Multimodal Emotion Recognition for AVEC 2016 Challenge

Filip Povolný
Phonexia
Brno, CZ
povolny@phonexia.com

Pavel Matějka*
Faculty of Information Technology
Brno University of Technology
matejkap@fit.vutbr.cz

Michal Hradiš
Faculty of Information Technology
Brno University of Technology
ihradis@fit.vutbr.cz

Anna Popková
Faculty of Information Technology
Brno University of Technology
xpopko00@stud.fit.vutbr.cz

Lubomír Otrusina
Faculty of Information Technology
Brno University of Technology
iotrusina@fit.vutbr.cz

Pavel Smrž
Faculty of Information Technology
Brno University of Technology
smrz@fit.vutbr.cz

ABSTRACT

This paper describes a systems for emotion recognition and its application on the dataset from the AV+EC 2016 Emotion Recognition Challenge. The realized system was produced and submitted to the AV+EC 2016 evaluation, making use of all three modalities (audio, video, and physiological data). Our work primarily focused on features derived from audio. The original audio features were complement with bottleneck features and also text-based emotion recognition which is based on transcribing audio by an automatic speech recognition system and applying resources such as word embedding models and sentiment lexicons. Our multimodal fusion reached CCC=0.855 on dev set for arousal and 0.713 for valence. CCC on test set is 0.719 and 0.596 for arousal and valence respectively.

CCS Concepts

•Social and professional topics → User characteristics;

Keywords

emotion recognition, valence, arousal, bottleneck features, neural networks, regression, speech transcription, word embedding

1. INTRODUCTION

This paper presents an emotion recognition system evaluated on the material defined within the Audio-Visual + Emotion Recognition Challenge (AV+EC 2016)[1] [25]. AV+EC is an annual challenge held since 2011. Its main purpose is emotion recognition from multimodal data — audio, video and physiological data. Emotion is understood as a value in two-dimensional arousal-valence continuous space [12].

The data comes with three sets of features for audio, video and physiological signals. Our main focus was on audio and video features. The work on physiological features was 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 [11] and have since been integrated into many top-performing ASR systems and their multilingual variants [11]. Recently, BN features (and general feature extraction schemes based on deep neural networks) were found very efficient in other areas of speech processing, such as language identification [19, 5] and speaker identification [4, 7]. Due to their ability to suppress nuisance variability in the speech data, we proved in AVEC2015 challenge [21] that these features are promising candidate also for emotion recognition.

In video, we have complemented the baseline features by activations of a convolutional neural network (CNN) trained to localize facial landmarks [26]. These activations encode geometrical information mixed with appearance information.

In addition we experimented with text based features, which we obtained from an automatic speech recognition system. We explored a lexicon-based approach as well as word embedding – a technique mapping words to vectors of real numbers in a space with lower dimension than the vocabulary size [2, 20].

The rest of this paper provides a description of experiments leading to our submission for the AV+EC 2016 challenge.

2. EMOTION FEATURES

2.1 Audio Features

Organizers provided a set of 102 dimensional audio features, known as Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS). The features were generated from short fixed length segments (3s) shifted by 40 ms [25].

In addition, we used two Stacked Bottle-Neck features as

*The author is further affiliated with Phonexia, Brno, CZ
[1]http://sspnet.eu/avec2016/

AVEC'16, October 16 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2988257.2988268

our main acoustic feature set trained as French only and second in Multilingual fashion (trained on several languages). We have seen very good results with this features in our AVEC 2015 submission [21].

The architecture for the feature extraction consists of two NNs. 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 (SBN) [19].

The NN input features are 24 log Mel Filter band 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 just a single F0 and Kaldi F0 estimator outputs 3 coefficients (Normalized F0 across a sliding window, probability of voicing and delta). Fundamental frequency variation (FFV) estimator [16] produces a 7-dimensional vector. Therefore, the whole feature vector has 24+2+1+3+7=37 coefficients [15].

The conversation-side mean subtraction is applied on the whole feature vector. 11 frames of log filter bank outputs and fundamental frequency features are stacked together. The Hamming window followed by the 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 [19].

