Objective Study of the Performance Degradation in Emotion Recognition through the AMR-WB+ Codec

Aaron Albin, Elliot Moore

Georgia Institute of Technology

aalbin3@gatech.edu, em80@gatech.edu

10.21437/Interspeech.2015-328

Abstract

Research in speech emotion recognition often involves features that are extracted in lab settings or scenarios where speech quality is high. However, a great deal of communication occurs through speech codecs, which alters the speech signal and features extracted from it. The purpose of this study is to report on the performance degradation in emotion recognition systems when speech is passed through a codec and to provide insight on features that are affected in relation to their relevance in emotion classification. Using two emotional databases and the AMR-WB+ codec, features that are the most and least significantly affected by the codec are investigated and classifier performances are compared among them in multiple experiments. The results show that clean-trained classifiers drop significantly in accuracy on codec speech, and vice versa for codec-trained classifiers on clean speech in a full feature set task. However, using an intersection feature set between two databases that is resilient to the codec process can provide comparable performance for clean and codec-trained classifiers on either type of speech. The results suggest that these sets of features seem to capture more relevant information about emotion classes, since the perception of emotion should not be altered by a codec.

Index Terms: emotion recognition, computational paralinguistics, speech codec, adaptive multi-rate wideband plus codec

1. Introduction

Many applications of emotion recognition have been proposed and investigated, such as in human and robotic interaction[1, 2, 3], media retrieval [4], health monitoring for depression [5], and call centers [6]. Much research in speech emotion recognition involves models designed to work on speech features extracted in lab settings that ensure the speech signal is preserved as closely as possible. However, a great deal of communication occurs that requires speech to pass through a codec such as in cell phones, teleconferencing, and voice over IP; this is only expected to grow each year, especially as mobile data demands increase [7]. The Extended Adaptive Multi-Rate Wideband (AMR-WB+) codec is one of the latest types explored here; it is designed for applications in content streaming, download services, and others and is an extension of the AMR-WB mode, which has already seen adoption by telecom companies [8, 9]. While monitoring emotion in speech covered through these media is necessary for building real-world emotion recognition models, there is little research addressing performance loss in codec speech directly. The work in this paper represents preliminary results for a larger study to investigate the impact of codecs on emotion classifiers trained in "clean" (i.e., on speech prior to being passed through the codec). This study aims to report on the performance degradation in emotion recognition systems when speech is passed through a codec and provide insight on features most or least affected by it.

2. Background

Many speech-based emotion recognition studies consider cepstral, pitch, intensity, and voice quality features [10]. Research has shown that those calculated as global statistics of all features extracted from an utterance seem to outperform local ones [11]. While there has not been much research on emotion recognition performance through codecs, there has been studies of speech and speaker recognition performance. Some features, like pitch contours, are robust to the GSM codec [15]. In [12], researchers attempted to recover some of the performance loss from the G.729 codec using codec parameters directly as well as through feature modification. Other research involved analysis in compressed bit stream domains [13, 14]. With the growth of emotion recognition and with applications involving speech processed through codecs, it is important to understand what codecs do to the original waveform.

2.1. Description of the AMR-WB+ Codec

The AMR-WB+ codec uses a hybrid of algebraic code-excited linear prediction (ACELP)[16] and transform coded excitation (TCX)[17]. Several processes in both ACELP and TCX could influence features extracted from emotional speech. In ACELP based codecs, indices from codebooks of vectors are transmitted in place of the residual of the speech. The codevector is chosen through an analysis-by-synthesis procedure to give the minimum mean squared error between synthetic and actual speech frames. An adaptive codebook is used to represent pitch periodicity in the excitation, which is subtracted from the target signal. An algebraic codebook then uses this result; it is constructed algorithmically in such a way that there are only at most four pulses per speech subframe. Hence any features that might be described by the excitation of the model of speech production would likely be altered. Quantization is applied throughout codec process on the target signal, LP coefficients, and gains for the codebooks. Additionally, there are several filters involved in the encoding and decoding process including: a pre-processing high pass filter, asymmetrical windowing, a perceptual weighting filter, an adaptive postfilter and tilt compensation filter, and adaptive gain control.

