EFFECTIVE EMOTION RECOGNITION IN MOVIE AUDIO TRACKS

Margarita Kotti[1] *and Yannis Stylianou*[1,2]

[1] Toshiba Research Europe Ltd., Cambridge Research Lab, Cambridge, U.K.
[2] Department of Computer Science, University of Crete, Greece

ABSTRACT

This paper addresses the problem of speech emotion recognition from movie audio tracks. The recently collected Acted Facial Expression in the Wild 5.0 database is used. The aim is to discriminate among angry, happy, and neutral. We extract a relatively small number of features, a subset of which is not commonly used for the emotion recognition task. Those features are fed as input to an ensemble classifier that combines random forests with support vector machines. An accuracy of 65.63% is reported, outperforming a baseline system that uses the K-nearest neighbor classifier and has an accuracy of 56.88%. To verify the suitability of the exploited features, the same ensemble classification schema is applied on the feature set similar those employed in Audio/Visual Emotion Challenge 2011. In the latter case, an accuracy of 61.25% is achieved using a large set of 1582 features, as opposed to just 86 features in our case that lead to a relative improvement of 7.15% in accuracy.

Index Terms— emotion recognition, speech features, random forests, support vector machines, ensemble classifiers

1. INTRODUCTION

Emotion recognition is an active research area with many applications such as human assistive systems [1], autonomous video summarisation [2], diagnosing patients mental illness, monitoring the drivers emotion variations to avoid accidents and helping the man-machine interactions [3]. Emotion recognition systems can also find applications in key event detection tasks [2], affective analysis in music [4] or dialogue management [5].

Although much research has been carried out in the last three decades [6], the problem is far from trivial. When human-computer interaction is based only on the audio channel, the problem becomes even more challenging, since the recognition is based solely on voice, which is the basic mean of human communication [7]. In fact, it is a complex, challenging task since emotion is implicitly conveyed through the external behavioral manifestations [8]. As emotional states do not have clear-cut boundaries and they often differ from person to person, sometimes even a human cannot easily classify natural emotions based on speech hue [7]. Much of the emotion recognition research uses the extraction of acoustic parameters from the speech signal as a method to capture changes in the acoustic waveform that are representative of emotional content [9]. Commonly extracted are the features related to pitch, formants, loudness, harmonic-to-noice-ratio, harmonic rations, jitter and shimmer [9]. Several classifiers have been used in the past, such as non-negative matrix factorisation [3], Gaussian Mixture Models [10], Hidden Markov Models [11], Support Vector Machines [12], Artificial Neural Networks (ANNs) either swallow or deep [13], Decision Trees or k-Nearest Neighbor distance classifiers [14].

In this paper we propose the use of a small set of features, namely: MFCCs, LPCs, ZCR, Spectal Flux, Spectral Rolloff, Chroma, and Clarity. This is the first time that Clarity is being used for the speech emotion recognition task, to the best of the authors' knowledge. Mean and standard deviation of the aforementioned features are provided as input to three independent classifiers namely (i) random forests, (ii) linear SVMs, and (iii) polynomial SVMs. Fusion is carried out at decision level, since previous research on the emotion recognition task indicates that decision fusion gives better results compared to feature level fusion [15]. Majority voting has an accuracy of 65.63% on the challenging and recently collected Acted Facial Expression in the Wild (AFEW) 5.0 database when discriminating angry, happy, and neutral.

In brief, the main contribution of our work is 3-fold: (i) the compact feature set, containing novel features, (ii) the database, and (iii) the ensemble classifier. Regarding the audio features, although all features, but Clarity have been used before for emotion recognition, usually a much larger set of audio features is employed. For example, the organisers of the Emotion Recogntion in the Wild challenge [16] who collected the AFEW 5.0 database extracts a large pool of 1582 features, as opposed to just 86 features in our case. Moreover, we select a constrained number of two functionals: mean and standard deviation, whereas it is a common approach to consider a much larger selection of them, such as skewness, kurtosis, and percentiles [3]. Additionally, to the best of the authors' knowledge this is the first time that the voice activity detection-inspired feature of Clarity is applied for the emotion recognition task. Experimental results have also shown the superiority of the proposed small-scale feature set, since using the feature set of 86 features lead to a relative improvement of 7.15% in accuracy compared to using the feature set of 1582 features. The second contribution refers to the database. This is the first time that the audio channel of AFEW 5.0 has been investigated in depth. Specifically, AFEW 5.0 is a very challenging database and most of the research effort goes towards the video channel. Last year a very limited number of teams that participated in the Third Emotion Recognition in the Wild Challenge considered the audio stream. Namely the authors of [17] and [15] consider the pre-extracted 1582 features, whereas alternative feature sets are investigated by the authors of [18]. However, in all those cases, the results refer to the fusion of audio and visual channels ranging from 31.54% to 33.96% for the emotional categories of angry, disgust, fear, happy, sad, surprise, and neutral. Thirdly, a novel ensemble classification scheme is employed. Compared to a baseline K-nearest neighbor classifier a relative improvement of 15.38% is accomplished.

