MEASURING THE IMPACT OF AUDIO COMPRESSION ON THE SPECTRAL QUALITY OF SPEECH DATA

Ingo Siegert, Alicia Flores Lotz, Linh Linda Duong, Andreas Wendemuth

Cognitive Systems Group, Faculty of Electrical Engineering and Information Technology
Otto-von-Guericke University, Universitätsplatz 2, 39106 Magdeburg, Germany
Email: ingo.siegert@ovgu.de

Abstract: Due to the grown use of speech technology, the need for efficient data storage becomes increasingly important. In this paper, we investigate whether well known audio-codecs for music data can be used to store speech data without introducing too many spectral errors. Our investigations are concluded with the recommendation to use Ogg Vorbis in its highest quality setting for data storage.

1 Introduction

Within the last years, deep learning has gained a lot of attraction as a sophisticated method for automatic speech recognition [1, 11]. In comparison to established methods as Support Vector Machines or Hidden Markov Models, deep learning requires much more data material [8]. Furthermore, speech corpora still rely on high quality data without compression errors which could distort the feature extraction and in the end bias the modeling. But as more and more data is needed it is worth to investigate whether audio compression is useful to store large amounts of data and whether the compression error of certain codecs is small enough to be neglected. This paper presented a study, comparing spectrograms of uncompressed and compressed acoustic data. Furthermore, the compression error rate is defined to draw conclusions on the compression quality. The analysis of the spectrogram allows us to identify both specific frequency sub-bands as well as specific acoustic parts that are influenced by the compression.

2 Audio Storage Formats and Compression Methods

For audio coding a wide range of codecs is available, ranging from non-compression over lossless, and lossy compression for general or special purposes (voice, mobile phones) [10]. The problem of lossy compression is a decreased quality, which cannot be corrected in the uncompression phase. Since lossless compression cannot simply replace lossy codecs because of low compression rates (cf. Section 4.1.), and in view of the overwhelming presence of popular lossy codecs, the need arises for an investigation which compares compression rate and error for various codecs of all natures. Since we are interested in speech data, our investigations are focussed on selected (music) audio codecs. Besides the well known codecs MP3 and AAC, we also investigated Ogg Vorbis, Speex and Opus, see Table 1.

Waveform Audio File Format (WAV) is a standard for storing audio bitstreams developed by Microsoft and IBM in 1991 [7]. It is mainly used to store raw and uncompressed audio data using linear pulse-code modulation (LPCM) as bitstream encoding. LPCM retains all of the samples of an audio track. An uncompressed Audio CD has a bit-rate of 1,411.2 kbit/s.

Free Lossless Audio Codec (FLAC) uses linear prediction for the lossless compression of digital audio [6]. The difference between the predictor and the actual sample data is calculated and stored using Golomb-Rice coding to reduce the needed bit size. Furthermore, for blocks

of identical samples, run-length encoding is used. FLAC supports different compression levels, which do not influence the quality, but size and speed of the compression.

MPEG-1/MPEG-2 Audio Layer III (MP3) MPEG-1/MPEG-2 Audio Layer III (MP3) is a lossy audio data compression codec [4]. It was developed by Fraunhofer Institute and released in 1993. For audio compression perceptual coding is used: certain parts of a sound considered to be beyond the auditory resolution ability discarded. The remaining information is afterwards stored in an efficient manner. The bit-rate ranges from 8[1] to 320 kbit/s.

Advanced Audio Coding (AAC) is a coding standard for lossy audio compression developed as successor of MP3 by the Moving Picture Experts Group [12]. Some of the improvements include: more sample frequencies, arbitrary bit rates and variable frame lengths, and higher blocksize for stationary and transient signals [4]. Due to higher flexibility than MP3, AAC achieves a more efficient compression. The bit-rate ranges from 16 to 320 kbit/s.

OGG Vorbis was developed as a patent-free alternative to MP3 and competitor to AAC by the Xiph.Org Foundation [14]. Vorbis uses a modified discrete cosine transform (MDCT). The noise floor and residue components of the resulting frequency-domain data are quantized and entropy coded using a codebook-based vector quantization algorithm. Ogg Vorbus uses a bit-rate in the range from 48 to 500 kbit/s.

