INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Speaker Dependency Analysis, Audiovisual Fusion Cues and A Multimodal BLSTM for Conversational Engagement Recognition

Yuyun Huang, Emer Gilmartin, Nick Campbell

Speech Communication Lab,
School of Computer Science and Statistics, Trinity College Dublin
huangyu@tcd.ie, gilmare@tcd.ie, nick@tcd.ie

Abstract

Conversational engagement is a multimodal phenomenon and an essential cue to assess both human-human and human-robot communication. Speaker-dependent and speaker-independent scenarios were addressed in our engagement study. Hand-crafted audio-visual features were used. Fixed window sizes for feature fusion method were analysed. Novel dynamic window size selection and multimodal bi-directional long short term memory (Multimodal BLSTM) approaches were proposed and evaluated for engagement level recognition.

Index Terms: Conversational Engagement Level Recognition, Multimodal BLSTM, Speaker (in)dependent analysis

1. Introduction

In human interaction both body behaviour and voice activity reflect how participants are engaging in the conversation. Research on social engagement is attracting considerable attention in the field of multimodal human speech communication as engagement is a key factor in inferring the quality of a human-human or human-machine dialogue. As a result, engagement assessment has several potential applications, in the dialogue, social, and health fields. More engaging applications could improve both user satisfaction and task success. To build these applications, automatic engagement assessment will be vital.

Social engagement is defined as *The process by which two (or more) participants establish, maintain and end their perceived connection. This process includes: initial contact, negotiating a collaboration, checking that other is still taking part in the interaction, evaluating whether to stay involved and deciding when to end the connection* [1]. Non-verbal audio-visual cues play a central rule in engagement recognition. In our work, we do not treat the beginning or ending stages, so our engagement annotation and processing are restricted to the context of an already established conversation.

Engagement detection is a relatively new concept – there is a shortage of freely available annotated datasets and robust prediction models are still being investigated. In general, research on automatic engagement prediction has been based on data-driven statistical approaches using non-verbal visual and audio cues. Engagement can be annotated as binary or scalar levels [2]. Gaze has been investigated as a feature in Bednarik et al. (2012) [3] and Zhao et al. (2016) [4], while Oertel's (2013) manually annotated gaze data obtained an accuracy of 71% in a 4-class classification, using fixed time windows [5]. Gaze-related visual features have been used to analyse the relationship between individual and group engagement in an eight-party multimodal corpus [5]. However, the accuracy of automatic gaze extraction is not reliable and thus may not suitable to use for automatic predict engagement. Salam et al. (2016) used head pose tracking and skeleton fitting based visual features to analyse engagement and its relationship to personality [6]. Audio cues such as fundamental frequency (F0), voice quality, jitter, and shimmer were used in Kim's (2016) work on engagement detection for children [7]. Hsiao et al. (2012) used mel-frequency cepstral coefficients (MFCCs) as low level features to compute higher level features to predict engagement levels [8]. Huang et al. (2016) used audio and visual hand-crafted features to train support vector machines (SVM) and convolutional neural networks (CNN) on video alone; the visual features included both appearance (using local binary patterns) and geometric features computed from the location of facial landmarks. Pitch level, MFCCs and loudness were used as audio cues. To model audio features, temporal dynamic, distance and angle features were computed from the amplitude in a fixed window size [9]. In contrast to the fixed window approach, unimodal audio low level features were used by Yu (2004) to recognise speech emotions using SVM and the obtained speech emotion states were used as input to a higher level hidden Markov models (HMM) to model social engagement's temporal and interactive aspects dynamically [10].

Although speaker independent and dependent factors in engagement were not differentiated in previous work, we think much automatic engagement recognition thus far may be quite speaker-dependent. In this work, we begin to fill this gap by comparing the depth of speaker dependency clearly. We address the question of using fixed or dynamic windows in our models. We propose a new dynamic window size approach using head pose signals for engagement level recognition, and empirically compare different fixed window size performance in the same conditions. We also propose a novel multimodal bi-directional Long short term memory (LSTM) for engagement recognition by using sequential features from different modalities.

Below, we analyse dyadic conversational engagement speaker dependency, investigate window size effects, and apply and propose deep learning approaches to model engagement sequentially. We also introduce a new facial energy image visual feature, and add previously unexplored visual features for engagement prediction including Histograms of Oriented Gradient (HOG) and Gabor filters.