The rest of the paper is organised as follows: The proposed ensemble classification method is described in Section 2 along with the exploited audio features. Emphasis is given on Clarity that has not been previously used for the emotion recognition task. Experiments using this set of 86 features along with the proposed ensemble classifier are detailed in Section 3, where also the AFEW 5.0 database

is summarised. Discussion is carried out in Section4, where a comparison with a baseline KNN classifier as well as with features similar to those employed in Audio/Visual Emotion Challenge (AVEC) 2011 as extracted by openSMILE/openEAR is performed. Finally, conclusions and future work are presented in Section 5.

2. METHOD

In this Section we provide a short mathematical foundation of the proposed method. The proposed method combines signal processing for feature extraction from the speech signal with machine leaning for emotion classification.

Regarding the classification part, this is done by an ensemble classification schema, also known as classification committee, that combines 2 independent classifiers, namely Random Forests (RF) and Support Vector Machines (SVMs). Let us consider the classification task for a set of training data $\mathbf{X} = \{(\mathbf{x}(i), t(i))|i = 1, ..., N\}$, where $\mathbf{x}(i) \in \mathbb{R}^n$ is a feature vector and $t(i)$ is the class label.

For the random forest case, an ensemble of randomly trained decision trees is built. A decision tree is grown recursively by partitioning the training data $\mathbf{x}(i)$ to successive subsets that contain as many samples of the same class $t(i)$ as possible. So, at the root of the tree all training samples are present and then based on a splitting criterion, the samples are partitioned into two child nodes. The splitting criterion here is Gini's index

$$GInd = 1 - \sum_i (p|t(i))^2 \quad (1)$$

where $p|t(i)$ is the observed fraction of samples with class $t(i)$ that reaches the node. This procedure is recursively applied to each child node until all the records in a node J belong to the same class $t(i)$.

Random forest is comprised of B bagged trees, where all trees are randomly different from one another. This leads to decorrelation between the individual tree predictions and, in turn, results in improved generalization and robustness [19]. The algorithm can be seen in Algorithm 1. During testing, the decisions $D_i(\mathbf{y}), i = 1...B$

Algorithm 1 The Random Forest Algorithm

for $tree\_counter$=1 to B **do**
1. Draw bootstrap sample $\mathbf{x}_b(i)$ of training data $\mathbf{x}(i)$
2. Grow unpruned tree by
for each node **do**
1. Select m samples $\mathbf{x_{mb}}(i)$ out of $\mathbf{x}_b(i)$
2. Calculate the best split among the m according to a criterion
3. Split the node
end for
end for

of each independent tree are combined in a majority voting fashion so that the final decision D_{rf} over an unseen testing feature vector $\mathbf{y}$ is

$$D_{rf}(\mathbf{y}) = majority\_vote\{D_i(\mathbf{y})\}, i = 1...B \quad (2)$$

With respect to SVMs, they are binary maximum margin classifiers that try to find the hyperplane which optimally separates the data. The hyperplane can be described as $\mathbf{w}^T\mathbf{x} + b = 0$, where $\mathbf{w}$ is a weight vector estimated during training and b is the bias. The binary classification problem is described as

$$\mathbf{t}(i)(\mathbf{w}^T\mathbf{x}(i) + b) - 1 \geq 0 \quad s.t \quad \min \frac{1}{2}|\mathbf{w}|^2. \quad (3)$$

