AN UNSUPERVISED FRAME SELECTION TECHNIQUE FOR ROBUST EMOTION RECOGNITION IN NOISY SPEECH

Meghna Pandharipande, Rupayan Chakraborty, Ashish Panda, Sunil Kumar Kopparapu

TCS Research and Innovations Mumbai, Thane West-400601, Maharashtra, INDIA

ABSTRACT

Automatic emotion recognition with good accuracy has been demonstrated for clean speech, but the performance deteriorates quickly when speech is contaminated with noise. In this paper, we propose a front-end voice activity detector (VAD)-based unsupervised method to select the frames with a relatively better signal to noise ratio (SNR) in the spoken utterances. Then we extract a large number of statistical features from low-level audio descriptors for the purpose of emotion recognition by using state-of-art classifiers. Extensive experimentation on two standard databases contaminated with 5 types of noise (Babble, F-16, Factory, Volvo, and HF-channel) from the Noisex-92 noise database at 5 different SNR levels ($0, 5, 10, 15, 20$dB) have been carried out. While performing all experiments to classify emotions both at the categorical and the dimensional spaces, the proposed technique outperforms a Recurrent Neural Network (RNN)-based VAD across all 5 types and levels of noises, and for both the databases.

Index Terms— Speech emotion, Noisy speech, Voice activity detector, Emotion recognition

1. INTRODUCTION

Emotion recognition from the clean speech is an active research area, and the performance has been reported with good accuracies [1–3]. However, the interest is already shifting to identify emotion in a speech from the realistic environments, where the signals are expected to be noisy. Like other speech technologies, emotion recognition also suffers from the degradation in performance in noisy speech.

Emotion recognition in noisy speech has been addressed in the literature using different methods, like using a robust set of acoustic features, enhancing signals, eliminating noise, adapting models and compensating features etc. In [4, 5], authors used a large set of robust acoustic features with Information-Gain-Ratio (IGR) filter-selection for picking up attributes according to noise conditions. Emotion recognition from noisy speech has also been performed by using Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) based feature reduction with sequential forward selection (SFS) [6]. Noise robust feature selection using weighted discrete-KNN has been found to be beneficial for detecting emotion in noisy speech [7].

In [8], histogram equalization is used to reduce the difference between feature vectors (pitch and MFCC) in clean and noisy conditions. Authors have used speech enhancement algorithms, based on spectral subtraction and the masking property to improve emotion classification for white noise contaminated speech in [9]. In [10], authors investigated audio signal denoising methods in cepstral and log-spectral domains and compared them with standard techniques by using average acoustic distance metrics. Further, they were applied to automatic recognition of natural and spontaneous emotions under simulated smart-phone-recorded noisy conditions. In [11], authors presented a noise cancellation framework based on the adaptive thresholding in wavelet domain as a front end to the speech emotion recognizer. In [12], spectral subtraction, Wiener filter and Minimum Mean Square Error (MMSE) based noise reduction has been used for the analysis of emotional speech. An approach of robust emotion recognition in noisy speech is proposed in [13], by using a weighted sparse representation model, which is based on the maximum likelihood estimation. In [14], authors have achieved better robustness with Deep Belief Networks (DBN) for emotion recognition.

In this paper, we propose to use a front-end signal processing block for discarding the noise affected and silence frames in spoken utterances. In particular, the frame selection is based on a VAD, which helps in identifying the frames with high SNR levels. Since noise/silence frames are not expected to contribute positively to emotion recognition, we expect that significant performance gain can be achieved by discarding frames with the high noise level. Compared to the noise robust techniques described in the above paragraphs, the proposed approach is advantageous in several ways. First, it can be used along with any of the previously mentioned noise robustness techniques. Second, unlike the model adaptation or feature compensation technique, the proposed method is not specific to a particular model or feature. Third, the computational complexity of the proposed method is extremely low compared to other noise robustness techniques or even other supervised VADs. It should be noted that the proposed VAD is not designed to counter the Lombard effect in noisy speech [15]. Nonetheless, it will improve the performance of

ISBN 978-90-827970-1-5

Fig. 1. Noisy speech emotion recognition system

the system by discarding spurious frames and it can be used along with the techniques designed specifically for the Lombard effect.

