AUDIO COMPRESSION AND ITS IMPACT ON EMOTION RECOGNITION IN AFFECTIVE COMPUTING

Alicia Flores Lotz[1], *Ingo Siegert*[1], *Michael Maruschke*[2], *Andreas Wendemuth*[1]
[1]*Institute for Information and Communications Engineering, Cognitive Systems Group, Otto-von-Guericke University Magdeburg, www.cogsy.de*
[2]*Institute of Communications Engineering, Leipzig University of Telecommunications*
alicia.lotz@ovgu.de

Abstract: Enabling a natural (human-like) spoken conversation with technical systems requires affective information, contained in spoken language, to be intelligibly transmitted. This study investigates the role of speech and music codecs for affect intelligibility. A decoding and encoding of affective speech was employed from the well-known EMO-DB corpus. Using four state-of-the-art acoustic codecs and different bit-rates, the spectral error and the human affect recognition ability in labeling experiments were investigated and set in relation to results of automatic recognition of base emotions. Through this approach, the general affect intelligibility as well as the emotion specific intelligibility was analyzed. Considering the results of the conducted automatic recognition experiments, the SPEEX codec configuration with a bit-rate of 6.6 kbit/s is recommended to achieve a high compression and overall good UARs for all emotions.

1 Introduction

In affective computing the assumption is made that machines need to be capable of expressing and recognizing affects to enable a natural Human Computer Interaction. In the last few years, the focus shifted from "acted" to "in the wild" emotion analyses [1]. This term is often linked to the concept of realistic emotions, but however, mobile applications with all their limitations are still out of focus. Most approaches for an automatic speech-based affect recognition still utilize uncompressed, high-quality speech [2], although speech compression techniques have shown a significant impact on acoustic characteristics [3].

Until now, the effects of speech compression on affective speech have been rarely addressed. To do so, one can rely on the well-established base emotions which have obvious ground truth labels, where recognition results on uncompressed data are well known and which therefore can serve as benchline. One of the first studies presented in [4] analyzes the impact of various speech related codecs on emotion recognition performance of Gaussian Mixture Models (GMMs) using only fear-type emotions. The authors of [5] use GMMs as well to analyze the impact of compression on emotion recognition using different acoustic feature types. Unfortunately, both studies focused on the pure recognition results without investigating the underlying spectral error introduced by the compression, although it is known from speech-based emotion recognition studies, that the spectral information is very important to achieve high performances [6]. Additionally, from codec design research, the listening quality assessment is often used to analyze the speech quality [7]. Both evaluations will be investigated within this paper using appropriate measures. Another neglected aspect is the human recognition ability of affect in compressed speech. It gives a feeling of the degradation of acoustic characteristics, regardless of the possible technical insufficiencies of the current imperfect automatic emotion recognition. Both aspects, the spectral error and the human affect recognition ability of compressed speech, are investigated in the current paper and set in relation to the automatic emotion recognition results. Furthermore, the emotion specific recognition ability is analyzed.

2 Audio Coding Formats

This study focuses on three audio codecs. The codecs were chosen such that they comprise different purposes, for music (MP3 [8]), for speech recognition (Speex [9]) and for mobile telephony (AMR-WB [10]). As reference format the standard Waveform Audio File (WAV) [11] is used. Table 1 gives an overview on the chosen codecs and selected bit-rates.

MPEG-1/MPEG-2 Audio Layer III (MP3) is a lossy audio codec, developed by Fraunhofer Institute and released in 1993 [8]. The audio compression is obtained by perceptual coding: certain parts of the original sound signal, considered beyond the auditory resolution ability, are discarded. Afterwards, the remaining information is stored in an efficient manner using Huffman-coding. The bit-rates range from 8 to 320 kbit/s.

Speex (SPX), started as a project in 2002 by the Xiph.Org Foundation [9] is as well a lossy codec. In contrast to MP3, it uses the "analysis-by-synthesis" method, where Linear predictive coding (LPC) coefficients, estimated using the Code-Excited Linear Prediction (CELP) speech coding algorithm, are transmitted and then a synthesis is performed to reconstruct the speech signal. The SPX codec allows bit-rates from 3.95 to 44.2 kbit/s.