2. Methodology

Figure 1 shows the flowchart of feature-based 'shallow' learning using a support vector machine. The raw data comprised video recording and their stripped audio. The raw data was used to generate auditory and visual features and then processed for engagement level classification.

The real time OpenFace toolkit [11] and Chehra scripts [12] were used to extract aligned frontal face images, facial landmarks, facial action units (FACs), and head pitch, roll and yaw, from the video data.

http://dx.doi.org/10.21437/Interspeech.2017-1496

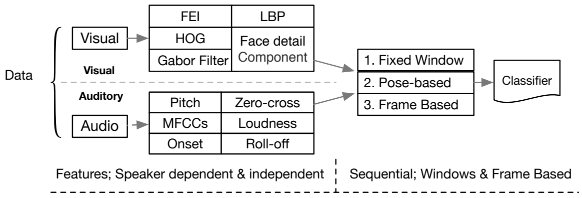

Figure 1: *Flow Chart of Feature Based Learning*

2.1. Visual Cues

2.1.1. Local binary patterns + PCA

Circular local binary patterns (CLBP) were used to compute rotation invariant texture features. Ahonen et al. [13] introduced CLBP by dividing images into $m \times n$ sub-areas and concatenating the local binary pattern histograms to form CLBP operators. We followed Ahonen's method to compute LBP operators. The input image was divided into 4×4 sub-fields. For a pixel point (x_c, y_c), the coordinates of its circularly symmetric neighbours can be computed by $x_p = x_c + Rcos(\frac{2\pi p}{P})$, $y_p = y_c - Rsin(\frac{2\pi p}{P})$, where R is the radius of the circle and P is the number of samples. Figure 2 shows the visualisation of LBP feature. The input facial image was resized to 112×112 pixels, the number of CLBP operators is $4 \times 4 \times (2^8) = 4096$, Principal Component Analysis (PCA) was used for dimensionality reduction.

2.1.2. Gabor filter + PCA

The Gabor transform is a windowed Fourier transform or short time Fourier transform. The orientation of Gabor filters and their frequency representations are similar to the human visual system and can be used for texture representation or recognition. Gabor filter based texture features are invariant to scale, rotation and translation. They are also insensitive to illumination changes. In our work, we used Gabor filters with 5 frequency scales with 8 orientations. The dimensions of the feature vector were $112 \times 112 \times 5 \times 8 = 501760$. We used a downsampling approach in the Matlab signal processing toolkit to decrease feature vector dimensionality by a factor of 16, resulting in 1960 features. PCA was used for further dimensionality reduction.

2.1.3. HOG + PCA

Histograms of Oriented Gradient (HOG) descriptors were proposed for human detection by Dalal and Triggs [14], and have been used widely in feature representation and recognition tasks. HOG's basis premise is that the appearance and shape of a local object can be described well by the distribution of local intensity gradients or edge directions. Figure 2 shows the visualisation of HOG features. We used the Matlab build-in HOG feature extraction function and set cell size as (5×5) with default values of block size, block overlap and number of orientation histogram bins in our experiments. The input image was resized to 60×60, obtaining a HOG feature length of 4356. Further PCA was used to reduce dimensionality.

2.1.4. Shape and Geometry: Facial Energy Image (FEI)

Considering speaker dependency issues, normalised shape and geometry features may be insensitive to speaker-specificity compared to texture features. In this work we estimated the silhouette of facial components based of the shape of facial landmark locations – face, mouth, nose, eyes and eye brows. We first identified regions of interest from facial landmarks and set pixel values inside and outside these polygons with binary codes 0 or 1 to form facial mask images. In order to model engagement dynamically, an energy image was computed by calculating the average of total selected facial mask images sequentially. $FEI(x, y) = \frac{1}{N}\sum_{t=1}^{N} M_t(x, y)$, where N is the number of total mask images, M represents the facial mask images and (x,y) are the coordinates of the 2D image. Figure 2 illustrates the FEI generation from time sequential binary silhouette images. PCA was also used to reduce feature size.

