	0.625	0.344	0.525	0.176
Single-task**	0.390	0.343	0.296	0.174
Multi-task	0.699	0.376	0.596	0.293

*Parameters tuned for each modality,
**Identical parameters as in multitask.

6 Fusion

Tuning and optimization of all individual systems was performed, as described in Sect. 4. Because of different sets of features and also different dimensions of emotion, we ended up with systems different in the pre-processing, training and post-processing. For fusion, we chose the two best single systems on arousal and valence separately, the real outputs (not weighted averages) from those best single systems were inputs for the fusion:

Table 3. Parameters of the best single systems used for final fusion.

	Arousal		Valence	
	Audio	Bottlenecks	Video geometric	Video appearance
Annotators	1 + 3	1 + 3	1 + 2 + 3	1 + 2 + 3
PCA	From 102 to 13	Not reduced (80)	Not reduced (316)	From 84 to 70
Length of context	141 (5.6 s)	161 (6.4 s)	3 (0.1 s)	30 (1.2 s)
DCT coefficients	20	7	No reduction	3
Downsampling	-	-	-	2

Table 4. Fusion system, AV+EC 2015 baseline results are in brackets

CCC	Arousal	Valence
Development	0.772 (0.476)	0.518 (0.461)
Test	0.660 (0.444)	0.504 (0.382)

– For *arousal*, we fused systems using audio features and Bottle-Necks. The fusion was a linear regression.
– For *valence*, both video system outputs were used. The fusion was done on the score level with a neural network with one hidden layer (topology: 486–243–1).

The system parameters selected for the final fusion are listed in Tables 3 and 4 contains the final fusion results. In all cases, comparison to baselines is favorable.

7 Conclusions

While the whole AV+EC evaluation was of interest for us, our focus was on the audio, as this is the main modality we are working with—most of our work is done in cooperation with contact centers that have no access to video or physiological data. Arousal is well recognizable from audio feature sets provided with AV+EC 2015 baselines, and we have improved the results by working on the context, regressor training and post-processing. The AV+EC features also combine favorably with the newly introduced Bottle-Neck features. For valence, the best evaluation results were obtained with video features, but the two-output Bottle-Neck structure is not far behind. We have also confirmed that simple regressors (linear or NN) can be used for the emotion prediction task.

References

1. Ringeval, F., Schuller, B., Valstar, M., Jaiswal, S., Marchi, E., Lalanne, D., Cowie, R., Pantic, M.: Av+ec 2015: the first affect recognition challenge bridging across audio, video, and physiological data. In: Proceedings of AVEC 2015, Satellite Workshop of ACM-Multimedia 2015, Brisbane, Australia, October 2015

2. Ringevaland, F., Sonderegger, A., Sauer, J., Lalanne, D.: Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In: Proceedings of Face and Gestures, Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE) (2013)
3. Grézl, F., Egorova, E., Karafiát, M.: Further investigation into multilingual training and adaptation of stacked bottle-neck neural network structure. In: Proceedings of Spoken Language Technology Workshop, pp. 48–53 (2014)
4. Matějka, P., Zhang, L., Ng, T., Mallidi, H.S., Glembek, O., Ma, J., Zhang, B.: Neural network bottleneck features for language identification. In: Proceedings of Odyssey 2014, pp. 299–304 (2014)
5. Fér, R., Matějka, P., Grézl, F., Plchot, O., Černocký, J.: Multilingual bottleneck features for language recognition. In: Proceedings of Interspeech 2015, pp. 389–393 (2015)
6. Cumani, S., Laface, P., Kulsoom, F.: Speaker recognition by means of acoustic and phonetically informed GMMs. In: Proceedings of Interspeech 2015 (2015)
7. Heldner, M., Laskowski, K., Edlund, J.: The fundamental frequency variation spectrum. In: Proceedings of FONETIK (2008)
8. Karafiát, M., Veselý, K., Szoke, I., Burget, L., Grézl, F., Hannemann, M., Černocký, J.: BUT ASR system for BABEL surprise evaluation. In: 2014 IEEE Spoken Language Technology Workshop (SLT), NV, USA, December 2014
9. Harper, M.: The BABEL program and low resource speech technology. In: ASRU 2013, December 2013
10. Garcia-Romero, D., McCree, A.: Insights into deep neural networks for speaker recognition. In: Proceedings of Interspeech 2015 (2015)
11. Gunes, H., Schuller, B.: Categorical and dimensional affect analysis in continuous input: current trends and future directions. Image Vis. Comput. Affect Anal. Continuous Input **31**(2), 120–136 (2013)