Although VAD has been used in emotion recognition task (e.g. [16] [17]), it has not been studied for its effectiveness in noisy speech scenarios. In this paper, we show that the proposed method results in significant performance gain, even without any other noise robustness techniques. We also compare the performance of our technique with that of the RNN based VAD (from openSMILE toolkit) [18]. The latter is a data-driven approach and uses noise models trained from a large set of noisy signals and is thus computationally complex. Despite being unsupervised and computationally simple, the proposed method outperforms the RNN based VAD by a significant margin.

The rest of the paper is organized as follows. Section 2 presents emotion recognition system, along with our proposed VAD algorithm and the RNN-based VAD. In section 3, we explain the experimental setup, databases, results, and analysis. We conclude in Section 4.

2. VOICE ACTIVITY DETECTOR BASED EMOTION RECOGNITION SYSTEM

The noisy speech emotion recognition system is designed as depicted in Figure 1. As mentioned earlier, the main contribution of this paper is to propose, design and implement the front-end of the system, i.e. the VAD-based frame selection block. At the output of this block, noisy frames from the input spoken utterance are discarded based on the relative energy based algorithm and expected to reduce the overall noise energy from the noisy speech.

2.1. Proposed VAD for noise reduction

The proposed VAD relies on tracking the noise floor and then selecting the speech frames which have the desired energy levels. This continuous and adaptive estimation of the noise floor and the ability of the VAD to set the desired energy value for frames to be selected make it eminently suitable for front-end processing in emotion recognition. The proposed VAD is described below:

The input speech signal is divided into frames of a certain duration (say 20 ms). The noise floor n_f is initialized to a certain value (in our case, we have used the average energy of the first four frames). This represents the average energy of a frame of noise. Next, energy x_f is computed for each frame of the input speech signal. However, we follow a different process to compute the frame energies. To compute the energy of i^{th} frame, we compute the average energy of frames $(i-k)$ through $(i+k)$. Next, for frames $i = 1, 2, .., T$, a ratio r is computed as x_f/n_f. A frame is rejected if the ratio falls below a certain threshold θ. If a frame is rejected, then the frame energy contributes to the noise floor adaptation in the following way. If the rejected frame energy is less than a certain range of the floor value (i.e., if $x_f \leq w_h n_f$, where w_h is a multiplication factor), then the rejected frame energy goes to an array N of P values arranged in First-In-First-Out manner. The new noise floor is then computed as the average of the P values in the array N. Thus, the noise floor is updated for every frame and the value of the noise floor changes with time.

Although above described VAD shares some attributes with the VAD described in [19], it has certain new and important traits that make it more robust in noisy speech emotion recognition. First, the computation of the frame energy is done over a neighborhood. This prevents sudden and transient fluctuation in energy due to the noise clicks and peaks from affecting the frame energies. Second, the change in noise floor is more gradual due to the array of P values. This lessens the effect of the spuriously rejected frames on the noise floor. Third, we are not considering buffer frames, which makes the VAD more efficient. The buffer frames prevent clipped words and hence improve intelligibility of the spoken utterance. However, although clipped words may degrade speech recognition, they are unlikely to affect emotion recognition. Therefore, we have not considered buffer frames in this paper.

2.2. Data-driven statistical VAD

Data-driven VAD is a statistical classifier which is trained to identify speech/non-speech frames from the acoustic features. We compare the proposed VAD with an RNN based VAD which uses traditional frame-wise features, capable of learning the dynamics of the inputs and use previous inputs adaptively for the decision of the current frame [18]. In particular, LSTM-RNN is used for their ability to model long-range dependencies between the inputs in comparison with other common data-driven VAD approaches (GMM or ANN), which do not consider any temporal dependencies. The LSTM-RNN is able to capture context information by introducing the concept of a memory cell which can be read, written and reset depending on the feature context and previous outputs. The networks used here for VAD have an input layer that takes the low-level acoustic feature vectors (RASTA-PLP with cepstral coefficients and their first order delta coefficients), multiple hidden layers, and an output layer with a single linear unit. The networks are trained as regressors to output a voicing score for every frame in the range $[-1; +1]$; $+1$ indicating voicing, -1 indicating silence or noise. Features extraction

ISBN 978-90-827970-1-5 © EURASIP 2018

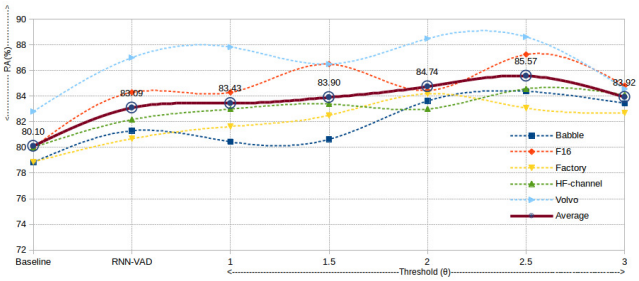

Fig. 2. Parameter tuning using data from EMO-DB

is done with openSMILE toolkit and z-normalization was applied to all features (mean 0, variance 1) [20].

