The Effect of MP3 Compression on Automatic Voice Comparison

Timo Becker[1], *Franz Broß*[2] *and Torsten Meier*[2]
[1]*Federal Criminal Police Office, Germany*
timo.becker@bka.bund.de
[2]*University of Applied Sciences Koblenz, Germany*
bross-franz@t-online, tmeier@fh-koblenz.de

MP3 is a lossy compression algorithm for audio recordings which is used mainly for the purpose of saving disc space or reducing transmission bandwidth. When used in forensic applications, MP3 imposes degradations on the audio file quality which sometimes can be heard or seen in a spectrogram. MP3 can be used with several parameters which specify the degree of compression and hence the degree of degradation of the audio signal. For the application of automatic voice comparison systems, it is necessary to know under which circumstances compression by MP3 affects system performance.

In an experiment, recordings of 102 male Romanian speakers recorded in a spontaneous speaking style were used. The single channel recordings are stored with 8 KHz sampling rate and 16 bit. For every speaker there exists a test and a training recording with an average net duration of 95 seconds. While test recordings represent simulated questioned recordings of offenders, training recordings represent simulated known recordings of suspects. The original studio recordings were compressed using the LAME encoder (http://lame.sourceforge.net) with the constant bitrates 8, 16 and 32 kbit/s. Files were encoded with the -h option for high quality encoding. The default 8 kbit/s encoding based on LAME uses a 3 kHz low pass filter to give more bits to the lower frequencies and to avoid artifacts (http://lame.sourceforge.net). For comparison reasons, additional 8 kbit/s encoding was performed where the low pass filtering was omitted. This was accomplished by using the -k option.

Together with the unaffected studio recordings there exist five different versions of each recording. All five recordings were used to compute likelihood ratio estimates with the automatic voice comparison system SPES which is used by the Federal Criminal Police Office of Germany (Becker et al. 2010). The system is designed for forensic voice comparison tasks and hence is based on recordings from real cases, but it was not adjusted to the special settings of the experiment in the present case. Features are analysed in the telephone frequency band only and the background data was kept constant for all tests. The five conditions were combined so that 5 x 5 = 25 different conditions were investigated.

Note that the evaluation corpus does not reflect a typical forensic setting since the language and the recording conditions do not match the required conditions for an analysis by SPES. Because of this, system validity was not evaluated. The discrimination performance of the system was investigated by measurement of the ROCCH-Equal Error Rate (Brümmer 2010). The results are shown in table 1. Note that the results do not reflect system performance in forensic settings, but allow comparison of the different compression rates only.

Table 1. ROCCH-Equal Error Rates for different MP3 compression rates and studio recordings (8 kbit/s processing without low pass filtering is marked by the –k option in parenthesis)

		training recordings				
		8 kbit/s	**8 kbit/s (-k)**	**16 kbit/s**	**32 kbit/s**	**studio**
test recordings	**8 kbit/s**	0.052	0.057	0.027	0.025	0.025
	8 kbit/s (-k)	0.065	0.071	0.032	0.029	0.029
	16 kbit/s	0.031	0.035	0.024	0.022	0.023
	32 kbit/s	0.028	0.032	0.023	0.025	0.025
	studio	0.029	0.032	0.025	0.027	0.030

The performance is worst when using 8 kbit/s recordings for both test and training recordings. For these recordings, the default processing including the 3 KHz low pass filtering shows better performance than the processing where the low pass filtering was omitted (-k option). When not using 8 kbit/s compression rates (grey shaded area in table 1), the ROOCH-Equal Error Rate is 3% or less. Interestingly, here, results for 16 and 32 kbit/s recordings show slightly better discrimination performance than results for studio recordings only. This includes the mixed conditions where studio recordings were compared with compressed recordings. However, the error rate estimates are based on a small evaluation corpus, so small differences should be treated with care.

The results show that the discrimination performance of the automatic voice comparison system SPES is only degraded substantially when using 8 kbit/s constant bit rate compression for both test and training recordings. However, these results are based on studio recordings and do not allow any conclusion about the application of successive compression algorithms. Also, there is no possibility to reliably determine the compression history of a recording after the fact when analysing forensic recordings. Because of this, an analysis should not be conducted whenever there is a notion of the involvement of an 8 kbit/s MP3 compressed recording unless the automatic voice comparison system can be adapted to such a scenario.

Other lossy compression algorithms and numerous combinations of successive application exist. Together with the development of algorithms and the distribution of lossy compression file formats (which is mainly based on popularity), the most common ones should be analysed because it can be expected to find them in forensic casework. This paper shows the influence of the popular MP3 compression on automatic voice comparison. More parameter settings, other lossy compression algorithms (like e.g. the AMR file format) and also frequent combinations of compression algorithms (like e.g. recordings of GSM-transmitted speech which have been compressed by MP3) still need to be investigated to estimate their influence in forensic applications.

References

Becker, Timo, Jessen, Michael, Alsbach, Sebastian, Broß, Franz und Meier, Torsten (2010). SPES: The BKA Forensic Automatic Voice Comparison System. Proceedings of Odyssey: The Speaker and Language Recognition Workshop, 58-62.

Brümmer, Niko (2010). Measuring, refining and calibrating speaker and language information extracted from speech. Dissertation. University of Stellenbosch