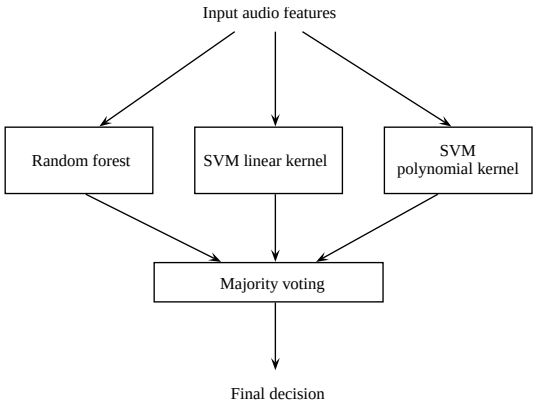

Fig. 1. The proposed ensemble classifier. Random forests and SVMs makes their own independent decisions which are then combined by majority voting.

The final ensemble classifier works at fusing the decisions of each independent classifier i.e. (i) random forest, (i) linear SVM, and (iii) polynomial SVM at decision level. The final label is the one obtained by the majority of the classifiers. Uncorrelated errors of individual classifiers can be eliminated by averaging. The proposed ensemble classifier is depicted in Figure 1.

Regarding the feature extraction part of this paper, it uses a relatively small number of features that are fed as input to an ensemble classification schema. We tested Clarity, that is commonly used in voice activity detection [20], [21] as a candidate feature for emotion recognition. The reason for that choice is that emotion recognition is a complex, versatile task, so alternative features may capture supplementary aspects of emotion expression. Since after a decade of research on emotion recognition the golden set from an endless list of non-linguistic features has not been found yet [7], it is worth testing for non trivial solutions.

Here, Clarity is defined as the relative depth of the minimum average magnitude difference function valley in the plausible pitch range [21]:

$$C(\tau) = 1 - \frac{A(\tau, k_{min})}{D(\tau, k_{max})} \quad (4)$$

where τ and k are frame and autocorrelation lag indices, respectively, and

$$A(\tau, k) \approx \beta(k)\sqrt{2[r_{xx}(\tau, 0) - r_{xx}(\tau, k)]} \quad (5)$$

where r_{xx} is the autocorrelation and

$$k_{min} = \underset{2ms \leq k \leq 16ms}{\operatorname{argmin}} A(t, k) \quad (6)$$

where k_{max} is defined as in Eq. (6), but with argmax and $\beta(k)$=0.8 is a scaling factor.

3. EXPERIMENTS

3.1. Database

We used the database collected for the third Emotion Recognition in the Wild (EmotiW) challenge 2015 [16]. This is an audio-visual data corpus comprising of scenes collected from movies, thus showing close-to-real-world conditions. AFEW is developed in a semi-automatic manner, parsing the subtitles for presence of

keywords related to emotion. The emotional categories are: angry, disgust, fear, happy, sad, surprise, and neutral. The emotions are annotated by 3 annotators; clips have a duration of 300-5400 ms, and the train (723 samples) and validation (383 samples) sets are publicly available and more information can be found here: https://cs.anu.edu.au/few/emotiw2015.html.

For this work, we limit ourselves to the emotional categories of angry, happy, and neutral. This subset is selected from a practical point of view, since it is fundamental to know whether the expressed emotion is negative or positive. Possible applications include a call-centre environment, where such an emotion recognition schema can be used to improve the quality of service. Furthermore, by discriminating negative from non-negative emotions, human-computer interaction designers will be able to recognize which parts of the interface are problematic, in the sense that they evoke negative emotions [22]. With respect to the audio, this is extracted from the audio-visual clips as monochannel wav files of a 48kHz sampling rate. Since the audio clips are not recorded it restricted lab conditions, they may contain for example background noise, music, or speech, as well as overlapping speakers and reverberation.

3.2. Proposed system

A pool of 86 features is extracted for this paper. This consists of the low level descriptors and functionals depicted in Table 1. Specifically, we have 43 low level descriptors * 2 functionals = 86 features. Since we compute the energy of the signal, we can disregard the first MFCC. Those features are fed as input to the proposed ensemble classification schema.

Referring to the novelty of this paper with respect to feature extraction from the speech signal, our contribution is two-fold. Firstly, this paper suggests the use of audio features that have not been widely used for speech emotion recognition. The second contribution lies in the use of a small feature collection. As detailed in Section 4, in order to prove the suitability of this small feature set, we compare it against the use of 1582 features, extracted using the Emotion and Affect Recognition (openEAR) [23] toolkit backended with openSMILE [24]. This large set of audio features is similar to the features employed in AVEC 2011 [25].