The first-stage NN has four hidden layers with 1500 units each except the BN layer. The size of the BN layer is 80 neurons and it is the third hidden layer. Its outputs are stacked over 21 frames and down-sampled (every 5th is taken) and entered into the second-stage NN with the same structure as the first-stage NN. Outputs from 80 neurons in the BN layer form the final BN features for the recognition system [15].

We trained 2 systems with this topology, first only on French data (which match the data from the challenge) and second on 5 languages as multilingual bottleneck features.

For the training of the French recognition system, we used the 21 hours of transcribed data from BISON project[2] and 23 hours from EVALDA project[3]. Bottleneck features derived from this system are denoted as *BN-FR*.

To train the multilingual system, the IARPA Babel Program data[4] were used. We used 11 languages to train our multilingual SBN feature extractor: Cantonese, Pashto, Turkish, Tagalog, Vietnamese, Assamese, Bengali, Haitian, Lao, Tamil, Zulu. Details about the characteristics of the languages can be found in [13]. The training speech was force-aligned using our BABEL ASR system [15]. Bottleneck features derived from this system are denoted as *BN-Multi*.

2.2 Text Based Features

People's emotion can be perceived through different modalities, most acknowledged ones being hearing and vision. However, the semantic of the words used can also be an important aspect to take into consideration in emotion detection. The words chosen can say a lot on the current state of emotion of the person indeed.

Automatic speech recognition was applied to the audio data and several approaches to extract features from the resulting texts were attempted. These included word embedding, lexicon based sentiment detection in French and two standard English sentiment tools applied to automatic translations of the transcripts.

2.2.1 Automatic Speech Recognition

The French speech-to-text transcription system used to generate the automatic word hypotheses has the same basic structure as the American English one described in [8] except that an MLP is used to estimate the HMM state likelihood. The French system (developed in collaboration with Vocapia Research), first separates non-speech and speech portions of the audio file and then applies a maximum-likelihood segmentation/clustering process [10], to associate labels with segment clusters, where each cluster ideally represents one speaker.

The acoustic models are speaker-adaptive (SAT) and Maximum Mutual Information (MMIE) trained on about *1200* hours of audio data from a variety of broadcast sources and cover 33k context-dependent phones. They are gender-independent, word-position dependent tied-state, left-to-right phone HMMs with about 10k tied pdfs estimated with a DNN. The states are tied by means of a divisive decision tree with questions concerning the phone position, the phone identity and distinctive features and the neighboring phones in order to reduce model size and increase triphone coverage. The acoustic features are obtained by combining bottle-neck MLP outputs applied to raw PLP and TRAP features [6].

The system has a 250k word pronunciation lexicon, represented with 34 phones including specific units for silence, breath noise and filler words [9].

N-gram language models are trained on over 2 billions words of text from a large number of sources. Unpruned component LMs trained on different subsets of the training texts are interpolated for the final language models used for both decoding and lattice rescoring.

Word decoding is carried out in two passes, where each decoding pass produces a word lattice with cross-word, word-position dependent acoustic models, followed by consensus decoding with a 4-gram language model and pronunciation probabilities. Unsupervised acoustic model adaptation is performed for each segment cluster using CMLLR [18].

While the ASR word error rate (WER) is not known on this data, an older version of the system obtained a WER in the range of 9-28% (average 15%) across a variety of styles of broadcast data in the Quaero 2011 test [17].

2.2.2 Word Embedding (WE)

Word embedding or word2vec is a technique which maps words to vectors of real numbers in a space with lower dimension than the vocabulary size [2, 20]. Usual dimension are ranging between 80 and 2000. Most of the new word embedding techniques rely on a neural network architecture where bottleneck layer does the compression to the final vector.

We used a French word embedding model[5] built using Word2Vec [20] on the frWak[6] corpus [1]. The model had 200 embedding dimensions with a cutoff of 0 and the cbow algorithm.