Adaptive Multirate Wideband (AMR-WB), the ITU-T Recommendation G722.2, was developed by 3GPP and ETSI for 3G systems [10]. The compression algorithm also uses a "analysis-by-synthesis" method. Based on Algebraic CELP on the down-sampled and pre-processed signal, the LPC parameters and several codebook parameters are transmitted for decoding. To reconstruct the speech signal a synthesis based on these parameters is performed. The AMR-WB codec operates with bit-rates from 6.6 to 23.85 kbit/s.

Table 1 – Overview of selected audio codecs and used bit-rates.

Name	WAV	MP3	Speex	AMR-WB
Released	1991	1993	2003	2001
Compression	No	Yes	Yes	Yes
Loss-less	–	No	No	No
Bit-rate [kbit/s]	256	8, 16, 24, 32, 64, 96	6.6, 11.11, 22.09	6.6, 12.65, 23.85
Filesize [%] of WAV	100	3.34, 6.47, 9.71, 12.94, 26.24, 39.36	2.76, 4.34, 8.62	2.83, 5.18, 9.57

3 Study Design

3.1 Database

To make the results comparable and ensure a high quality of the emotional speech samples, the Berlin Database of Emotional Speech (EMO-DB) is used [12]. The database contains 10 different sentences with neutral semantic content uttered by 10 actors (5 male, 5 female) in seven base emotions (anger, boredom, disgust, fear, joy, neutral and disgust). By conducting a perception test 494 samples with a naturalness over 60% and emotion recognizability over 80% were selected. This reduction of recordings, unfortunately, led to an unbalanced distribution of emotions. An overview of the used subsets of the database is shown in Table 2.

Table 2 – Overview of the used EMO-DB samples.

		total	anger	boredom	disgust	fear	joy	neutral	sadness
automatic emotion recognition	# recordings	494	127	79	38	55	64	78	53
	# male/female	-	5/5	5/5	4/4[1]	5/5	5/5	5/5	5/5
human labeling	# recordings	26	4	4	4	4	3	4	3
	# male/female	-	2/2	2/2	2/2	2/2	1/2	2/2	1/2

All 494 samples were used for calculation of the quality measures, described in the next section, and the automatic emotion recognition experiment. For the human labeling a more

[1] As two speaker do not have samples of disgust, the recall is defined as 0. This also influences the calculation of the mean UAR, but as this holds for all recognition experiments, we acknowledge this as a static error.

balanced subset of these samples was used, containing 26 samples generated from four sentence spoken each by one speakers (2 male, 2 female) in nearly all available emotional states. A more detailed description of the human labeling can be found in section 4.2. For each of the samples contained in the dataset, compressed versions, using the settings given in Table 1, were generated.

3.2 Measures

For the evaluation of the results four measures were used, which will be shortly described.

The **Mean Opinion Score (MOS)** is a subjective measure to rate the quality of a speech signal. It is scaled from 0 "bad" to 5 "excellent" quality. To obtain this value a Perceptual objective listening quality assessment (POLQA) with regard to the ITU-T recommendation P.863 is carried out. This is an objective method to predict the overall listening speech quality as perceived by humans in an ITU-T P.800 Absolute Category Rating listening-only test, [7], [13]. POLQA supports two operating modes: narrowband (300 to 3400 Hz) and super-wideband (50 to 1400Hz). For both methods the prediction algorithm reaches a saturation level at a certain MOS value. In this investigation the super-wideband mode was used, which saturates at 4.75.

As second measure the **Compression Error Rate (CER)** is used [14]. It measures the absolute difference between the original and compressed spectrogram in terms of the samples' variations. The spectrograms are standardized and the CER is determined by calculating the root mean squared error of the absolute error over each window and afterwards computing its mean over all windows. To make the results comparable over each degraded speech sample of a codec and bit-rate, the CER is divided by the averaged maximum error over all samples. This leads to values ranging from 1 "no difference" to 0 "max. difference".

The **Unweighted Average Recall (UAR)** is utilized to measure the mean average recall of one class in a classification task [15]. In our case is used to calculate the accordance of the true known emotions with both the results of the human labeling and the automatic emotion recognition experiment. As ground truth the acted emotions obtained from the corpus description were used. In the human labeling the Unweighted Average Recall (UAR) is calculated for each labeler and then averaged over all ($\overline{UAR_h}$). For the emotion recognition experiment the UAR is calculated for each validation step and afterwards averaged over all steps ($\overline{UAR_a}$).