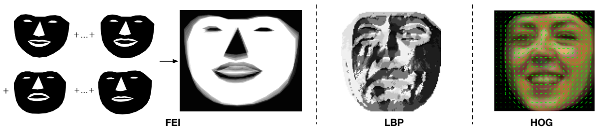

Figure 2: *Visualisation of FEI, LBP and HOG*

2.2. Auditory Cues

Auditory features were extracted using OpenSmile [15] and Essentia [16]. Audio files were re-sampled at 22050 Hz, and set the same hop length to 512.

Loudness, intensity: Loudness was estimated using OpenSmile, treating loudness (l) as the normalised intensity (I) raised to a power of 0.3, $l = (I/I_0)^{0.3}$ [15].

Pitch, Fundamental Frequency: we used the implementation based on the Summation of the Residual Harmonics method [17], which is robust to noise.

Mel-frequency cepstral coefficients (MFCCs): A filterbank of 40 bands ranging from 0Hz to 11000Hz was used to compute mel-frequency cepstrum coefficients. We extracted 13 coefficients including log of the frame energy. MFCC+delta was used as features.

Zero-crossing rate, roll-off frequency, onset strength envelope: the first two spectral features and one rhythm feature were also extracted.

We also applied a shape and angle approach [9] on the curves to calculate features from extracted auditory data. The locations of four descriptive points showed in Figure 3 (c) were used to compute shape and angle features in a sub-segment (window) of period t, and a sliding window with a step size of $t/2$. Two points in each half t with the largest and smallest values were selected. Visual features were combined together and a correlation feature selection (CFS) approach was used [18] to select useful features, then auditory and visual features were fused sequentially.

2.3. Dynamic and Sequential Modelling: Windows

Window based approach: we used static and dynamic approaches here. In the static approach we computed the average within a fixed window size of time t. In the dynamic approach, we generated a set of different window sizes based on head motion computed by head pitch and yaw values, as they may indicate dynamic segments such as nodding. From t_{n-1} to t_n to pitch or yaw value changes ±20 degree as one window size of $(t_n - t_{n-1})$, while with maximum time limit to 1.0 second. Figure 3(a) shows pitch, yaw and roll values on current frame, in (b) illustrates yaw changes 20 degrees.

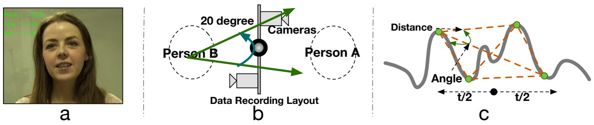

Figure 3: *Head Motion and Curve Angle/distance Feature*

2.4. Dynamic Sequential Modelling: Multimodal-BLSTM

LSTM: To model the engagement signal temporally and learn it over longer or variable time windows, a deep learning LSTM approach was investigated. Visual and audio features were used as feature-level fused multimodal cues to feed the LSTM. An LSTM network comprises many memory cells. The memory unit inside the memory cell elects to store or erase (diminish) past data to ensure the LSTM has the ability to model long term dependencies in sequential signals and overcome the recurrent neural network's (RNN) gradient vanishing problem. The remember and forgot selection is modelled using the operation of multiplicative gating. The following equations and Figure 4 describes signal model LSTM.

$$g_t = tan(W_{xg} * X_t + W_{hg} * h_{t-1} + b_g), \tag{1}$$

$$i_t = \varphi(W_{xi} * X_t + W_{hi} * h_{t-1} + b_i), \tag{2}$$

$$f_t = \varphi(W_{xf} * X_t + W_{hf} * h_{t-1} + b_f), \tag{3}$$

$$o_t = \varphi(W_{xo} * X_t + W_{ho} * h_{t-1} + b_o), \tag{4}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot g_t, \tag{5}$$

$$h_t = o_t \odot \varphi(C_t), \tag{6}$$

$$y_t = softmax(W_y * h_t). \tag{7}$$

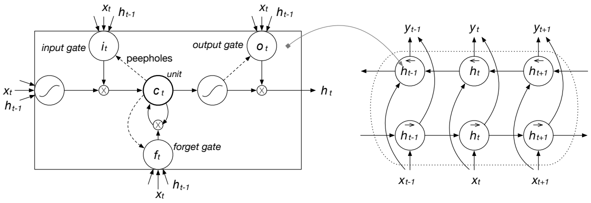

Figure 4: *LSTM Cell and BRNN*

2.4.1. BLSTM

BLSTM expands regular LSTM to look at both forward hidden and backward hidden sequences. Combining Bidirectional-RNN and LSTM by using LSTM for the hidden layers to form BLSTM [19] as illustrated in Figure 4. Equation h_t can be modified to look in both directions $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$.