2.2.3 Lexicon-based approach

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

[3]http://www.elra.info/en/projects/archived-projects/evalda/

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

[5]http://fauconnier.github.io/index.html#wordembeddingmodels

[6]http://wacky.sslmit.unibo.it/doku.php?id=corpora

We realized a valence and arousal estimation from text analysis, based on the transcriptions obtained from the ASR system previously described. Several methods have been approached, among them a lexicon-based system. A French lexicon of emotional words have been extracted from the emoBase platform[7] which stores resources gathered for the ANR project EMOLEX[8]. The semantic (high intensity, verbal demonstration, etc.) and emotion (joy, disappointment, contempt, etc.) labels, as well as the collocation information (as for context) provided by the corpus have been interpolated to estimate valence and arousal values for each entry of the lexicon. The latter were then applied to the training and development datasets, expecting high precision results.

2.3 Video Features

Organizers provided features including two types of facial descriptors: appearance and geometric based [22, 25]. 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 3s segments with 40 ms were used. The problem we experienced with the video features was that for parts of the data, the face was not recognized and no information was provided. For certain records, the amounts of unrecognized frames were up to 40%.

We have complemented the baseline video features by activations of a convolutional neural network (CNN) trained to localize facial landmarks [26] on the AFLW dataset. The regression network has 4 convolutional layers followed by a fully connected layer with absolute hyperbolic tangent activation. A final fully connected layer outputs x and y coordinates of 5 facial landmarks. It is necessary to use a pretrained network due to the very small size of the AVEC dataset, and the facial landmark localization task should be suitable for emotion recognition considering the good performance of the baseline geometric features on the valence task.

We have extracted the activations of the last convolutional layer (Video *CNN-L4*) and the first fully connected layer (Video *CNN-L5*) from the baseline facial regions enlarged by factor of 1.3 and rescaled to 40×40 pixels. The CNN-L4 features should contain more appearance information while the CNN-L5 should encode more geometric information.

2.4 Physiological Features

Physiological sets included Electrocardiogram (ECG) derived features, based on heart rate, its measure of variability, and derived parameters and statistics, and Electrodermal activity (EDA), skin conductance response (SCR), skin conductance level (SCL), as well as a number of derived parameters [22]

3. EXPERIMENTS AND ANALYSIS

3.1 Database

The data-set comes from RECOLA multimodal database [23]. It contains spontaneous interactions in French. Participants were recorded in dyads during a video conference while resolving of a collaborative task (âĂIJWinter survival taskâĂİ). 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 minutes of each recordings were rated by 6 French-speaking emotion raters in the continuous arousal-valence space, leading to 135 minutes of data with emotion ground truth. The database is freely available[9] and full details are provided in [23].

[7]http://emolex.u-grenoble3.fr/emoBase/index.php
[8]http://emolex.eu/

3.2 Evaluation and baselines

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

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

CCC produces values from -1 to 1. 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 [25]. These serve as baselines for our work and are mentioned in the tables in brackets.

3.3 Feature pre-processing

There are several steps to prepare features for regressor training. Each step has a different setting for different input features and different modality. Table 1 shows in condense form the settings of each pre-processing block which are described in more details below in this section.

At first, Principal Component Analysis (PCA) is used for dimensionality reduction. And the resulting features are normalized to have zero mean an unit variance.

In our experiments, we trained 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. This context is different for each modality.

We chose to provide the context primarily by stacking together features from a temporal neighborhood. The features themselves have quite smooth trajectories, so we do not need to take every frame but rather skip some frames, in order to keep the size of frame feature vectors manageable - we call it *sub-sampling*.

Further context is provided by computing local statistics for each feature. We compute mean, variance, maximum and minimum for each feature from a temporal window. Finally, we apply PCA again to reduce the size of feature vectors and we optionally normalize the features again to zero mean and unit variance.

We experimented with the delay applied to the gold-standard and optimal numbers in seconds for each modality are presented in the work too, This shift is consistent with previous works [14, 25].

All parameters in the Table 1 were obtained by grid search with the performance measured as CCC on the development partition of the AVEC 2016 database. Dash "-" in the Table 1 means that this step was skipped for particular subsystem.