3. EXPERIMENTS AND RESULTS

3.1. Databases

To test our proposed technique in noisy speech emotion recognition task, we have used 2 standard emotional databases, namely (1) Berlin emotional database (Emo-DB) [21, 22] (2) Interactive emotional dyadic motion capture database (IEMOCAP) [23, 24]. Emo-DB dataset consists of 535 emotional speech utterances, where 10 professional actors (5 male and 5 female) participated to act for 7 emotions (anger, boredom, disgust, anxiety, happy, sad and neutral). IEMOCAP database is an interactive, spontaneous, multimodal database of approximately 12 hours of audiovisual data that consists of dyadic sessions where interactions took place on improvisations or scripted scenarios. Multiple annotators annotated the sample in both categorical and dimensional space. To corrupt the clean emotional speech from the above-mentioned databases, we used noise samples from Noisex-92 database [25]. In particular, we have taken 5 different types of noise (Voice babble, Factory noise, HF radio channel, F-16 fighter-jets, and Volvo 340). FANT toolkit is used for contaminating noise to the clean speech samples for 5 SNR levels (0, 5, 10, 15, and 20dB) [26].

3.2. Parameters tuning for proposed VAD

Since the proposed VAD has the flexibility to adjust its parameters, we conducted a set of experiments to reach the optimal parameter values. First, we varied frame duration from 10ms to 40ms, with 10ms shift. We found that the best performance was obtained for 20 ms frame duration. The number of frames considered in the neighborhood for computing the average energy was also fixed empirically, and the best results were obtained when $k = 2$ was considered. Similarly, to arrive at the optimal threshold value (i.e. θ) we conducted a series of experiments and obtained emotion Recognition Accuracy (RA). Figure 2 plot the RAs for various threshold values. As a reference, we also plotted RA for the baseline and RNN-VAD. Note that the baseline results are obtained without VAD or any other noise reduction technique. To select the optimal value for θ, we calculated the average RA over all types of noise and found the best RA (i.e. 85.57%) for $\theta = 2.5$, which is shown in the Fig. 2. It is worth noting that although these parameters were tuned on Emo-DB database, the same parameters hold equally well for IEMOCAP database. Thus, these parameters are not database specific.

3.3. Emotion recognition system

Once the noisy speech is passed through VAD for noise reduction, we extract large set of high level audio features (high level descriptors (HLDs)) from the popular set of low level audio descriptors (i.e. LLDs) like Log-energies, voice probability, frequency-band energies, MFCCs, pitch, F0, ZCR, LSP, FFT, $melspec$ magnitude so that it carries more relevant information about the emotion compared to just using LLDs. Since the HLDs are statistics (up to fourth order) of LLDs overall smaller frames (20 ms) in a spoken utterance, the dimension of the acoustic features remains same (i.e. 6552) for all utterances. We have used emo_large configuration file from the openSMILE toolkit for extraction of acoustic features (HLDs from LLDs) [20]. Then the acoustic features are fed to the Support Vector Machine (SVM) classifier for recognition, and 5-fold stratified cross-validation is used in all experiments. LibSVM classifier with *linear kernel* from WEKA machine learning toolkit has been used [27].

3.4. Results and analysis

In Table 1, we put the recognition accuracies (RA in %) for the emotions classified in categorical space, 7 classes for Emo-DB and 9 classes for IEMOCAP, 5 types of noises (Babble, F-16, Factory, HF-channel, Volvo) with 5 SNR levels (0, 5, 10, 15, 20 dB). We also put RA results for data-driven VAD from openSMILE toolkit for the purpose of comparing. RA for the baseline system that does not use any VAD is also tabulated as a reference. As observed, the proposed VAD based emotion recognition system always produces superior result compared to the RNN-based VAD and the reference baseline system. For Emo-DB database, an absolute improvements in RA (in %) of 5.54 (Babble), 7.24 (F-16), 4.19 (Factory), 4.56 (HF-channel), 5.68 (Volvo) with a mean of 5.54 average over noises are observed by using our proposed VAD over the baseline. Similarly for the same database, absolute improvements in RA (in %) of 3.07 (Babble), 2.95 (F-16), 2.38 (Factory), 2.39 (HF-channel), 1.48 (Volvo) with a mean of 2.39 average over noises are observed by using our proposed VAD over the RNN-based VAD. As shown in Table 1, for IEMOCAP database, an absolute improvements in RA (in %) of 8.51 (Babble), 3.26 (F-16), 6.44 (Factory), 5.41 (HF-channel), 7.99 (Volvo) with a mean of 6.44 are observed by using our proposed VAD over the baseline. Similarly for the same database, an absolute improvements in RA (in %) of