Clearly, features extracted from codec speech side would be affected. Thus, it is interesting to know whether classifiers training on clean speech suffer a performance loss on codec speech and vice versa. Additionally, there ought to be some features that work well on both types of speech, considering that perception of speech should not change with the codec.

3. Methodology

To investigate the impact of codec processing on speech features, two databases of labeled emotional speech are chosen. Features are extracted from both clean and codec versions. Experiments are conducted to compare clean and codec test sets.

3.1. Databases

Two databases, the Berlin Database of Emotional Speech (BDES) [18] and USC Interactive emotional dyadic motion capture (IEMOCAP) database [19] are chosen. The BDES recordings have 10 speakers (equal male and female), who produce a total of 535 German utterances in 7 different emotional states: angry, happy, feared, sad, disgusted, bored and neutral. The IEMOCAP database contains audiovisual and motion capture data annotated with categorical and dimensional labels. Only speech and categorical labels are used. Ten actors (equal male and female) perform improvised or scripted dialogue that elicit emotions. As in other work [20, 21], happiness and excitement categories are merged to balance data of different class labels. Only labels of happy, angry, sad, and neutral are used for a total of 5536 utterances; each class has approximately equal samples.

3.2. Codec Parameters and Feature Extraction

The AMR-WB+ Codec software used [22] is capable of operating in both AMR-WB and AMR-WB+. The lowest bitrates are selected for each, 6kbps and 24kbps which allow for bandwidths of 7.2kHz and 16kHz respectively. Voice activity detection was disabled and no frame erasures are simulated.

Feature extraction is performed with openSmile [23] using the feature set from the INTERSPEECH 2010 Paralinguistic Challenge [24]. There are ten types: loudness, 15 MFCCs, log power of eight mel frequency bands, eight line spectral pair frequencies, smoothed f0 contour, envelope of f0 contour, voicing probability of f0 candidate, local jitter, differential jitter, and shimmer. Each of these has corresponding delta coefficients appended and 21 functionals such as means, extrema positions, higher order moments, quartile percentages, and linear regression parameters for a total of 1582 features. Feature extraction was performed for both the BDES and IEMOCAP databases on the clean and codec speech for rate 6 and rate 24.

After processing speech through the codec, for each database, 1582 t-tests are performed for the features between the clean and codec sets to see which had statistically significant differences between them, with a p-value of .05 as a threshold. In future work, Wilcoxon signed-rank tests might be more appropriate since there is no assumption about normality of the data; however, t-tests are still robust to non normal data with large sample sizes. To account for differences in databases, the intersection set of features between IEMOCAP and BDES is used and denoted as SIG (i.e. common features between databases with statistically significant differences between clean and codec speech). Similarly the intersection of features with p-values greater than .05 is denoted NONSIG. SIG and NONSIG intersection sets are made for both bitrates. Table 1 shows the number of features resulting from the intersections. Intuitively, since SIG features are significantly altered by the codec, they might cause noticeable differences in classifier performance between clean and codec data; similarly, since NONSIG features are not as affected by the codec, they might exhibit less of a difference in performance between clean and codec test data. Therefore, several experiments will be conducted to examine classification performance for each set in both bitrates.

P-Values $\leq .05$			
	BDES	IEMOCAP	Intersection: SIG
Rate 6	1306	1446	1237
Rate 24	1202	1432	1129
P-Values $>.05$			
	BDES	IEMOCAP	Intersection: NONSIG
Rate 6	276	136	67
Rate 24	380	150	77