[9]https://diuf.unifr.ch/diva/recola/

Table 1: Parameters of the single systems used for final fusion. The right side of the table lists the operations in the order they were computed. PCA - number of PCA components; norm. - per feature zero and mean normalization; stack - stacks frames from local window with temporal sub-sampling; stat. - computes statistics in local window per feature (min., max., mean, median); - shift features by n frames.

features	task	fusion	dev. CCC	PCA	norm.	stack	stat.	shift	PCA	norm.
audio BN FR 80	arousal	1 and 3	0.83	8	yes	201 (sub 5)	-	60	-	-
audio eGeMAPS			0.72	-	yes	-	-	60	-	-
audio BN l3			0.83	8	yes	60 (sub 20)	-	60	-	-
video appearance			0.44	40	yes	-	-	90	-	-
video geometric			0.44	64	yes	-	-	40	-	yes
ECG			0.31	-	yes	21	-	40	-	-
HRHRV			0.39	8	yes	-	-	0	-	-
SCL			0.13	4	yes	21	-	10	-	-
SCR			0.17	-	yes	41	-	40	-	-
video appearance	velence	1 and 3	0.39	64	yes	-	-	50	-	yes
video geometric			0.54	64	yes	-	-	90	-	-
audio BN-FR			0.50	8	yes	181 (sub 5)	-	50	-	-
audio BN-Multi			0.50	64	yes	161 (sub 5)	-	30	-	-
audio eGeMAPS			0.47	64	yes	-	-	100	-	-
ECG			0.27	8	yes	40	-	60	-	-
HRHRF			0.39	8	yes	-	-	30	-	yes
SCL			0.31	-	yes	-	-	20	-	yes
audio BN-FR	arousal	2 and 3	0.81	4	yes	120 (sub 20)	60	60	64	-
audio BN-Multi			0.83	8	yes	120 (sub 20)	-	90	-	-
audio eGeMAPS			0.79	-	yes	120 (sub 20)	-	90	64	-
ECG			0.32	24	yes	60 (sub 10)	-	30	64	-
Text ASR+WE			0.63	8	yes	120 (sub 20)	10	60	64	-
EDA	velence	2 and 3	0.32	-	yes	-	150	30	64	yes
HRHRV			0.19	-	yes	120 (sub 20)	-	30	-	-
SCL			0.21	-	yes	120 (sub 20)	-	30	-	-
video appearance			0.35	-	yes	-	150	60	64	yes
video geometric			0.58	4	yes	60 (sub 20)	60	60	64	-

3.4 Classifiers and Fusion

Linear regression is used on all single systems for arousal and valence. Liner regression is used also for the fusion. We have experimented with many classifiers (NN, RNN, LSTM, BLSTM) and their settings ... but we did not see any gain.

3.5 Individual systems

Table 2 summarizes our best results of single systems on development data for both modalities (Arousal and Valence). First part of the Table describes our systems with baseline features. We provide baseline results from organizers [25] for comparison, the numbers in brackets.

Second part of the Table 2 is reserved for our own feature extraction. We show results for our two systems based on Bottleneck features (BN) which outperform baseline audio system. Next line is reserved for text based word embedding system described in more details in Section 3.7.2. Last two lines are results from our two video features reaching CCC=0.617 on arousal and CCC=0.497 on valence.

Last line of Table 2 show our best fusion results.

Scores from all subsystem were smoothed with median filter with length 2.4 sec.

3.6 CCC as objective function

In general, regression systems should be trained with the same objectives as those used for evaluation. CCC is fully differentiable and can be easily integrated into gradient descent learning.

We trained diverse linear regressors using different regularizations and feature preprocessing pipelines to assess the effect of the objectives. We optimized the regressors using AdaDelta algorithm [27] with mini-batches of 256 frames. Figure 1 shows that CCC objective consistently improves results in the emotion recognition task compared to mean squared error loss (MSE). The average CCC improvement for arousal and valence is 0.048 (median 0.056) and 0.062 (median 0.066), respectively. Similar trend in performance was already reported in [14, 24].