$$y_t = softmax(W_{\overrightarrow{h_t}y} * \overrightarrow{h_t} + W_{\overleftarrow{h_t}y} * \overleftarrow{h_t}). \tag{8}$$

2.4.2. Multimodal Modal-level Fusion

We used a 'full cross-modal weight sharing' solution among different modalities from Ren et al. (2016) [20] and extended the sharing weights strategy to a bidirectional-LSTM for our multimodal engagement task. The selectively sharing strategy in the forward pass is explained in following equations and illustrated in Figure 5. Memory units are not shared inside memory cells, weights with superscript s (represents each modality) are not shared to enable more focused learning objectives and make it easier to learn within the modality s. Other weights marked with light colour in the equations and Figure 5 are shared and play a central role to share across modalities.

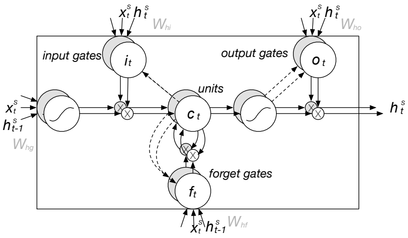

Figure 5: *Multimodal Weight Sharing*

$$g^s_t = tan(W^s_{xg} * X^s_t + W_{hg} * h^s_{t-1} + b^s_g), \tag{9}$$

$$i^s_t = \varphi(W^s_{xi} * X^s_t + W_{hi} * h^s_{t-1} + b^s_i), \tag{10}$$

$$f^s_t = \varphi(W^s_{xf} * X^s_t + W_{hf} * h^s_{t-1} + b^s_f), \tag{11}$$

$$o^s_t = \varphi(W^s_{xo} * X^s_t + W_{ho} * h^s_{t-1} + b^s_o), \tag{12}$$

$$C^s_t = f^s_t \odot C^s_{t-1} + i^s_t \odot g^s_t, \tag{13}$$

$$h^s_t = o^s_t \odot \varphi(C^s_t), \tag{14}$$

$$y^s_t = softmax(W_y * h^s_t) \quad or \tag{15}$$

$$y^s_t = softmax(W^s_y * h^s_t), \quad s = 1 \quad to \quad n \tag{16}$$

3. Experiments and Evaluations

3.1. Data Collection and Annotation

To analyse human-human dyadic face-to-face conversation, an audio-video corpus of 12 dyadic conversations were recorded with length ranging from 8 minutes to 22 minutes (16 speakers). We added more data from the Cardiff Conversational Database (CCDF) [21] of short conversations with duration of around 3 to 5 minutes (6 speakers). Data were annotated by two engagement annotators with interannotator agreement or kappa coefficient of 0.81. We analysed 3 levels of engagement – natural, engaged, highly engaged – in this work. In total, 17 speakers' data were selected.

In section 3.2 and 3.3 an SVM with RBF kernel was used and a grid-search approach was used to find the best C and gamma parameters, while a learning curve plots training and validation accuracy was used to avoid any parameter causing over-fitting or under-fitting issues. Validation procedures are described in each section.

3.2. Speaker-independent and Speaker-dependent

To compare speaker dependent and independent scenarios, the following four configurations were applied to compare the depth of engagement speaker dependency. These configurations were inspired by Rybka and Janicki's work [22].