Table 1: Number of features grouped by p-values

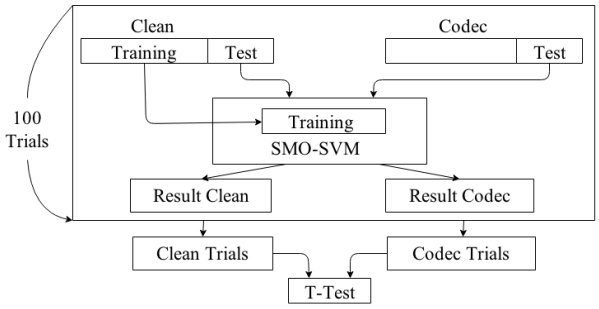

Figure 1: *Experiment for a clean-trained classifier.*

3.3. Classification Experiment

To measure the effect of the codec on classification accuracy, an outline for a set of experiments is created that compares either a clean or codec training set to a clean and codec test. A dataset of either clean or codec features is chosen and a 70% split is performed for training, keeping the number of instances of each class proportional. Two test sets are constructed from the 30% left over and the same 30% of utterances from the other, keeping the utterances the same between test sets. A sequential minimal optimization (SMO) linear support vector machine is trained using a polynomial kernel and tested on both clean and codec test sets. No conventional feature selection methods are chosen; the goal is not to build an optimal classifier, but to observe classifier degradation between test sets. For 100 trials of 70 train / 30 clean and 30 codec test splits, 100 accuracy results are recorded for both the clean and codec test sets and then a paired t-test is performed on them. Figure 1 shows an example of an experiment for a clean-trained classifier. With this evaluation framework, a set of experiments is created which will use both the BDES and IEMOCAP database, at rates 6 and 24.

- Exp. A: Train clean data w/ all feat., test both
- Exp. B: Train codec data w/ all feat., test both
- Exp. C.1: Train clean data w/ SIG feat., test both
- Exp. C.2: Train codec data w/ SIG feat., test both
- Exp. D.1: Train clean data w/ NONSIG feat., test both
- Exp. D.2: Train codec data w/ NONSIG feat., test both

Exp. A results should prove there is a drop in accuracy when training on clean speech and testing on codec speech; such a drop in emotion recognition accuracy is expected, given similar drops in accuracy in speech and speaker recognition. Exp. B will show the performance of codec-trained classifiers. These models are expected to do well on codec speech since they are trained upon it; however, if they perform worse on

clean speech, this suggests that many features are not discriminating among emotions, since both test sets should be perceptually similar. Exps. C.1 and C.2 will show how the SIG intersection set discussed in Table 1 affects the performance of clean-trained and codec-trained classifiers. Since, this set of features is affected by the codec, a difference in accuracy is expected between clean and codec test results. Likewise, Exps. D.1 and D.2 will show how the NONSIG intersection set affects the performance of the classifiers. Since NONSIG features are more resilient to the codec, the difference in accuracy between clean and codec test results should be lower.

4. Results

4.1. Exp. A: Training on Clean, All Features

Table 2 shows results of Exp. A with clean-trained classifiers; for both databases and rates, a clean-trained classifier suffers a statistically significant performance loss on codec speech. Classification of codec speech is lower on the IEMOCAP set than BDES, dropping nearly as low as chance (14.3% for BDES, and 25% for IEMOCAP). This is likely because the utterances of IEMOCAP are more natural and varied than that of BDES, which, by contrast, uses only shorter sentences that express the emotion in an explicit, scripted manner. Comparing rates, drop in accuracy between clean and codec test sets is less for rate 24 than rate 6, and could be expected if one assumes that higher rate codec speech is "closer" to clean speech than a lower rate.

Rate 6		
Test	BDES	IEMOCAP
Clean	85.03%	56.34%
Codec	70.82%	28.45%
P-Value	7.38E-62	5.05E-95
Rate 24		
Test	BDES	IEMOCAP
Clean	85.66%	56.45%
Codec	75.43%	32.79%
P-Value	4.57E-50	2.98E-90

Table 2: Exp. A Results - all features, clean-trained classifiers

4.2. Exp. B: Training on Codec, All Features

Exp. B shows that codec-trained classifiers perform well on codec speech but poorly on clean speech. Since codec speech should not sound perceptibly different, then many of the features are being significantly altered by the codec process and as a result, do not provide true discriminating power among emotion classes. Table 3 verifies this, showing that the difference between clean and codec test sets is significantly different. Surprisingly, despite the higher bitrate, the classifier has a much larger degradation testing on a clean speech for rate 24 than rate 6. This performance is slightly worse than chance on IEMOCAP. A higher bitrate codec-trained classifer performs poorer on the clean testing set than one at a lower bit rate.