Speex is an open source lossy audio compression format designed directly for speech applications by the Xiph.Org Foundation [19]. It is based on the (Code-excited Linear Prediction) CELP speech coding algorithm. Speex is now cosidered obsolete; its successor is the Opus codec, which surpasses Speex's performance in all areas. The bit-rate ranges from 2 to 24 kbit/s.

Opus is the successor of Speex and also an open source lossy audio compression format deveoped by the Xiph.Org Foundation [18]. Opus is a hybrid codec combined from two differeno-riginally separate algorithms: the speech-oriented SILK and the low-latency CELT. SILK based on Linear Predictive Coding (LPC) and CELT uses MDCT in combination with CELP frequency domain. For frequency ranges above 8kHz a combination of both SILK and CELP used for encoding. The bit-rate ranges from 8 to 128 kbit/s.

Windows Media Audio (WMA) is a proprietary audio data compression technology developed by Microsoft. The default lossy compression method of WMA is based on the same principle as MP3: After a conversion into a frequency-amplitude domain masking effects (near frequencies which are not distinguishable), or sounds that are not audible (hearing threshold) are deleted. The bit-rate ranges from 24 to 448 kbit/s[2].

Table 1 - Overview of selected Audio Codecs

Name	WAV	FLAC	MP3	Vorbis	AAC	Speex	Opus	WMA
Released	1991	2001	1993	2000	1997	2003	2012	1999
Compression	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Loss-less	–	Yes	No	No	No	No	No	No
Bit-rate (kbit/s)	1,411.2	935	16–320	48–500	16–320	2–24	8–128	32–448
Encoder	–	flac	lame	oggenc	ffmpeg	speexenc	opusenc	ffmpeg
Decoder	–	ffmpeg	lame	oggdec	ffmpeg	speexdec	opusdec	ffmpeg

3 Study Design

Usually, the codec quality is rated in double-blind listening tests. A small sample of unknown music files (<20) should be identified as being the original or the encoded version. Another

[1] We skipped the 8kbit/s compression rate in our investigation, as the encoder limits the sampling rate to 8kHz. Also, in listening experiments this compression rate achieves very bad ratings (cf. [15]).

[2] Higher bit-rates up to 768kbit/s are available using the WMA Pro version.

possibility for speech quality assessment is the ITU-standard PESQ, the Perceptual Evaluation of Speech Quality [13]. PESQ is used for objective quality testing in telephony systems. But, as it only characterizes the voice quality it does not take into account spectral changes affecting the emotional content. In our study we wanted to perform a more systematic analysis and chose the Berlin Database of Emotional Speech. This corpus offers a comparison platform of high quality acted samples of affective speech [5] and is well known in the speech and acoustic emotion recognition community. The recordings are done in an anechoic cabin with a sampling rate of 16kHz. Ten (five male, five female) professional actors speak ten German sentences with emotionally neutral content. It contains 494 phrases, where both naturalistic and pre-identified emotions are present. As emotional categories anger, boredom, disgust, fear, joy, neutral, and sadness are used.

To perform the compression analyses, we select all emotional neutral recordings of this database. From these 79 samples, different kinds of compressed versions for each of the different codecs and bit-rate settings were generated. Table 2 depicts all used compression settings.

Table 2 - Overview of utilized compression settings (bitrates) for each codec. For the values in brackets WMA encodes bigger files than the original WAV file.

FLAC	1	2	3	4	5	6	7	8								
bit-rate	2.15	3.95	5.95	7.15	8	10.15	11	14.15	15	16	18.2	24	32	40	48	56
MP3	–	–	–	–	–	–	–	–	–	X	–	X	X	X	X	X
Vorbis	–	–	–	–	–	–	–	–	–	–	–	–	–	–	X	–
AAC	–	–	–	–	–	–	–	–	–	X	–	–	X	–	–	–
Speex	X	X	X	X	X	X	X	X	X	–	X	X	–	–	–	–
Opus	–	–	–	–	X	–	–	–	–	X	–	–	X	–	X	–
WMA	–	–	–	–	–	–	–	–	–	–	–	X	X	–	X	–
bit-rate	64	80	96	112	128	160	176	192	224	256	320	384	416	448	500	
MP3	X	X	X	X	X	X	X	X	X	X	X	X	X	X	–	
Vorbis	X	X	X	X	X	X	–	X	X	X	X	–	–	–	X	
AAC	–	–	X	–	X	X	–	X	–	X	X	–	–	–	–	
Speex	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	
Opus	X	–	X	X	X	–	–	–	–	–	–	–	–	–	–	
WMA	X	–	X	–	X	X	–	–	–	X	X	–	–	X	–	