Additionally, we compared CCC loss to mean absolute error loss (MAE) which previous work suggests is more suitable for the valence estimation task [3] than MSE. In our experiments CCC consistently improved results over MAE for arousal and valence on average by 0.054 (median 0.052) and 0.081 (median 0.079), respectively. The corresponding scatter plots are shown in Figure 2.

3.7 Text Based Features

3.7.1 Lexicon based approach

As for the lexicon based approach, a closer look at the database and transcripts revealed that only two words from the resulting lexicon appeared in the voice transcripts, and these two apparent transcription errors. Several factors may

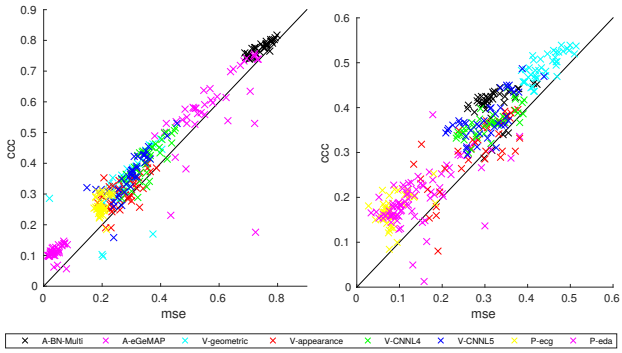

Figure 1: Scatter plots of MSE vs. CCC learning objective showing CCC scores of diverse linear systems on arousal (left) and valence (right).

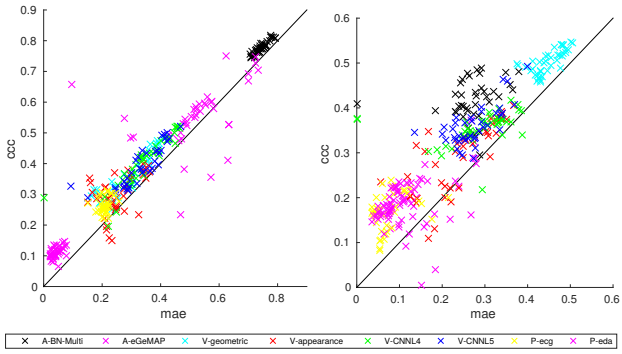

Figure 2: Scatter plots of MAE vs. CCC learning objective showing CCC scores of diverse linear systems on arousal (left) and valence (right).

Table 2: Comparison of single systems of different modalities, AV+EC 2016 baseline results are in brackets [25].

	Development	
CCC	Arousal	Valence
Audio	0.791 (0.796)	0.470 (0.455)
Video geometric	0.539 (0.379)	0.623 (0.612)
Video appearance	0.541 (0.483)	0.475 (0.474)
ECG	0.323 (0.271)	0.272 (0.153)
EDA	0.123 (0.077)	0.316 (0.194)
SCL	0.134 (0.101)	0.310 (0.124)
SCR	0.167 (0.071)	0.194 (0.110)
HRHRV	0.391 (0.382)	0.388 (0.293)
Audio BN-Multi	0.833	0.503
Audio BN-FR	0.830	0.497
Text ASR+WE	0.626	0.278
Video CNN-L4	0.595	0.497
Video CNN-L5	0.617	0.467
Fusion-Multimodal	0.855 (0.821)	0.713 (0.683)

have contributed to this: on casual examination, we observed that the transcripts were very low in verbal expressions of emotion and contained no apparent emotion-specific words; the lexicon may have been too generic, not covering such specific domain as the one from the provided corpus. More than analyzing the context around the words, a whole profiling of the domain should be done for such method to really be beneficial.

3.7.2 Word Embedding

This approach is more promising, since we can use the features and train classifier on the target database. The first results from this approach are in the Table 2 reaching on valence CCC=0.278 and CCC=0.626 for arousal.