4.3. Exp. C.1: Training on Clean, SIG Features

Table 4 shows that for both databases and both rates, clean-trained classifiers using the SIG intersection set perform significantly poorer on codec speech. Again, as in Exp. A, the drop in accuracy is less at rate 24 than at rate 6, but both drops are statistically significant.

Rate 6		
Test	BDES	IEMOCAP
Clean	70.21%	51.39%
Codec	80.50%	54.9%
P-Value	8.92E-49	4.98E-33
Rate 24		
Test	BDES	IEMOCAP
Clean	48.48%	23.26%
Codec	82.40%	56.00 %
P-Value	3.54E-73	1.13E-120

Table 3: Exp. B Results - all features, codec-trained classifiers

Rate 6		
Test	BDES	IEMOCAP
Clean	85.59%	58.04%
Codec	71.11%	27.69%
P-Value	2.62E-60	3.71E-101
Rate 24		
Test	BDES	IEMOCAP
Clean	84.59%	57.91%
Codec	73.25%	32.43%
P-Value	4.17E-55	2.09E-88

Table 4: Exp C.1 Results - SIG, clean-trained classifiers

4.4. Exp. C.2: Training on Codec, SIG Features

Table 5 shows the use of the SIG intersection set on codec-trained classifiers. Results show that these classifiers perform better on codec than clean speech. The average increase in accuracy between testing clean and testing codec is greater for rate 24 than for rate 6; however, for rate 24, as in Exp. B, the codec-trained classifier performs at chance level on IEMOCAP.

Rate 6		
Test	BDES	IEMOCAP
Clean	69.55%	52.00%
Codec	80.71%	56.95%
P-Value	4.41E-46	1.89E-45
Rate 24		
Test	BDES	IEMOCAP
Clean	40.95%	23.11%
Codec	82.16%	57.78%
P-Value	8.94E-85	2.14E-124

Table 5: Exp. C.2 Results - SIG, codec-trained classifiers

Exps. C.1 and C.2 show us again that classifiers trained on clean speech will have difficulty on codec speech and vice versa. Since the performance of this codec-trained classifier drops on clean speech, it suggests that this SIG intersection feature set does not discriminate among emotions well.

4.5. Exp. D.1: Training on Clean, NONSIG

Table 6 shows the results from training clean speech using features identified as NONSIG from Table 1. The t-test results suggest all but one of the classifier comparisons between testing on clean and codec speech is statistically significant (only the result from BDES at Rate 24 has a nonsignificant p-value). Additionally, overall classifier performance when tested on clean speech is significantly less than results reported in Tables 2-5

(except for IEMOCAP results in Table 5 for Rate 24). However, closer examination reveals that, while still showing some statistically significant differences in accuracies, the absolute difference in the average accuracy between testing on clean and codec speech is much reduced from what is reported in Tables 2-5. Also, when considering testing on the codec speech, Table 6 shows a significant increase in average accuracy over the reported results in Tables 2 and 4 (trained on clean speech). This is a significant finding because it suggests that these features are more robust to the speech codec and able to improve performance in the emotion recognition task.

Rate 6		
Test	BDES	IEMOCAP
Clean	45.36%	46.50%
Codec	41.46%	46.04%
P-Value	1.48E-17	1.27E-05
Rate 24		
Test	BDES	IEMOCAP
Clean	64.84%	48.79%
Codec	64.98%	49.29%
P-Value	.691	5.76E-08

Table 6: Exp D.1 Results - NONSIG, clean-trained classifiers

4.6. Exp. D.2: Training on Codec, NONSIG

Table 7 shows the results from training clean speech using the features identified as NONSIG. The t-tests suggest that the classifier comparisons between clean and codec are statistically significant; however, as in Exp. D.1, the absolute difference in average accuracy between testing and codec is again very much reduced. Indeed, the results of the codec-trained classifiers are comparable to that of clean-trained classifiers. The NONSIG codec-trained classifiers perform similarly on clean and codec speech. When considering testing on clean speech, Table 7 shows a significant increase in average accuracy over the results in Tables 2 and 4 (trained on codec speech). Again, this significant finding suggests that the NONSIG intersection set is both robust to the speech codec and provides performance improvements. The NONSIG set provides comparable performance for both clean and codec-trained classifiers.