To visualize the influence of the compression, the spectrograms of the decoded files were compared to their original WAV-file. Therefore, the spectrogram was calculated using a Hamming-window with a window size of 200 samples and an overlap of 80 samples followed by a short-time Fourier analysis on the windowed signal. A spectrogram as shown in Fig. 2(a) depicts the power spectral density for each window in [dB/Hz]. The absolute error between two windows of the decoded and original data is denoted in [dB].

Unfortunately, both encoding and decoding adds a specific amount of zeros to the beginning and ending of the audio data in order to properly apply the MDCT/filterbank routines [2, 17]. This so-called "algorithmic delay" is dependent on the codec implementation, the bit-rate, and the sampling frequency. Thus, to calculate the absolute error between two windows, this delay has to be removed beforehand. As this delay is not documented for every codec, a manual inspection to specify the correct delay was performed.

To specify the spectrograms error, several measures can be used. In [16], the spectral center of gravity is investigated. This measures the first spectral movement and is perceptually connected with the impression of the sound-"brightness". But as we are interested in the error differences between the original and encoded spectrum in terms of the sample standard variation, we determined a compression error rate (ER_c) using the absolute error obtained from the comparison of the spectrograms from the compressed and associated original file. The error rate was de-

termined by calculation of the root mean squared error (RMSE) of the absolute error over each window i ($i = 1, ..., m$) of the spectrogram and afterwards computing its mean over all windows:

$$RMSE(i) = \sqrt{\frac{\sum(error(i,:)^2)}{FS}} \quad (1) \qquad ER_c = \frac{\sum(RMSE(i))}{m} \quad (2)$$

FS corresponds to the number of samples along the frequencies.

The relation of the compression error rate and the compression ratio was finally used, to draw a recommendation of possible codecs for speech processing. As this investigation is intended to be used for offline speech processing, for instance, data storage, the compression times for encoding and decoding were not considered.

4 Results

Audio compression aims to minimize the space needed to store the data, therefore, we first analyzed the compression ratio achieved with the different codecs and compression ratios. Secondly, selected spectrograms were depicted and the observed error were discussed. Afterwards, the compression error rate was analysed in detail between different codecs and compression-rates. Finally, by comparing the compression error rate and the compression ratio, a recommendation for the compressed data storage was made.

4.1 Achieved Compression Ratio

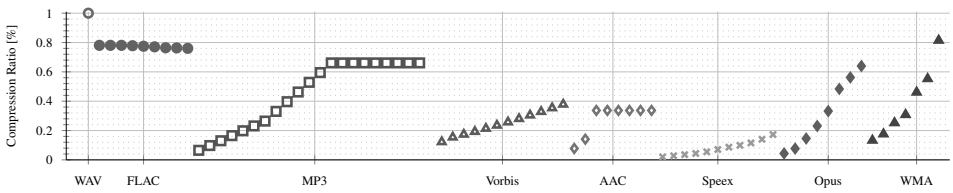
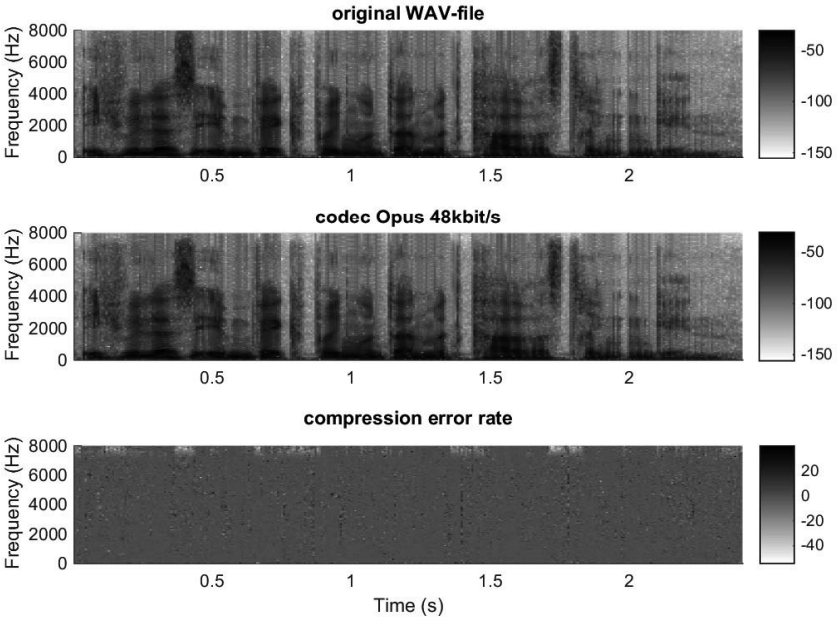