Spearman's rank correlation coefficient (R_s) rates the correlation between two ordinal and/or metric scaled measures [16]. Generally, values below 0 show an inverse dependency, from 0 to 0.2 poor to non relation, 0.2 to 0.5 a weak to moderate relation, 0.5 to 0.8 a clear relation and values above 0.8 a high to perfect relation.

3.3 Experimental Realization

Human Labeling was used to analyze the effect of compression on the human ability to recognize emotions. A labeling experiment employing seven native German speaking participants (2 females, 5 males), similar to the experiment in [17], was conducted. None of them had taken part in this kind of study before. At the beginning of the labeling, all participants had to go through a training phase by listening to selected samples of one speaker having the same sentence recorded in all seven emotional conditions. In this way, the understanding of how the different emotions are uttered by the actors was ensured for all participants. Samples of the training speaker were not used in the main study. Afterwards, the participants had to listen to a total of 338 samples. The samples were presented in a random order that was fixed over all participants. The labelers could listen to each sample several times, but they were not able to revise given labels. The whole process took on average 100 minutes. To facilitate the labeling process, after 26 sound-samples a 3 min music file was played. The labeling was conducted using an updated version of the software tool ikannotate [18].

Automatic Emotion Recognition was used to investigate the influence of compression. Therefore, state of the art experiments were conducted. A Support Vector Machine with linear kernel and a cost factor of 1 was utilized with WEKA [19]. The "emobase" feature set provided by openSMILE [20], comprising 988 acoustic characteristics, was used. Additionally, standardization was utilized to normalize the data [21]. For the various automatic recognition experiments the train and test sets comprise the same codecs. This secured that the possible acoustic degradation introduced by the codecs was the same for training and testing of the classifiers. To accomplish speaker independence, a Leave-One-Speaker-Out validation scheme was used. The classifier was trained on all speakers except one. This remaining one was afterwards used for testing. The procedure was repeated for every speaker.

4 Results

In the following, first a comparison of the quality measures using Spearman's rank correlation coefficient is presented. Then the results of the human labeling and the automatic emotion recognition are introduced separately before concluding with a comparison of both methods.

4.1 Quality Assessment (MOS vs. CER)

For the quality assessment the R_s was utilized. First, a direct comparison between MOS and CER was conducted. Afterwards, the mean UAR of the human labeling and the automatic recognition experiment were incorporated, to state if and how well these measures correlate.

Table 3 – Mean $\overline{UAR_h}$,$\overline{UAR_a}$ and quality assessment measures (CER & MOS) over all emotions.

codec	AMR	AMR	AMR	MP3	MP3	MP3	MP3	MP3	MP3	SPX	SPX	SPX	WAV
bit-rate	6.6	12.65	23.85	8	16	24	32	64	96	6.6	11.11	22.09	256
$\overline{UAR_a}$ [%]	73.70	76.11	75.17	70.19	69.93	75.59	79.67	78.85	79.34	77.83	74.29	75.28	77.25
$\overline{UAR_h}$ [%]	89.46	90.82	90.31	88.95	90.65	91.50	92.01	93.37	92.01	86.56	92.52	92.69	94.39
CER	0.9749	0.9787	0.9661	0.9710	0.9771	0.9809	0.9836	0.9937	0.9971	0.9600	0.9674	0.9757	1.0000
MOS	2.71	3.51	3.79	1.68	2.82	3.51	4.04	4.55	4.61	2.07	2.96	3.91	4.59