Rate 6		
Test	BDES	IEMOCAP
Clean	42.92	46.44
Codec	41.50	46.07
P-Value	1.40E-03	1.55E-04
Rate 24		
Test	BDES	IEMOCAP
Clean	64.76	48.77
Codec	65.63	49.30
P-Value	7.22E-03	1.59E-07

Table 7: Exp. D.2 Results - NONSIG, codec-trained classifiers

Figures 2 and 3 summarize the experiments for rates 6 and 24, respectively. The variability between clean and codec test sets is lowest for the NONSIG classifiers. Between rates, higher bitrate codec-trained classifiers perform poorly, especially on IEMOCAP using all features and SIG features, as Exps. B and C.2 show. For the NONSIG feature set, higher bitrate codec-trained classifiers perform slightly better than lower rate codecs.

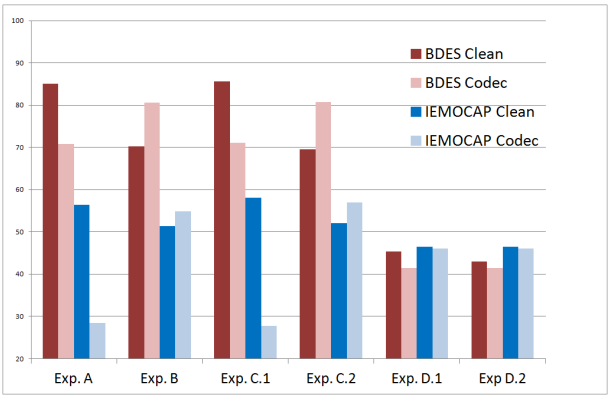

Figure 2: *Rate 6 Accuracy Results*

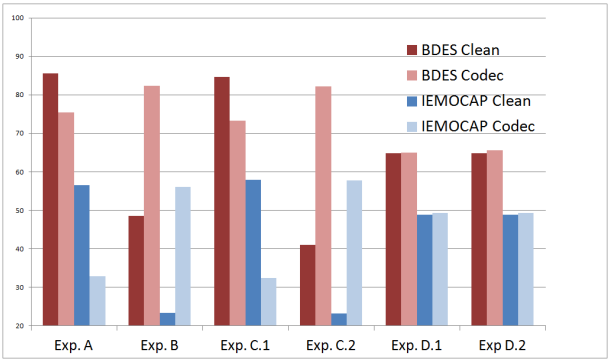

Figure 3: *Rate 24 Accuracy Results*

4.7. A Closer Look at the NONSIG sets

The NONSIG sets for rate 6 and 24 have 67 and 77 features respectively and 33 features in common. The categories of the intersection are of every type except voicing probability, log power mel frequency bands, and line spectral pairs. The functionals for the remaining categories are mostly global statistics including uplevel times, kurtosis, linear regression parameters, and higher order moments. 16 of these global features are delta functionals. Since the perception of emotional content should not be affected by the different bitrates, these results suggest that emotion is better tracked by changes in differential descriptors rather than more local statistics like the mean or standard deviation, as suggested in previous research [11].

5. Conclusions

This paper has verified the expectation that accuracy in emotion classification is significantly degraded when processed through a codec. However, several experiments have also shown features of speech that are not affected by the speech codec process and can produce emotion classification accuracies well above chance (though still not optimal). Additionally, these features showed minimal differences in average classification accuracy regardless of the training model being based on the clean or codec speech. Such a result is a significant impetus to further investigate features that are resilient to speech codecs at various bit rates while still producing high classification accuracy for the design of practical emotion recognition systems in a wide range of communication environments.

6. References

[1] M. Rabiei and A. Gasparetto, "A system for feature classification of emotions based on speech analysis; applications to human-robot interaction," in *Robotics and Mechatronics (ICRoM), 2014 Second RSI/ISM International Conference on*, Oct 2014, pp. 795–800.