3.8 Undefined regions

There is about 50% of speech in each audio file, the rest is silence. There is about 60% of detected face in video, the rest is unrecognized. It is obvious that we can not recognize emotion from audio if there is silence, and similar remark is applicable for video too. We present an analysis of training and evaluation of the system on all data, then *Defined* region (speech or face is detected) and *notDefined*) regions (silence/unrecognized face). Table 3 present results of such experiment for arousal. The system for video fulfilled our expectation that training and testing on matched data is better and provides performance gain, whereas scoring in *notDefined* region yield to poor performance. The audio system does not behave the same way and we are currently investigating on the reasons of this.

3.9 Fusion

Table 4 presents the results of our best fusions and submitted systems. We have submitted 3 systems to the challenge. All of them are fusions of several subsystems of different modalities. Table 1 gives the lists of subsystems belonging to each submission. Fusion 1 for arousal is from first block from the Table 1 and for valence from the second block. Fusion 2 is from third block for arousal and fourth for valence. The last Fusion for arousal contain all subsystems from Fu-

Table 3: Analysis of training and evaluating the system on all data, *Defined* region(speech or face detected) and *notDefined* region (silence/unrecognized face).

Arousal		CCC on Dev	
train	test	Audio	Video
all	all	0.836	0.448
all	Defined	0.837	0.464
all	notDefined	0.761	0.348
Defined	all	0.830	0.541
Defined	Defined	0.754	0.576
Defined	notDefined	0.770	0.327

Table 4: Results of final fusions submitted to AV+EC 2016.

	Development		Test	
CCC	Arousal	Valence	Arousal	Valence
Baseline [25]	0.820	0.702	0.682	0.638
Fusion 1	0.851	0.656	0.706	0.584
Fusion 2	0.852	0.589	0.708	0.505
Fusion 3	0.855	0.713	0.719	0.596

sion 1 and 2 and the same apply for valence. All our fusions consists of many systems and we will continue to work on analyzes which subsystems contribute the most.

Our fusion is better than the baseline from [25] except the results for valence on test set. We trained our systems and fusion on the train part of the AVEC database. Our fusion is not able to get the same gain as the baseline fusion for valence.

All systems are trained with only one output which is gold standard. We have experimented with other settings, separate raters etc, but did not get any improvement. The objective function is CCC.

Median filter from 100 frames (4 second) is applied on the top of the fusion scores.

Last thing which, unfortunately did not end up in the final fusion, are the statistics of *notDefined* regions. Easiest way of incorporating such statistics was to define a confidence vector for each single system with such statistics. This confidence vector is a binary vector with 1 at *Defined* regions and 0 at *notDefined* regions. Such confidence vector can be used in the fusion and tells us which system produce meaningful result for particular frame. We got a slight improvement (2% relative) on development set if such vectors are input vectors to the fusion. We are still experimenting how to use this information in the fusion and improve overall results.

4. CONLUSION

We substantially improved our system from last year submission [21]. This year we experimented again mainly with the audio modality. We improved our bottleneck feature system from CCC=0.699 [21] to CCC=0.833 on development set.

We also newly experimented with text based features. The automatic speech recognition was used to get text transcriptions. First results CCC=0.278 for valence and CCC=0.626 for arousal was obtained with word embedding approach. This modality is new to this field and our plan is to experiment more in that direction: comparing different speech transcription systems, applying sentiment recognition to the French corpus by building a French sentiment model, or else translating the text into English in order to use one of the main English sentiment detector tools available.

To summarize our effort and compare it to baseline system, our single best system for both modalities are better than the baseline single best systems, for arousal it is even better then the baseline fusion. The CCC of our single best systems is 0.833 for arousal and 0.623 for valence on development set. Our final linear fusion reached CCC 0.855 and 0.713 on Arousal and Valence on development set and 0.713 and 0.596 on test set respectively for Arousal and Valence.

5. ACKNOWLEDGMENTS

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.

6. ADDITIONAL AUTHORS

Additional authors: Ian Wood (NUIG, IE), Cécile Robin (NUIG, IE), Lori Lamel (LIMSI, CNRS, Université Paris Saclay, FR).

7. REFERENCES

[1] M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226, 2009. 00587.