Low Level Descriptor	Functionals
Energy	Mean and standard deviation
MFCCs (1-12)	
LPCs (0-13)	
ZCR	
Spectral Flux	
Spectral Rolloff	
Chroma Vector (0-11)	
Clarity	

Table 1. Extracted audio features

For training we use the training set (394 clips for the emotional categories of angry, happy, and neutral) and for validation and testing the validation set (additional 190 clips) of the AFEW 5.0 database. The test set of AFEW 5.0 database is obviously not publicly available. Validation is needed in the proposed approach to determine the optimal parameters, such as the order of the polynomial kernel. For that reason we retain 10 files per emotional category from the initial AFEW 5.0 validation set, leaving the additional 160 files available for testing. Validation determined the number of trees to be 100, the optimal polynomial kernel order to be 3, and allows 1% of the training examples to be out-layers in both the polynomial and the linear kernel. Accuracy is 58.13% for the linear SVM and 56.87% for the polynomial one.

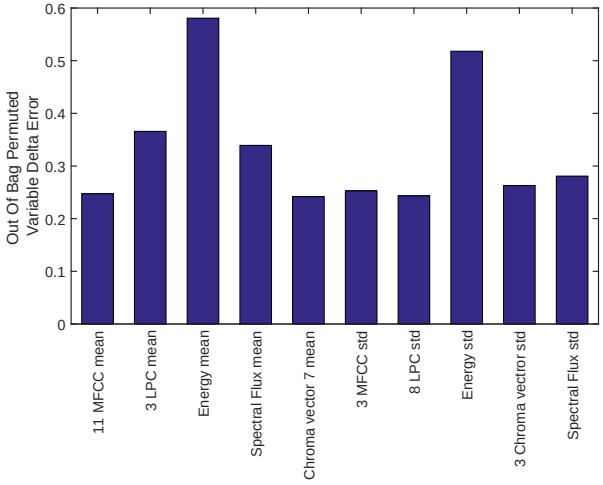

Fig. 2. Increase in prediction error if the values of the feature are permuted across the out-of-bag observations. The 10 largest values are reported.

Table 2. Confusion matrix for the proposed approach (i.e. 86 feature set and classification ensemble)

True Emotion \ Predicted Emotion	Angry	Happy	Neutral
Angry	43	7	4
Happy	13	26	14
Neutral	3	14	36

Regarding the random forest, as said, it uses 100 classification trees, with 72 nodes per tree, on average. All input features are sampled with replacement. Each tree is constructed using a different bootstrap sample from the training data that includes two thirds of the features, so the remaining one-third is left out, thus constituting the out-of-bag features. The number of features to select at random for each decision split is 10. The cost for misclassification is the same across the three classes. Prior probability is 0.333 for each class (empirical probability). Accuracy is 58.57% for the random forest. The ensemble classifier provides an accuracy of 65.63%, whereas the detailed confusion matrix can be seen in Table 2.

To provide an insight in the importance of the features for the random forest, we compute the increase in prediction error if the values of that feature are permuted across the out-of-bag observations. The increase in the prediction error if the values of that feature are permuted across the out-of-bag observations is computed for every tree, then averaged over the entire ensemble and divided by the standard deviation over the entire ensemble. For this work, the 10 most informative parameters are depicted in Figure 2. As can be seen from this Figure, the energy plays the most important role, followed by LPC and MFCC related parameters as well as Spectral Flux and Chroma vector ones.

Further experimentations concluded that if Clarity is removed, then accuracy drops to 63.12%. McNemar test for 95% confidence

Low Level Descriptor	Functionals
Loudness, delta coefficients	Absolute position of the maximum/minimum value, mean, slope and offset of a linear approximation of the contour, linear/quadratic error, standard deviation, skewness, kurtosis, 25%,50%, 75% percentile, inter-quartile ranges, outlier-robust maximum/minimum value of the contour, percentage of time the signal is above (75%/90% * range + min)
MFCCs (0-14), delta coefficients	
Logarithmic power of Mel-frequency bands (0-7), delta coefficients	
Line spectral pair frequencies (0-7), delta coefficients	
Envelope of the smoothed fundamental frequency contour, delta coefficients	
Voicing probability, delta coefficients	
F0, delta coefficients	
Local jitter, delta coefficients; jitterDDP, delta coefficients	
Local shimmer, delta coefficients	

Table 4. Pre-extracted audio features in the AFEW 5.0 database using openEAR back-ended with openSMILE.