Figure 1 - Achieved average compression ratio for each codec and bit-rate. The bit-rate is increasing from left to right, see Table 2.

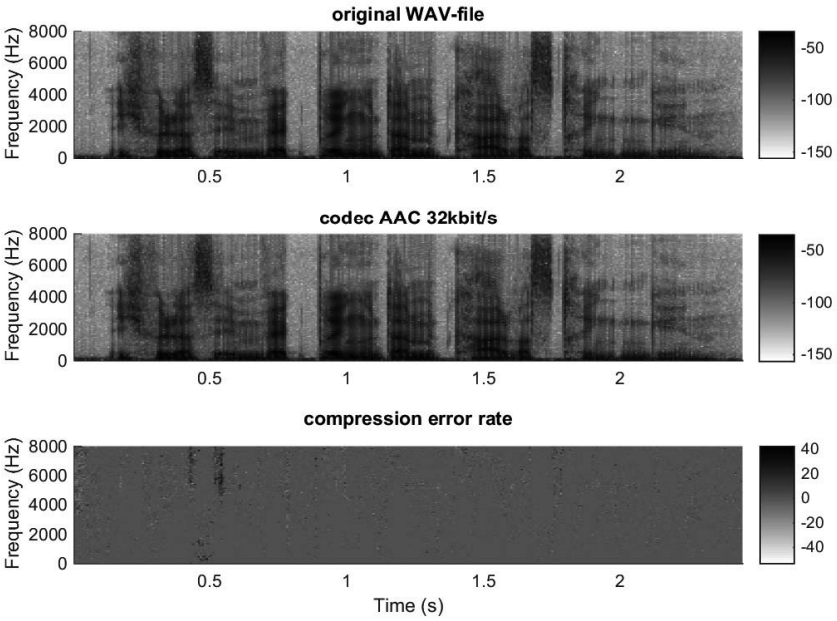
Comparing the different compression ratios (cf. Fig. 1), we see that FLAC has the smallest reduction with only 76.04% at best. The well known codecs and direct competitors MP3 and WMA comprise similar compression ratios with a small advantage for MP3: 6.51% (16kbit/s) to 66.11% (160kbit/s and above) for MP3 and 13.30% (24kbit/s) to 81.44% (160kbit/s) for WMA. Surprisingly, for higher bit-rates (256-448kbit/s) the WMA-encoded file is up to three times larger than the original file. For MP3 encoded files, the compression ratio remains the same for bit-rates from 160 to 448kbit/s. Using Opus similar compression ratios in comparison with MP3 could be achieved. A much better compression for high bit-rates can be gained using the successor codecs. The compression ratio for Vorbis is at 37.84% for 500kit/s and at 33.70% for 160kbit/s using AAC. The speech-specialized codec Speex achieves the lowest compression rates below 17.24%, reaching 2.00% at best.

4.2 Spectrogram evaluation

As FLAC is a loss-less codec, the spectrograms of the original and decoded files should be identical. This was the case for all considered files and thus verifies the results of the spectrogram calculation for further investigations on the different audio codecs.



(a) File: 11b09Na, Codec: Opus with 48kbit/s bit-rate



(b) File: 15b09Nb, Codec: AAC with 32kbit/s bit-rate

Figure 2 - Spectrograms of the original and decoded WAV-file in [dB/Hz] using Opus-codec with a bit-rate of 48kbit/s and its resulting absolute error in [dB].

In the following two spectrograms are exemplarily plotted (cf. Fig. 2). In the upper spectrograms of the original and decoded files the light white to gray areas depict a low intensity of the speech material. The lower subfigure shows the absolute error of the spectrograms. Here, in the gray regions the absolute error is zero. For dark regions the decoder over-estimates the frequency ratio of the original WAV-data. Contrarily, the decoder under-estimated the frequency ratio for the light regions.