[2] Y. Bengio, H. Schwenk, J.-S. Senécal, F. Morin, and J.-L. Gauvain. *Neural Probabilistic Language Models*, pages 137–186. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[3] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen. Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, AVEC '15, pages 65–72, New York, NY, USA, 2015. ACM.

[4] S. Cumani, P. Laface, and F. Kulsoom. Speaker recognition by means of acoustic and phonetically informed gmms. In *Proceedings of Interspeech 2015*, 2015.

[5] R. Fér, P. Matějka, F. Grézl, O. Plchot, and J. Černocký. Multilingual bottleneck features for language recognition. In *Proceedings of Interspeech 2015*, pages 389–393, 2015.

[6] P. Fousek, L. Lamel, and J.-L. Gauvain. Transcribing broadcast data using mlp features. *InterSpeech*, 8:1433–1436, 2008.

[7] D. Garcia-Romero and A. McCree. Insights into deep neural networks for speaker recognition. In *Proceedings of Interspeech 2015*, 2015.

[8] J. Gauvain, L. Lamel, and G. Adda. The LIMSI Broadcast News Transcription System. *Speech Communication*, 37(1-2):89–108, 2002.

[9] J.-L. Gauvain, G. Adda, M. Adda-Decker, A. Allauzen, V. Gendner, L. Lamel, and H. Schwenk. Where are we in transcribing french broadcast news? In *InterSpeech*, pages 1665–1668, 2005.

[10] J.-L. Gauvain, L. Lamel, and G. Adda. Partitioning and transcription of broadcast news data. *ICSLP*, 98(5):1335–1338, 1998.

[11] F. Grézl, E. Egorova, and M. Karafiát. Further investigation into multilingual training and adaptation of stacked bottle-neck neural network structure. In *Proceedings of 2014 Spoken Language Technology Workshop*, pages 48–53, 2014.

[12] H. Gunes and B. Schuller. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 31(2):120 – 136, 2013. Affect Analysis In Continuous Input.

[13] M. Harper. The BABEL program and low resource speech technology. In *ASRU 2013*, Dec 2013.

[14] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, AVEC '15, pages 73–80, 2015.

[15] M. Karafiát, K. Veselý, I. Szoke, L. Burget, F. Grézl, M. Hannemann, and J. Černocký. But ASR system for BABEL surprise evaluation 2014. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, NV, USA, Dec 2014.

[16] M. H. Kornel Laskowski and J. Edlund. The fundamental frequency variation spectrum. In *Proc. FONETIK*, 2008.

[17] L. Lamel. Multilingual speech processing activities in quaero: Application to multimedia search in unstructured data. In *Baltic HLT*, pages 1–8, 2012.

[18] C. Leggetter and P. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9(2):171–185, 1995.

[19] P. Matějka, L. Zhang, T. Ng, H. S. Mallidi, O. Glembek, J. Ma, and B. Zhang. Neural network bottleneck features for language identification. In *Proceedings of Odyssey 2014*, pages 299–304, 2014.

[20] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013. 00754.

[21] A. Popková, F. Povolný, P. Matějka, O. Glembek, F. Grézl, and J. H. Černocký. Investigation of bottle-neck features for emotion recognition. In *Text Speech and Dialog (TSD)*, 2016.

[22] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic. Av+ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In *Proc. AVEC 2015, satellite workshop of ACM-Multimedia 2015*, Brisbane, Australia, Oct. 2015.

[23] F. R. A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *Proc. Face and Gestures 2013, Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE)*, 2013.

[24] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. Nicolaou, S. B., and S. Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proceedings of the 41st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016.

[25] M. F. Valstar, J. Gratch, B. W. Schuller, F. Ringeval, D. Lalanne, M. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. AVEC 2016 - depression, mood, and emotion recognition workshop and challenge. *CoRR*, abs/1605.01600, 2016.

[26] Y. Wu and T. Hassner. Facial landmark detection with tweaked convolutional neural networks. *CoRR*, abs/1511.04031, 2015.

[27] M. D. Zeiler. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701, 2012.