Table 3. Confusion matrix for the KNN classifier using the 86-dimensional feature set

		Predicted Emotion		
		Angry	Happy	Neutral
True Emotion	Angry	38	7	9
	Happy	13	24	16
	Neutral	7	17	29

Table 5. Confusion matrix for the pre-extracted features using openEAR and the ensemble classification schema

		Predicted Emotion		
		Angry	Happy	Neutral
True Emotion	Angry	40	9	5
	Happy	13	26	14
	Neutral	6	15	32

interval showed no statistical significant difference among any combination of the 3 different classifiers.

4. DISCUSSION

Aiming to prove the efficiency of the proposed approach with respect to (i) the classification method and (ii) the extracted features, additional experiments took place. Regarding (i) we utilised a base-line KNN classifier and for (ii) we exploited the feature set that was pre-extracted by the emotion in the wild challenge organisers using openEAR back-ended with openSMILE. Regarding the AFEW 5.0 splits, the same training, testing, and validation sets were utilised, as in the proposed approach.

4.1. Comparison with KNN

For this comparison, we substituted the ensemble classification schema with a base-line KNN classifier. The rest of the experimental protocol remains the same as in the proposed approach. So, the feature set is comprised of the 86 features described in Table 1. The optimal number of neighbors in the validation set was found to be 16 and the selected distance function is one minus the correlation between the feature vectors. An accuracy of 56.88% was reached, that equals to an absolute deterioration of 15.38%. The confusion matrix can be seen in Table 3.

4.2. OpenEAR/openSMILE pre-extracted features

To validate that the proposed feature set is a suitable choice, we compared the accuracy of the proposed ensemble classification system when the classifiers' input of 86 features is substituted by the 1582 features that are pre-extracted as part of the AFEW 5.0 database. Feature extraction was carried out by means of the open-source toolkit openEAR [23] toolkit back-ended with openSMILE [24]. This large set of audio features is similar to the features employed in AVEC 2011 [25]. The extracted features were selected based on a) their potential to index affective physiological changes in voice production, b) their proven value in former studies as well as their automatic extractability, and c) their theoretical significance [9]. Specifically, AFEW 5.0 offers pre-extracted the low level descriptors and functionals depicted in Table 4. Features listed in Table 4 are smoothed by a moving average filter with window length 3. The statistical function of the percentage of time that the signal is above a threshold is computed only for those features, where that is meaningful. A couple of more F0 features are computed, namely segment duration and number of onsets.

When the proposed ensemble classification system is trained on those features and tuned using the same validation set as in the proposed approach, accuracy equals 61.25%, a 7.15% relative deterioration, compared to the proposed approach. This could potentially be attributed to the fact that the extracted features are too many for discriminating among 3 classes of 394 training examples. In other words, it is our coarse speculation that the emotional space representation using all the pre-extracted features may potentially lead to an insufficient number of representative examples for each class, since each example is 1582-dimensional and there are 118 to 145 training clips per class. The detailed confusion matrix can be seen in Table 5.

5. CONCLUSIONS AND FUTURE WORK

This paper deals with the problem audio emotion recognition in movie clips. It presents a classification committee that takes an audio stream as input and recognises an emotional category among happy, angry, and neutral for output. Individual classifiers are random forests and SVMs, both linear and polynomial, the decision of which as fused. The database used to test the efficiency of the proposed method is the challenging AFEW 5.0 that contains high background noise/music and overlapping speakers. The size of the extracted audio features set is limited to 86. An accuracy of 65.63% is reported, outperforming a big audio feature set, comprising of 1582 features inspired by the AVEC2011 challenge.

In the future, we aim to investigate further audio features that are not traditionally used for speech emotion recognition. However, the aim is to retain the extracted feature set cardinality small, by replacing some of the existing features. To further improve performance and boost robustness, we plan to investigate more sophisticated classification committee realisations, namely Bayesian model averaging.

6. REFERENCES

[1] R. W. Picard, *Affective Computing*, MIT Press, Cambridge, MA, USA, 1997.