[2] A. Austermann, N. Esau, L. Kleinjohann, and B. Kleinjohann, "Fuzzy emotion recognition in natural speech dialogue," in *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on*, Aug 2005.

[3] D.-S. Kwon, Y. K. Kwak, J. Park, M. J. Chung, E.-S. Jee, K.-S. Park, H.-R. Kim, Y.-M. Kim, J.-C. Park, E. H. Kim, K. H. Hyun, H.-J. Min, H. S. Lee, J. W. Park, S. H. Jo, S.-Y. Park, and K.-W. Lee, "Emotion interaction system for a service robot," in *Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on*, Aug 2007.

[4] M. Pantic and L. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE*, vol. 91, no. 9, Sept 2003.

[5] E. Moore, M. Clements, J. Peifer, and L. Weisser, "Investigating the role of glottal features in classifying clinical depression," in *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE*, vol. 3, Sept 2003.

[6] C. Vaudable and L. Devillers, "Negative emotions detection as an indicator of dialogs quality in call centers," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, March 2012.

[7] "Erricson mobility report june 2014," accessed: 2015-3-18. [Online]. Available: http://www.ericsson.com/res/docs/2014/ericsson-mobility-report-june-2014.pdf

[8] "Amr-wb+," accessed: 2015-3-18. [Online]. Available: http://www.voiceage.com/AMR-WBplus.html

[9] B. Klug, "T-mobile announces amr-wb (hd voice) calls active on its network," accessed: 2015-3-18. [Online]. Available: http://www.anandtech.com/show/6594/tmobile-announces-amrwb-hd-voice-calls-active-on-its-network

[10] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge." in *INTERSPEECH*, vol. 2009, 2009, pp. 312–315.

[11] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, 2011.

[12] T. Quatieri, R. Dunn, D. Reynolds, J. Campbell, and E. Singer, "Speaker recognition using g.729 speech codec parameters," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, vol. 2, 2000, pp. II1089–II1092 vol.2.

[13] M. Petracca, A. Servetti, and J. De Martin, "Low-complexity automatic speaker recognition in the compressed gsm amr domain," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, July 2005.

[14] N. Nandan and G. Saha, "On the performance of ip and mobile based automatic speaker verification," in *Communications (NCC), 2012 National Conference on*, Feb 2012, pp. 1–5.

[15] S.-H. Chen and H.-C. Wang, "Improvement of speaker recognition by combining residual and prosodic features with acoustic features," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, vol. 1, May 2004, pp. I–93–6 vol.1.

[16] R. Salami, C. Laflamme, J.-P. Adoul, A. Kataoka, S. Hayashi, T. Moriya, C. Lamblin, D. Massaloux, S. Proust, P. Kroon, and Y. Shoham, "Design and description of cs-acelp: a toll quality 8 kb/s speech coder," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, no. 2, Mar 1998.

[17] R. Lefebvre, R. Salami, C. Laflamme, and J.-P. Adoul, "High quality coding of wideband audio signals using transform coded excitation (tcx)," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, vol. i, Apr 1994, pp. I/193–I/196 vol.1.

[18] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[19] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[20] M. Li, A. Metallinou, D. Bone, and S. Narayanan, "Speaker states recognition using latent factor analysis based eigenchannel factor vector modeling," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, March 2012, pp. 1937–1940.

[21] S. Mariooryad and C. Busso, "Exploring cross-modality affective reactions for audiovisual emotion recognition," *Affective Computing, IEEE Transactions on*, vol. 4, no. 2, pp. 183–196, April 2013.

[22] "3gpp specification detail: Audio codec processing functions; extended adaptive multi-rate - wideband (amr-wb+) codec; transcoding functions," accessed: 2015-3-18. [Online]. Available: http://www.3gpp.org/DynaReport/26290.htm

[23] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the International Conference on Multimedia*, ser. MM '10. New York, NY, USA: ACM, 2010, pp. 1459–1462.

[24] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. A. Müller, and S. S. Narayanan, "The interspeech 2010 paralinguistic challenge." 2010.