Table 3 presents the resulting MOS, CER, $\overline{UAR_h}$ and $\overline{UAR_a}$ values. By conducting a top-down ranking the ranks within the measures were determined, ranking the highest value with the best rank "1". The comparison of MOS and CER results in $R_{s(MOS/CER)} = 0.7730$ and therefore specifies a clear relation between these measures. The correlation between the other measures is as follows: $R_{s(MOS/\overline{UAR_h})} = 0.7989$, $R_{s(CER/\overline{UAR_h})} = 0.6878$, $R_{s(MOS/\overline{UAR_a})} = 0.6850$, $R_{s(CER/\overline{UAR_a})} = 0.5495$. Thus, a clear relation of the measures is recognized for all cases. As mentioned in the introduction, the spectral information has a high impact on the performance of speech-based emotion classifiers. Therefore, it was assumed that the CER can be used to forecast a trend of the performance of the classifier better, compared to MOS. The results, unfortunately, do not confirm this assumption, as MOS reaches a slightly higher correlation for both human labeling and automatic emotion recognition. This means, that not only spectral but also other features (e.g. prosodic) are strongly influenced by the compression. Also noticeable is the moderate correlation between $\overline{UAR_h}$ and $\overline{UAR_a}$ ($R_s = 0.4154$). This led to the assumption that no conclusion on the performance of the emotion classifier can be drawn from the results of the human labeling.

4.2 Human Labeling Results

As stated in Table 3 the $\overline{UAR_h}$ achieves for all codecs an average recall of over 86%. Considering the ranking of the different codecs and bit-rates it can be seen, that codecs with a high filesize reduction (cf. Table 1) result in a lower UAR and higher standard deviation. The UAR rises consecutively with increasing bit-rates within different codecs, except for MP3 with a bit-rate of 96 kbit/s. In Figure 1 this evolution is depicted.

Table 4 – Mean of the $\overline{\text{UAR}_h}$ for each analyzed bit-rate and individual emotion. The lowest recognition results of each emotion are highlighted in dark color.

codec	AMR	AMR	AMR	MP3	MP3	MP3	MP3	MP3	MP3	SPX	SPX	SPX	WAV
bit-rate	6.6	12.65	23.85	8	16	24	32	64	96	6.6	11.11	22.09	256
anger	96.43	96.43	96.43	100	100	100	96.43	100	96.43	100	100	96.43	100
boredom	78.57	85.71	85.71	85.71	85.71	96.43	92.86	92.86	92.86	82.14	96.43	89.29	92.86
disgust	82.14	89.29	89.29	82.14	85.71	85.71	82.14	92.86	82.14	82.14	89.29	85.71	96.43
fear	82.14	82.14	82.14	82.14	85.71	82.14	85.71	92.86	85.71	75.00	89.26	89.29	85.71
joy	90.48	90.48	95.24	90.48	95.24	95.24	95.24	90.48	95.24	90.48	90.48	95.24	95.24
neutral	96.43	96.43	92.86	96.43	96.43	85.71	96.43	89.29	96.43	100	96.43	92.86	100
sadness	100	95.24	90.48	85.71	85.71	95.24	95.24	95.24	95.24	76.19	85.71	100	90.48

Table 4 shows the results of the human labeling specified over each analyzed bit-rate for each emotion. All emotions can be detected with a high UAR of at least 75%. Only in three cases an average recall of over 80% could not be achieved. In both cases these results were obtained by the lowest bit-rate of two different codecs (SPX and AMR-WB). For these codec configurations the UAR shows the lowest value for certain emotions, compared to the remaining codec configurations. The entries with the lowest recognition results for each emotion over every codec and bit-rate are marked in dark color. In case of AMR-WB and SPX with a bit-rate of 6.6 kbit/s this is the case for 4 out of 7 emotions. It should be also noticed, that for two emotions (anger and joy) a UAR over 90% is always achieved. For these emotions the labelers had a high consistency in their decision making during the labeling process.

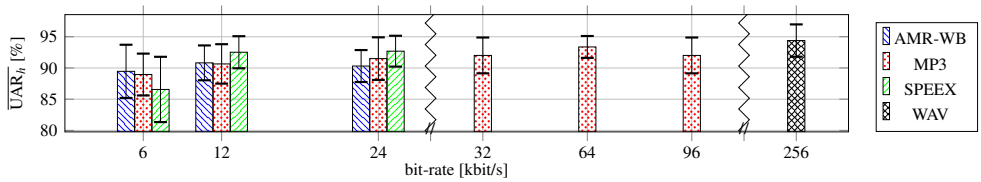

Figure 1 – Mean and standard deviation of the $\overline{\text{UAR}_h}$ for each analyzed bit-rate over all emotions.