[2] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1553–1568, Nov. 2013.

[3] P. Song, S. Ou, W. Zheng, Y. Jin, and L. Zhao, "Speech emotion recognition using transfer non-negative matrix factorization," in *ICASSP 2016*, Mar. 2016, pp. 5180–5184.

[4] Y. E. Kim, E. M. Schmidt, R. Migneco, B.G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Music emotion recognition: A state of the art review," in *ISMIR 2010*, Utrecht, The Netherlands, Aug. 2010, pp. 255–266.

[5] Y. Xiaobu, "Lightly-supervised utterance-level emotion identification using latent topic modeling of multimodal words," in *ICSI/CCI 2015*, Jun. 2015, pp. 297–308.

[6] F. B. Pokorny, F. Graf, F. Pernkopf, and B. W. Schuller, "Detection of negative emotions in speech signals using bags-of-audio-words," in *ACII 2015*, Sep. 2015, pp. 879–884.

[7] C.N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011," *Artificial Intelligence Review*, vol. 43, no. 2, pp. 155–177, Feb. 2015.

[8] Z. Yang and S. Narayanan, "Lightly-supervised utterance-level emotion identification using latent topic modeling of multimodal words," in *ICASSP 2016*, Mar. 2016, pp. 2767–2771.

[9] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. Andre, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, 2016, to appear.

[10] T. Kostoulas, T. Ganchev, A. Lazaridis, and N. Fakotakis, *Text, Speech and Dialogue: 13th International Conference, TSD 2010, Brno, Czech Republic, September 6-10, 2010. Proceedings*, chapter Enhancing Emotion Recognition from Speech through Feature Selection, pp. 338–344, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[11] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *ICASSP 2003*, Apr. 2003, vol. 2, pp. II–1–4 vol.2.

[12] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, Jul. 2010.

[13] 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," in *ICASSP 2016*, Mar. 2016.

[14] T. L. Pao, W. Y. Liao, Y. T. Chen, J. H. Yeh, Y. M. Cheng, and C. S. Chien, "Comparison of several classifiers for emotion recognition from noisy mandarin speech," in *IIHMSP 2007*, Nov. 2007, vol. 1, pp. 23–26.

[15] H. Kaya, F. Gürpinar, S. Afshar, and A. A. Salah, "Contrasting and combining least squares based learners for emotion recognition in the wild," in *ICMI '15*, New York, NY, USA, 2015, ICMI '15, pp. 459–466, ACM.

[16] A. Dhall, O.V. R. Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and image based emotion recognition challenges in the wild: EmotiW 2015," in *ICMI 2015*, NY, USA, 2015, pp. 423 – 426.

[17] J. Wu and H. Lin, Z.and Zha, "Multiple models fusion for emotion recognition in the wild," in *ICMI 2015*.

[18] B. Sun, L. Li, G. Zhou, X. Wu, J. He, L. Yu, D. Li, and Q. Wei, "Combining multimodal features within a fusion network for emotion recognition in the wild," in *ICMI 2015*, New York, NY, USA, 2015, ICMI '15, pp. 497–502, ACM.

[19] A. Criminisi, J. Shotton, and E. Konukoglu, "Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning," *Foundations and Trends in Computer Graphics and Vision*, vol. 7, no. 2-3, pp. 81–227, Feb. 2012.

[20] T. Drugman, Y. Stylianou, Y. Kida, and M. Akamine, "Voice activity detection: Merging source and filter-based information," *IEEE Signal Processing Letters*, vol. 23, no. 2, pp. 252–256, Feb. 2016.

[21] S. O. Sadjadi and J. H. L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 197–200, Mar. 2013.

[22] M. Kotti and F. Paternò, "Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema," *International Journal of Speech Technology*, vol. 15, no. 2, pp. 131–150, 2012.

[23] F. Eyben, M. Wllmer, and B. Schuller, "openEAR - introducing the munich open-source emotion and affect recognition toolkit," in *ACII 2009*, Sep. 2009, pp. 1–6.

[24] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE: The Munich versatile and fast open-source audio feature extractor," in *MM 2010*, New York, NY, USA, 2010, MM '10, pp. 1459–1462, ACM.

[25] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011 the first international audio/visual emotion challenge," in *ACII 2009*, Oct. 2011, pp. 415–424.