ISBN 978-90-827970-1-5 © EURASIP 2018

Table 1. Categorical emotion (RA in %) for Emo-DB and IEMOCAP databases (5 types of noise with 5 SNR levels), Baseline: using no VAD, RNN-VAD: openSMILE VAD, Prop-VAD: Our proposed VAD)

Emo-DB																										
	Babble					F16					Factory					HF-channel					Volvo				
	No COMP	RNN VAD	NMF	P-VAD	P-VAD + NMF	No COMP	RNN VAD	NMF	P-VAD	P-VAD + NMF	No COMP	RNN VAD	NMF	P-VAD	P-VAD + NMF	No COMP	RNN VAD	NMF	P-VAD	P-VAD + NMF	No COMP	RNN VAD	NMF	P-VAD	P-VAD + NMF
0db	69.6	72.52	76.44	75.85	77.23	68.03	72.33	73.08	74.71	74.81	64.85	68.15	70.46	69.58	70.91	69.53	79.81	79.06	81.69	81.91	80.56	84.85	85.23	85	85.77
5db	73.33	78.56	77.75	81.8	80.30	77	79.81	80.56	82	82.56	72.71	78.01	76.82	77.89	78.12	76.63	80.93	84.11	84	84.56	82.42	85.16	86.04	87.17	87.95
10db	78.83	81.3	81.68	84.37	84.95	80	84.29	82.99	87.24	87.92	78.87	80.68	82.30	83.06	83.56	80	82.17	86.54	84.56	85.12	82.8	87	87.92	88.48	88.99
15db	83.14	84.11	85.60	86.8	87.12	81.49	85.79	85.58	89.29	90.12	80.56	83.73	84.29	84.12	85.0	80.74	86.35	87.28	87.93	88.0	83.73	87.2	88.41	89.42	90.15
20db	83.85	86.35	86.11	88.85	88.99	82.43	87.1	88.92	89.86	90.73	83.78	85	84.85	87.55	88.12	81.3	86.54	87.98	88.61	88.95	84.11	87.38	89.91	90.35	90.93
Mean	78.83	81.3	81.51	84.37	83.71	80	84.29	82.22	87.24	85.22	78.87	80.68	79.74	83.06	81.14	80	82.17	84.99	84.56	85.70	82.8	87	87.50	88.48	88.75
IEMOCAP																									
0db	37.05	40.12	43.86	44.61	45.12	37.86	38.12	40.00	40.08	40.32	36.78	37.16	40.16	42.21	42.71	39.21	40.32	43.97	44.71	45.01	44.07	46.12	44.79	48.09	48.93
5db	40.9	42.32	45.98	46.12	46.87	39.45	39.45	40.15	40.96	41.03	37.7	38.92	41.78	43.11	43.93	41.09	42.21	45.15	46.43	46.73	44.73	46.66	48.12	51.75	51.87
10db	44.12	47.19	50.19	52.63	53.01	40.96	40.96	42.11	44.22	44.65	39.48	42.61	43.62	45.92	46.01	42.1	42.98	46.52	47.51	48.12	46.12	50.18	52.71	54.11	54.76
15db	48.95	50.67	53.67	55.76	55.81	42.31	42.31	45.32	47.62	48.11	40.87	43.11	46.12	47.66	48.12	43.25	44.67	47.91	48.16	48.76	48.05	52.67	55.16	56.37	56.95
20db	52.87	54.33	56.12	57.82	56.12	43	43	48.92	52.19	53.51	41.9	44.65	46.95	48.92	49.13	44.33	45.11	49.12	50.37	50.98	49.33	53.9	56.25	58.75	59.10
Mean	44.12	47.19	49.96	52.63	51.38	40.96	40.96	43.3	44.22	45.42	39.48	42.61	43.72	45.92	45.98	42.1	42.98	46.53	47.51	47.92	46.12	50.18	51.40	54.11	54.32