1. Speaker Independent Full (SIF): A Leave-one-speaker-out validation strategy was used. To keep the training and test data balanced, we randomly duplicated instances in both training and test sets for each engagement class. The random instance duplication was repeated 10 times and the results were averaged, based on an approach used in [23] for emotion recognition. Each speaker was used once in the validation set, for a total $C^1_{17} = 17$ validation iterations, then the results of all the

iterations were averaged.
2. Speaker Independent SI-(train, test): The SI-(train,test) is similar to the SIF, where (train,test) represents the number of speakers in training and test sets. In each case, the number of speakers in the test set varies as following: (16,1) equal to SFI, (15,2), (13,4), (11,6), (8,9), (5,12). The amount of instances in test class for each speaker's each class was limited to be balanced. Results were also the average of iterations, e.g. for (15,2), calculate mean of $C_{17}^2 = 136$ iterations.
3. Speaker Dependent A SDA-(train, test): As with SDA-(train,test), assuming that there are a total of k instances in a class's test set for given *speakers* in test set, n instances were moved to the training set and the remaining (k-n) instances were used as test data, while n instances were randomly removed from the training set to keep the size the same, in order to adapt the classier for the test given *speakers*.
4. Speaker Dependent B SDB-(train, test): Similar to SDA-(train,test), while k instances were selected from only *one* speaker in the test set regardless of number of speakers in the test set.

We used 700 instances for each class of each speaker, total $700 \times 3 \times 16 = 33600$ in the training set and $700 \times 3 \times 1 = 2100$ in the test set. For each configuration, the results are shown in Table 1.

Table 1: *Recognition Rates of Speaker (In)Dependent Cases*

(tr,te)	(16,1)	(15,2)	(13,4)	(11,6)	(8,9)	(5,12)
SI	0.502	0.491	0.482	0.473	0.454	0.420
SDA	0.800	0.805	0.780	0.804	0.812	0.811
SDB	0.800	0.604	0.472	0.467	0.437	0.412

To keep conditions uniform, we used raw fusion features without any normalisation. SIF achieved a accuracy of 50.2%, compared to the chance level of 33.3% (1/3). Leave-one-out semi-speaker dependent got a accuracy of 80.0%. The results show that social engagement speaker dependency needs to considered. For speaker dependent scenarios such as personal mobile devices, each speaker needs to be used in training the classier in order to get reliable speaker-specific engagement recognition results, as shown by the difference in SDA and SDB results. For the speaker independent scenario, fewer speakers in the training set decreased the prediction accuracy, therefore large corpora including many speakers are also an essential component for this work, even though they are costly.

To better analyse the speaker dependency issue, we used a simple normalisation approach for comparison. It was adapted from z-standardization [24]. Each feature element was reduced by the mean of natural engagement and divided by the natural engagement standard deviation as in equation 17. The ideal normalisation is to use each speaker's natural segments. As these segments are unknown in testing, we used mean and standard deviation values from natural segments in the training set. After normalisation, SIF obtained a slight higher accuracy of 0.521 (2% increase), and Leave-one-out SD also had a higher result of 0.8329 (3.2% increase). In conclusion, the speaker normalisation process works although the improvement is not remarkable.

$$f^s_{Normalised} = \frac{f^s - \mu^{TrainSet}_{Neutral}}{\sigma^{TrainSet}_{Neutral}} \tag{17}$$

3.3. Fixed and Dynamic windows

To evaluate the sliding window approach introduced in section 2.3 to model engagement dynamics, we used several different fixed window sizes. The dynamic window size was limited in the range 0.2 to 1.0 second. We first got features applied at different window sizes then selected and randomly duplicated to 700 instances of each class of each speaker for each size case. The raw features were normalised before applying the window. Leave-one-speaker-out validation strategies (*SD*: SDA-(16,1) & *SIF*) described in the previous section were used.

Table 2: *Accuracies of Different Window Size Cases*

Size	FB	0.2s	0.4s	0.7s	1.00s	DW
SIF	0.521	0.527	0.536	0.510	0.513	0.538
SD	0.832	0.833	0.838	0.831	0.825	0.842

As shown in Table 2, for a fixed window size approach, window size within 1 second has no slope changes, and 0.4 second windows obtained best accuracy. For dynamic window (DW) size approach, we obtained a higher accuracy of 0.842. Window based approaches can obtain higher results compared to frame based feature processing. Careful tuning of window size is needed for the purpose of a higher prediction rate.

3.4. MBLSTM

The same number of normalised instances (33600 for train and 2100 for test) were used to evaluate deep learning approaches. A baseline (BL) approach of non-normalized audio unimodal with SVM was tested as comparison. Results of normalised fusion cues with best accuracy using window based approach (FSVM), Feature-level fusion to feed LSTM and BLSTM, and modal-level multimodal (MBLSTM) are shown in Table 3.