Most important for a good understanding of speech are the lower cepstral formants occurring in the frequency bands up to 4kHz [9]. Therefore, we assumed that for codecs directly designed for speech application (Speex, Opus), higher order frequency bands are seen as less important and thus are compressed more than lower frequency bands to achieve the desired compression rate. MP3/WMA also make use of this assumption as stated in Section 2, but to a lower extent. When taking a look at the spectrograms of MP3/WMA or Opus a clear cut-off of the higher frequency bands (lowpass behavior) can be recognized (cf. Fig. 2(a)). The upper frequencies from approx. 7 to 8 kHz have been cut-off. Only for Speex this phenomenon was not observable as the algorithm uses a slightly different approach than Opus, even though, Opus is the successor of Speex and one could expect that they work in a similar way.

A second observation from spectrograms is that the error occurred mostly in less intensive regions of the audio signal. As the codecs are designed for music compression, which is assumed to not contain many silent parts, the encoder concentrates on correctly compressing parts of a certain intensity rather than silent or low intense parts. Furthermore, the higher the bit-rate, the better the less intensive regions get decoded, as depicted in Fig. 2(b).

4.3 Compression Error Rate

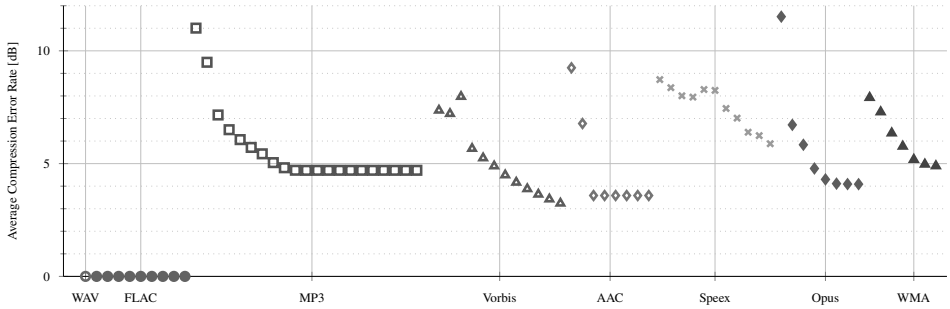


Figure 3 - Average compression error rate for each codec and bit-rate. The bit-rate is increasing from left to right, see Table 2.

In Fig. 3 the average compression error rate (in dB) calculated according to Eq. (2) for the different investigated codecs is depicted. As expected, FLAC has no error. Further, high compression rates result in higher errors than low compression rates. It is apparent that for MP3, AAC, WMA, Opus, and partly Vorbis the error rate reaches a saturation for higher compression ratios. We assume that a certain share of the error cannot be reduced, due to the codec algorithms specialized for music data.

Also, our compression error rate supports previous findings from listening evaluations [4, 15, 18]: Vorbis, ACC and Opus achieve better results than MP3. WMA falls behind MP3. Speex as a specialized codec has an error clearly above the best codecs, but at very low bit-rates, as it can be seen in Fig. 4.

4.4 Compression Ratio vs. Compression Error Rate

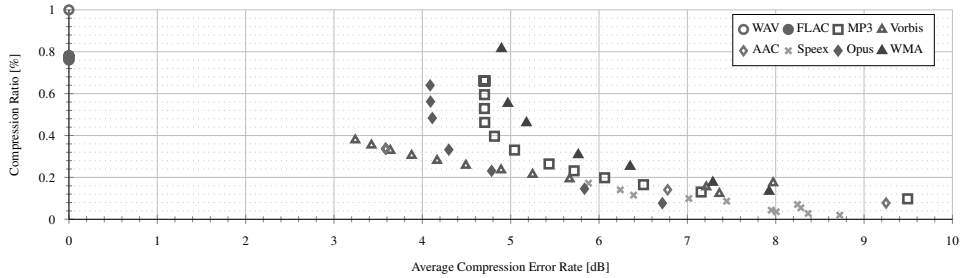


Figure 4 - Average compression ratio over average compression error rate for each codec and bit-rate.