4.3 Automatic Emotion Recognition Results

Using the average recall over every emotion a recommendation on the usage of codecs and their bit-rates was made. For the evaluation all results with an above average result were assumed to be suitable for the recognition of this emotion. In Table 5 these entries are marked in bright color. Considering all codecs an over-average result for all emotions could be achieved for MP3-32. Satisfactory results (6 out of 7 higher than average) were achieved for MP3 with the bit-rates 64 and 92 kbit/s. The worst result with no emotion being recognized above average is reached by MP3-16. This confirms, on the one hand, that higher bit-rates result in higher audio-quality and a better recognition of emotions, but on the other hand disregards the comparability of codecs with same bit-rates. For this reason a second recommendation was made only regarding codecs with similar bit-rates. For two codec configurations (SPX with 6.6 and 22.09 kbit/s) an over-average result was achieved for 5 out of 7 emotions. For Table 5 the dark colored entries denote the lowest recall for one emotion. SPX-22.09 includes one of these entries for the recognition of anger. To achieve an overall good recognition of the emotions SPX-6.6 can be recommended. Surprisingly this configuration also has the highest filesize compression of 2.76% of the original WAV-samples (cf. Table 1). Depending on the recognition task, it is also possible to use different compression codecs to recognize certain emotions. For the recognition of disgust and sadness SPX-6.6, for anger and boredom MP3-8, for fear and joy AMR-WB-12.65 and for neutral emotion SPX-11.11 give the best results. The recommendations drawn from Table 5 are also confirmed by the results depicted in Table 3 and Figure 2. The low standard deviation of SPX-6.6, compared to other configurations, emphasizes the given recommendation. It is in the same range as the standard deviation of the UAR obtained by the emotion recognition on the original WAV data samples and configurations with higher bit-rates.

Table 5 – Mean of the $\overline{\text{UAR}_a}$ for each analyzed bit-rate and individual emotion. The lowest recognition results of each emotion are highlighted in dark, results above average are highlighted in bright color.

codec	AMR	AMR	AMR	MP3	MP3	MP3	MP3	MP3	MP3	SPX	SPX	SPX	WAV
bit-rate	6.6	12.65	23.85	8	16	24	32	64	96	6.6	11.11	22.09	256
anger	90.81	90.72	90.46	92.03	85.88	87.38	91.29	88.86	92.77	87.61	90.15	85.49	92.00
boredom	87.13	85.42	84.06	88.92	82.74	82.85	90.81	88.56	87.10	86.78	85.20	88.88	86.85
disgust	39.71	42.57	41.14	42.52	33.09	50.47	50.71	49.71	53.57	52.14	38.71	39.71	53.57
fear	84.87	86.52	85.74	57.49	80.33	80.58	84.58	87.09	86.53	84.63	73.60	83.18	80.92
joy	62.24	68.50	59.89	50.48	56.00	65.76	70.21	65.44	69.94	67.13	66.81	64.02	65.65
neutral	79.53	78.01	81.37	78.96	78.16	79.44	88.59	84.60	80.28	81.03	88.77	84.14	82.71
sadness	71.59	81.03	83.53	80.90	73.30	82.66	81.51	87.66	85.16	85.52	76.79	81.51	79.05

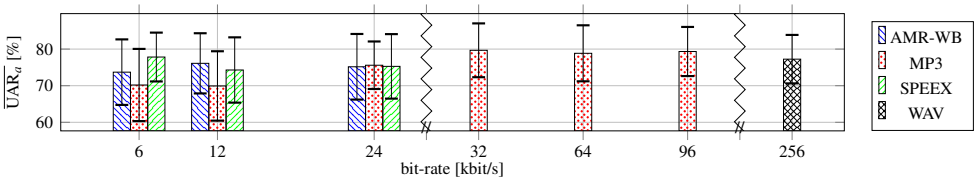

Figure 2 – Mean and standard deviation of the $\overline{\text{UAR}_a}$ for each analyzed bit-rate over all emotions.

4.4 Comparison of Human Labeling vs. Automatic Emotion Recognition

Comparing the results of the two previous sections, it was noticed, that human labeling achieves an overall higher recall for all emotions and codecs, than the automatic emotion recognition. This was expectable, as the human auditory system is able to distinguish emotions in a high resolution. Considering Figures 1 and 2, a clear difference can be seen regarding SPX-6.6. For this codec configuration the results of the human labeling show the lowest $\overline{UAR}$ with the highest standard deviation of all configurations. Surprisingly, in case of the automatic emotion recognition, this configuration shows the highest result, considering low bit-rates, with one of the lowest standard deviations.