Table 3: *Accuracies of Different Approaches*

	BL	FSVM	LSTM	BLSTM	MBLSTM
SIF	0.420	0.538	0.583	0.603	0.617
SD	0.683	0.842	0.862	0.870	0.883

4. Conclusions

Speaker dependency is an important factor which needs to be considered in engagement recognition and measurement. Both speaker-specific and speaker-independent models are useful for engagement recognition depending on the application scenarios. Results show improvement in specified validation strategies over previous work. Speaker-independent analysis requires more work, simple speaker normalisation (z-standardization) improves the prediction rate, while an unsupervised normalisation approach is essential as natural segments are unknown in test sets, meanwhile some speaker non-sensitive features will also be an advantage to deal with the issue. The window based approach was expected to obtain higher rates, but this may depend on the annotation scheme used, as computing average of long annotated segments may discard much meaningful information – for this case a dynamic window will be more suitable. Multimodal BLSTM obtained the best accuracy, demonstrating the value of further research into deep learning approaches in sequential engagement level recognition research.

5. References

[1] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artificial Intelligence*, vol. 166, no. 12, pp. 140 – 164, 2005.

[2] C. Oertel, F. Cummins, J. Edlund, P. Wagner, and N. Campbell, "D64: A corpus of richly recorded conversational interaction," *Journal on Multimodal User Interfaces*, vol. 7, no. 1-2, pp. 19–28, 2013.

[3] R. Bednarik, S. Eivazi, and M. Hradis, "Gaze and conversational engagement in multiparty video conversation: An annotation scheme and classification of high and low levels of engagement," in *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction*, ser. Gaze-In '12. New York, NY, USA: ACM, 2012, pp. 10:1–10:6.

[4] R. Zhao, T. Sinha, A. W. Black, and J. Cassell, "Automatic recognition of conversational strategies in the service of a socially-aware dialog system," in *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2016, p. 381.

[5] C. Oertel and G. Salvi, "A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 99–106.

[6] H. Salam, O. Celiktutan, I. Hupont, H. Gunes, and M. Chetouani, "Fully automatic analysis of engagement and its relationship to personality in human-robot interactions," *IEEE Access*, 2016.

[7] J. Kim, K. Truong, and V. Evers, "Automatic detection of children's engagement using non-verbal features and ordinal learning," in *Workshop on Child Computer Interaction*, pp. 29–34.

[8] J. C.-y. Hsiao, W.-r. Jih, and J. Y.-j. Hsu, "Recognizing continuous social engagement level in dyadic conversation by using turn-taking and speech emotion patterns," in *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[9] Y. Huang, E. Gilmartin, and N. Campbell, "Conversational engagement recognition using auditory and visual cues," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, 2016, pp. 590–594.

[10] C. Yu, P. M. Aoki, and A. Woodruff, "Detecting user engagement in everyday conversations," *arXiv preprint cs/0410027*, 2004.

[11] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016, pp. 1–10.

[12] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1859–1866.

[13] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *European conference on computer vision*. Springer, 2004, pp. 469–481.

[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[15] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[16] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. R. Zapata, and X. Serra, "Essentia: An audio analysis library for music information retrieval." Citeseer.

[17] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[18] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.

[19] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005.

[20] J. Ren, Y. Hu, Y.-W. Tai, C. Wang, L. Xu, W. Sun, and Q. Yan, "Look, listen and learna multimodal lstm for speaker identification," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[21] A. J. Aubrey, D. Marshall, P. L. Rosin, J. Vandeventer, D. W. Cunningham, and C. Wallraven, "Cardiff conversation database (ccdb): A database of natural dyadic conversations," in *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, ser. CVPRW '13. Washington, DC, USA: IEEE Computer Society, 2013, pp. 277–282.

[22] J. Rybka and A. Janicki, "Comparison of speaker dependent and speaker independent emotion recognition," *International Journal of Applied Mathematics and Computer Science*, vol. 23, no. 4, pp. 797–808, 2013.

[23] C. Busso, S. Mariooryad, A. Metallinou, and S. Narayanan, "Iterative feature normalization scheme for automatic emotion detection from speech," *IEEE transactions on Affective computing*, vol. 4, no. 4, pp. 386–397, 2013.

[24] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, March 2005.