When comparing the compression ratio and the error rate (cf. Fig. 4), it can be seen that the new generation codecs (AAC, Opus, Vorbis) improve both aspects, the compression ratio and the compression error rate. Comparing MP3 and WMA as direct competitors, MP3 outperformes WMA in both aspects. Speex achieves the highest compression ratio for the same error rate as other codecs. Thus, the specialization on speech can be seen. But the resulting error rate is too big for the purpose of high quality data storage.

FLAC does not produce any error, but the compression ratio is only about 76.04%. The best compression ratio/error rate is surprisingly achieved by Vorbis. Using the highest quality setting results in a compression ratio of 37.84% causing only an average compression error rate of 3.24dB.

5 Conclusion and Outlook

In this paper, we analyzed the impact of audio compression on the frequency spectrum using different codecs and bit-rates. Our results confirm previous investigations on the compression quality of the different codecs. We defined the compression error rate, to specify the compression error. Concluding our investigations, we recommend to use FLAC for all cases where the accuracy matters. In cases where a slight error is acceptable, we recommend Vorbis at 500 kbit/s. most of the investigated codecs use perceptual coding, the next step is to investigate the impact of compression on the feature extraction and automatic speech and emotion recognition further on works as e.g. [3]. This will be work for future investigations.

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).

References

[1] ANAGNOSTOPOULOS, T. and C. SKOURLAS: *Ensemble Majority Voting Classifier for Speech Emotion Recognition and Prediction*. Journal of Systems and Information Technology, 16, 02 2014.

[2] APPLE INC.: *Audio Priming - Handling Encoder Delay in AAC*. Mac Developer Library, 2015.

[3] BESACIER, L., C. BERGAMINI, D. VAUFREYDAZ and E. CASTELLI: *The effect of speech and audio compression on speech recognition performance*. In *IEEE Fourth Workshop on Multimedia Signal Processing*, pp. 301–306, 2001.

[4] BRANDENBURG, K.: *MP3 and AAC Explained*. In *17th International Conference: High-Quality Audio Coding*, Aug 1999.

[5] 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.

[6] COALSON, J. and X. FOUNDATION: *FLAC – Free Lossless Audio Codec*, 2014.

[7] CORPORATION, I. and M. CORPORATION: *Multimedia Programming Interface and Data Specifications 1.0*. Techn. Rep., August 1991.

[8] DENG, L. and D. YU: *Deep Learning: Methods and Applications*. Foundations and Trends in Signal Processing, 7:197–387, 2014.

[9] EPPINGER, B. and E. HERTER: *Sprachverarbeitung*. Carl-Hanser-Verlag, Munich, Germany, 1993.

[10] FFMPEG: *General Documentation. Supported File Formats, Codecs or Features.*. Techn. Rep., ffmpeg.org, January 2016. Version 2.8.5.

[11] GLÜGE, S., R. BÖCK and A. WENDEMUTH: *Segmented-Memory Recurrent Neural Networks versus Hidden Markov Models in Emotion Recognition from Speech*. In *Proc. of the 3rd IJCCI*, pp. 308–315, Paris, France, 2011.

[12] ISO: *Information technology – Generic coding of moving pictures and associated audio information – Part 7: Advanced Audio Coding (AAC)*. ISO 13818-7:1997, International Organization for Standardization, Geneva, Switzerland, 1997.

[13] ITU-T: *Perceptual objective listening quality assessment*. Recommendation P.863, International Telecommunication Union, Geneva, 2014.

[14] MOFFITT, J.: *Ogg Vorbis – Open, Free Audio – Set Your Media Free*. Linux Journal, 2001(81es), Jan. 2001.

[15] ROGOWSKA, A.: *Audibility of Lossy Compressed Musical Instrument Tones*. In *Audio Engineering Society Convention 138*, May 2015.

[16] SON, R. J. J. H. VAN: *A Study of Pitch, Formant, and Spectral Estimation Errors Introduced by Three Lossy Speech Compression Algorithms*. Acta Acustica united with Acustica, 91(4):771–778, 2005.

[17] TAYLOR, M.: *LAME Technical FAQ*. Techn. Rep., 2009.

[18] VALIN, J., K. VOS and T. TERRIBERRY: *Definition of the Opus Audio Codec*. RFC 6716, September 2012.

[19] XIPH.ORG FOUNDATION: *Speex: A free codec for free speech*, 2014.