Table 2. Dimensional emotion recognition accuracies (Correlation coefficients (CC) and Mean absolute error (MAE) for IEMOCAP database (5 types of noise with 5 SNR levels), Baseline: using no VAD, RNN-VAD: openSMILE VAD, Prop-VAD: Our proposed VAD

		Babble			F16			Factory			HF-channel			Volvo		
		Baseline	RNN-VAD	Prop-VAD	Baseline	RNN-VAD	Prop-VAD	Baseline	RNN-VAD	Prop-VAD	Baseline	RNN-VAD	Prop-VAD	Baseline	RNN-VAD	Prop-VAD
0DB	**CC**	0.56	0.57	**0.58**	0.56	0.56	**0.57**	0.54	0.54	0.54	0.57	0.57	**0.58**	0.58	0.58	0.58
	MAE	0.54	**0.56**	0.55	0.54	0.54	0.54	0.55	0.55	**0.56**	0.53	0.53	0.53	0.53	0.53	0.53
5DB	**CC**	0.57	0.57	**0.58**	0.57	0.58	**0.59**	0.56	0.56	0.56	0.58	0.58	**0.59**	0.59	0.59	0.59
	MAE	0.53	0.53	**0.54**	**0.53**	**0.53**	0.52	**0.54**	**0.54**	0.53	**0.53**	0.52	0.52	0.52	0.52	0.52
10DB	**CC**	0.58	0.58	**0.59**	0.58	0.58	**0.6**	0.57	0.58	**0.59**	0.71	0.71	0.71	0.59	0.59	**0.6**
	MAE	0.52	0.52	0.52	**0.53**	**0.53**	0.52	0.53	0.53	0.53	0.4	0.4	0.4	0.52	0.52	0.52
15DB	**CC**	0.59	0.6	**0.61**	0.58	0.59	**0.6**	0.58	0.59	**0.59**	0.71	0.71	**0.72**	0.6	0.61	**0.61**
	MAE	**0.53**	0.52	0.51	**0.53**	0.52	0.51	**0.53**	**0.53**	0.52	**0.4**	0.39	0.39	**0.52**	**0.52**	0.51
20DB	**CC**	0.59	0.61	**0.62**	0.59	0.59	**0.61**	0.59	0.59	**0.6**	0.71	0.72	**0.73**	0.6	0.61	**0.62**
	MAE	**0.52**	0.51	0.5	**0.52**	0.51	0.5	**0.53**	0.52	0.52	**0.39**	**0.39**	0.38	**0.52**	0.51	0.5

5.44 (Babble), 2.6 (F-16), 3.31 (Factory), 4.53 (HF-channel), 3.93 (Volvo) with a mean of 3.93 average over noises are observed by using our proposed VAD over the RNN-based VAD.

In Table 2, we show the performance of mean absolute error (MAE) and correlation coefficients (CC) for the emotions classified in dimensional space (3 dimensions: arousal, valence and dominance) for IEMOCAP database, 5 types of noises (Babble, F-16, Factory, HF-channel, Volvo) with 5 SNR levels (0, 5, 10, 15, 20 dB). However, because of the space constraint, instead of putting all the performance values individually for all 3 dimensions, we tabulated the average of performance metrics (MAE and CC) for 3 dimensions. We have also put results for data-driven VAD from openSMILE toolkit for the purpose of comparing. Performance for the baseline system that does not use any VAD is also tabulated as a reference. Again, similar to the Table 1, we observed that our proposed VAD based emotion recognition system, in general, produces superior result compared to the RNN-based VAD and the reference baseline system.

In addition to all above-mentioned experiments for noisy emotional speech, we also conducted experiments for clean speech using our proposed VAD to check its contribution. It is worth mentioning that our proposed VAD based system is able to achieve absolute 1% improvement in RA over the baseline result.

4. CONCLUSIONS

In this paper, we propose a novel and useful front-end voice activity detector (VAD)-based unsupervised method to select the frames with relatively better SNR in the spoken utterances for robust emotion recognition in noisy speech. We experimentally validated the usefulness of our proposed VAD by conducting extensive experimentations on two standard emotional databases contaminated with 5 different types of noise (with 5 SNR levels). The proposed technique outperforms the RNN-based VAD across all types and levels of noises, and for both the databases. This is all the more interesting, the proposed VAD is considerably simpler than the RNN based VAD. Moreover, we got better performance, overall, in all experiments to classify emotions both in the categorical and the dimensional spaces. This study indicates that frame selection has an important contribution to emotion recognition from speech.