Furthermore, the two classification approaches can be examined comparing their correlation coefficients for each emotion over all codecs (e.g. row "anger" in Table 4/Table 5 ($R_{s(anger)}$)) and for each codec over all emotions (e.g. column "AMR-WB-6.6" in Table 4/Table 5 ($R_{s(AMR-6.6)}$)). Regarding R_s for each emotion, the values range from -0.25 to 0.04. They characterize an anti-correlation for anger (-0.25) and no correlation for all other emotions. This implies, that there is no relation between human labeling and the automatic recognition of certain emotions using different codecs (e.g. anger achieves the best results with MP3-96 for the automatic recognition and the worst result for human labeling). For the codec evaluation R_s ranges in different areas for each codec (AMR-WB: -0.05 to 0, MP3: 0.33 to 0.67, SPX: 0.05 to 0.71 and WAV: 0.25). MP3 shows the highest correlation, AMR-WB achieved the lowest and SPX has the highest variation within the considered bit-rates. The high relation using MP3 is explicable by the codec's compression algorithm. It uses perceptual coding, discarding parts of the signal considered beyond the auditory resolution ability. Furthermore, it is specified for the optimal compression of music and most music is based on inducing emotions in the listener.

5 Conclusion and Outlook

The investigations presented in this paper examine several problem statements. First, it is examined if it is possible to forecast a trend of the performance of a speech-based emotion classifier using suitable quality measures (POLQA-MOS and CER). As the spectral information has a high impact on the performance of an emotion classifier, it was assumed, that the CER is more suitable to do so. A clear relation between the MOS and CER could be shown using Spearman's rank correlation coefficient. In case of the human labeling and automatic recognition of emotions both measures show a clear relation, still MOS achieved slightly higher values than CER.

Second, the problem of emotion recognition on degraded speech is addressed: If compression of the data is needed, which codec achieves the best UAR for emotion recognition? The study showed, that in case of an overall good UAR for each emotion and a low filesize compression (bit-rates over 24 kbit/s), the MP3 codec with bit-rates of 32 kbit/s and higher is suitable for a satisfactory result. In case of a high filesize compression (low bit-rates), the SPX codec with a bit-rate of 6.6 kbit/s should be used. This codec configuration also shows a low standard deviation and the highest filesize reduction (2.76% of WAV). Finally, it is noticed, that the human labeling achieves an overall higher recall than the emotion recognition experiment. This is expectable, as the human auditory system is able to distinguish emotions in a high resolution. Surprisingly, the codec configuration recommended for high filesize reduction for automatic emotion recognition shows a contrary behavior in the results of the human labeling. As an outlook it can be stated, that a deeper understanding of the codec's compression algorithm is needed to evaluate the results in detail.

Acknowledgments

The work presented in this paper was done within the Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" (www.sfb-trr-62.de) funded by the German Research Foundation (DFG). We would further like to thank SwissQual AG (a Rhode & Schwarz company), in particular Jens Berger, for supplying the POLQA testbed.

References

[1] DHALL, A., R. GOECKE, G. T., and N. SEBE: *Emotion recognition in the wild. Journal on Multimodal User Interfaces*, 10, pp. 95–97, 2016.

[2] ZENG, Z., M. PANTIC, G. I. ROISMAN, and T. S. HUANG: *A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. IEEE Trans. Pattern Anal. Mach. Intell.*, 31, pp. 39–58, 2009.

[3] BYRNE, C. and P. FOULKES: *The 'mobile phone effect' on vowel formants. International Journal of Speech, Language and the Law*, 11(1), pp. 83–102, 2004.

[4] GARCÍA, N., J. C. VÁSQUEZ-CORREA, J. D. ARIAS-LONDOÑO, J. F. VÁRGAS-BONILLA, and J. R. OROZCO-ARROYAVE: *Automatic emotion recognition in compressed speech using acoustic and non-linear features*. In *20th Symposium on Signal Processing, Images and Computer Vision (STSIVA)*, pp. 1–7. Bogota, Colombia, 2015.