ISBN 978-90-827970-1-5 © EURASIP 2018

5. REFERENCES

[1] M. El Ayadi, Mohamed S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, pp. 572–587, 2011.

[2] T. Vogt and E. Andre, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *ICME*, 2005, pp. 474–477.

[3] E. Mower, M. Mataric, and S. S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 5, pp. 1057–1070, 2011.

[4] B. Schuller, D. Arsi, F. Wallhoff, and G. Rigoll, "Emotion recognition in the noise applying large acoustic feature sets," in *Speech Prosody*, 2006.

[5] B. Schuller, D. Seppi, A. Batliner, A. K. Maier, and S. Steidl, "Towards more reality in the recognition of emotional speech," in *ICASSP*, 2007, pp. 941–944.

[6] M. You, C. Chen, J. Bu, J. Liu, and J. Tao, "Emotion recognition from noisy speech," in *2006 IEEE ICME*, July 2006, pp. 1653–1656.

[7] 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 *IIH-MSP*, Nov 2007, vol. 1, pp. 23–26.

[8] Lukasz Juszkiewicz, *Improving Noise Robustness of Speech Emotion Recognition System*, pp. 223–232, Springer International Publishing, Cham, 2014.

[9] C. Huang, G. Chen, Hua Yu, Y. Bao, and Li Zhao, "Speech emotion recognition under white noise," *Archives of Acoustics*, vol. 38, no. 4, pp. 457–463, 2013.

[10] Jouni Pohjalainen, Fabien Fabien Ringeval, Zixing Zhang, and Björn Schuller, "Spectral and cepstral audio noise reduction techniques in speech emotion recognition," in *ACM on Multimedia Conference*, USA, 2016, MM '16, pp. 670–674, ACM.

[11] A. Tawari and M. M. Trivedi, "Speech emotion analysis in noisy real-world environment," in *ICPR*, Aug 2010, pp. 4605–4608.

[12] F. Chenchah and Z. Lachiri, "Speech emotion recognition in noisy environment," in *ATSIP*, March 2016, pp. 788–792.

[13] Xiaoming Zhao, Shiqing Zhang, and Bicheng Lei, "Robust emotion recognition in noisy speech via sparse representation," *Neural Computing and Applications*, vol. 24, no. 7, pp. 1539–1553, Jun 2014.

[14] Rajib Rana, "Emotion classification from noisy speech - A deep learning approach," *CoRR*, vol. abs/1603.05901, 2016.

[15] M. Aiswarya, D. Pravena, and D. Govind, "Identifying issues in estimating parameters from speech under lombard effect," in *SIRS*, 2017.

[16] I. Luengo, E. Navas, and I. Hernez, "Combining spectral and prosodic information for emotion recognition in the interspeech 2009 emotion challenge," in *INTERSPEECH*, 2009.

[17] M. Kockmann, L. Burget, and J. Honza ernock, "Brno university of technology system for interspeech 2009 emotion challenge," in *INTERSPEECH*, 2009.

[18] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies," in *IEEE ICASSP*, May 2013, pp. 483–487.

[19] D. A. Reynolds, *A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification*, Ph.D. thesis, Georgia Institute of Technology, 9 1992.

[20] "opensmile," `http://www.audeering.com/research/opensmile`.

[21] "Berlin database of emotional speech," `http://www.emodb.bilderbar.info/`.

[22] F. Burkhardt, A. Paeschke, M. Rolfes, Walter F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *INTERSPEECH 2005, Portugal, September 4-8*, pp. 1517–1520.

[23] C. Busso, M. Bulut, Chi-Chun Lee, Abe Kazemzadeh, E. Mower, S. Kim, Jeannette N. Chang, S. Lee, and S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[24] "Interactive emotional dyadic motion capture database," `http://sail.usc.edu/iemocap/`.

[25] "Noisex-92 database," `http://spib.rice.edu/spib/select_noise.html`.

[26] "Fant - filtering and noise adding tool," `http://dnt.kr.hsnr.de/download/`.

[27] "Weka machine learning toolkit," `http://www.cs.waikato.ac.nz/ml/weka/`.

ISBN 978-90-827970-1-5 © EURASIP 2018