[5] ALBAHRI, A., M. LECH, and E. CHENG: *Effect of speech compression on the automatic recognition of emotions. International Journal of Signal Processing Systems*, 4(1), pp. 55–61, 2016.

[6] SCHULLER, B., A. BATLINER, S. STEIDL, and D. SEPPI: *Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Commun*, 53(9-10), pp. 1062–1087, 2011.

[7] ITU-T: *Methods for objective and subjective assessment of speech quality (POLQA): Perceptual Objective Listening Quality Assessment*. REC P.863, International Telecommunication Union (Telecommunication Standardization Sector), 2014. URL `http://www.itu.int/rec/T-REC-P.863-201409-I/en`.

[8] BRANDENBURG, K.: *MP3 and AAC Explained*. In *17th AES International Conference: High-Quality Audio Coding*. Florence, Italy, 1999.

[9] VALIN, J.-M.: *Speex: A free codec for free speech*. In *Proc. of the linux.conf.au*. Geelong, Australia, 2006.

[10] ITU-T: *Wideband Coding of Speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)*. REC G.722.2, International Telecommunication Union (Telecommunication Standardization Sector), 2003. URL `https://www.itu.int/rec/T-REC-G.722.2-200307-I/en`.

[11] IBM CORPORATION AND MICROSOFT CORPORATION: *Multimedia programming interface and data specifications 1.0*. Tech. Rep., 1991. URL `https://www.aelius.com/njh/wavemetatools/doc/riffmci.pdf`.

[12] BURKHARDT, F., A. PAESCHKE, M. ROLFES, W. SENDLMEIER, and B. WEISS: *A Database of German Emotional Speech*. In *Proc. of the INTERSPEECH 2005*, pp. 1517–1520. Lisbon, Portugal, 2005.

[13] ITU-T: *Methods for subjective determination of transmission quality*. REC P.800, International Telecommunication Union (Telecommunication Standardization Sector), 1996. URL `https://www.itu.int/rec/T-REC-P.800-199608-I/en`.

[14] SIEGERT, I., A. F. LOTZ, L. L. DUONG, and A. WENDEMUTH: *Measuring the impact of audio compression on the spectral quality of speech data*. In O. JOKISCH (ed.), *Elektronische Sprachsignalverarbeitung 2016. Tagungsband der 27. Konferenz*, vol. 81 of *Studientexte zur Sprachkommunikation*, pp. 229–236. TUDpress, Leipzig, Germany, 2016.

[15] ROSENBERG, A.: *Classifying skewed data: Importance weighting to optimize average recall*. In *Proc. of the INTERSPEECH-2012*, pp. 2242–2245. Portland, USA, 2012.

[16] SPEARMAN, C.: *The proof and measurement of association between two things*. *American Journal of Psychology*, 15, pp. 88–103, 1904.

[17] SIEGERT, I., A. F. LOTZ, M. MARUSCHKE, O. JOKISCH, and A. WENDEMUTH: *Emotion intelligibility within codec-compressed and reduced bandwidth speech*. In *ITG-Fachbericht 267:Speech Communication*, pp. 215–219. VDE VERLAG GMBH Berlin Offenbach, Paderborn, Germany, 2016.

[18] BÖCK, R., I. SIEGERT, M. HAASE, J. LANGE, and A. WENDEMUTH: *ikannotate – a tool for labelling, transcription, and annotation of emotionally coloured speech*. In *Affective Computing and Intelligent Interaction*, vol. 6974 of *LNCS*, pp. 25–34. Springer, Berlin, Heidelberg, Germany, 2011.

[19] HALL, M., E. FRANK, G. HOLMES, B. PFAHRINGER, P. REUTEMANN, and I. H. WITTEN: *The weka data mining software: An update*. *SIGKDD Explor. Newsl.*, 11(1), pp. 10–18, 2009.

[20] EYBEN, F., M. WÖLLMER, and B. SCHULLER: *openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor*. In *Proc. of the ACM MM-2010*, p. s.p. Firenze, Italy, 2010.

[21] ZHANG, Z., F. WENINGER, M. WÖLLMER, and B. SCHULLER: *Unsupervised learning in cross-corpus acoustic emotion recognition*. In *Proc. of the IEEE ASRU-2011*, pp. 523–528. Waikoloa, USA, 2011.