The Effects of mp3 Compression on Acoustic Measurements

Robert Fuchs and Olga Maxwell

Englisches Seminar, Westfälische Wilhelms-Universität Münster, Germany
robert.fuchs@uni-muenster.de

Abstract

Recordings for acoustic research should ideally be made in a lossless format. However, in some cases pre-existing data may be available in a lossy format such as mp3, prompting the question in how far this compromises the accuracy of acoustic measurements. In order to answer this question, we compressed 10 recordings of read speech in variable compression rates (16-320 kbps), and reconverted them to wav in order to examine the effect of compression on the commonly used suprasegmental measures of (fo), pitch range and pitch level.

Results suggest that at compression rates between 56 and 320 kbps, measures of fundamental frequency and most measures of pitch range and level remain reliable, with mean errors below 2 % and often better than that. The skewness of the distribution of fo measurements, however, shows much greater measurement errors, with mean errors of 6.9 % - 7.6 % at compression rates between 96 kbps and 320 kbps, and 44.8 % at 16 kbps. We conclude that mp3 compressed recordings can be subjected to the acoustic measurements tested here. Nevertheless, the indeterminacy added by mp3 compression needs to be taken into account when interpreting measurements.

Index Terms: Acoustic measurements, compression, lossy format, mp3, fundamental frequency, pitch range, pitch level

1. Introduction

Sociophonetic researchers as well as forensic and clinical practitioners are frequently faced with the task of measuring the acoustic properties of speech recordings in order to make empirical and exact comparisons between speakers. If recordings are made specifically for this purpose, modern computer technology allows researchers to gather large amounts of data with a sampling rate of 44.1 kHz and store them as wav files. However, pre-existing data, perhaps gathered for other purposes or with an aim to reduce data storage space, might only be available in compressed, i.e. lossy, formats such as mp3. For example, the International Dialects of English Archive (IDEA, http://www.dialectsarchive.com/) offers more than 1,000 recordings of speakers of English from almost 100 countries, with the primary aim of familiarising actors and business people with accents that they need to study. IDEA could potentially be a valuable resource for sociophonetic research, except for the fact that it is not known in how far the fact that all recordings are stored in the mp3 files influences acoustic measurements and whether the quality of the audio recording will be adequate for a fine-grained acoustic analysis.

It is therefore imperative to determine whether, and to what degree, the acoustic reduction processes effective in mp3 compression influence and perhaps compromise the accuracy of acoustic measurements of speech. The present paper addresses this question by evaluating the accuracy of measurements of intonation at various compression rates in mp3 audio files. Fundamental frequency and related measures are of particular relevance, given cross-linguistic and cross-dialectal differences in the manipulation of fo and its phonetic gradience that can be used to signal sentence modality or information structure [1,2].

2. Audio Compression

This section briefly points out how mp3 and other compression algorithms work, and what this could mean for acoustic measurements.

3. Previous Research

The question of how lossy formats of audio storage influence acoustic measurements and auditory evaluations has previously been addressed by a small number of studies [3–8]. All studies agree that acoustic measurements are affected to some degree, but draw varying conclusions. For example, [3] found that Ogg Vorbis compression at 40 and 80 kbps and mp3 at 192 kbps affected his data in two ways: (1) Large differences of more than 9 semitones(st) occurred in less than 3 % of all vowels for fo, 1 % or less for F1 and F2, and less than 0.3 % for F3 in all conditions. (2) After removing large jumps from the data, mean error in st was smaller than 0.7 st for fo, and ranged from 1 to 0.3 st for F1, F2, F3 and CoG, with higher errors for more extreme compression rates. Notably, a change in microphone produced more jumps in F1, F2 and F3, and, after removal of these from the data, higher mean errors in F1, F2, F3 and CoG than compression. On this basis, [3] suggests that the use of audio data derived from mp3 files is relatively unproblematic.

By contrast, [4] found that F1 was raised, F2 was raised for front vowels and lowered for back vowels, and F3 and F4 were also altered, leading the authors to warn against the use of acoustic data from this source. However, since the audio data for this study was first recorded with a hand-held camera, then uploaded to Youtube, and finally downloaded as an mp3 file, it is not clear at what stage the acoustic properties of the recordings were altered. Moreover, [8] found that jitter and shimmer measurements were affected to such a degree that differences between normal and pathological speech that were significant in the original recordings were obscured in the mp3 condition (encoded at 128 kbps).

In summary, while it is clear that mp3 compression influences various acoustic measurements, there is conflicting evidence on whether they are compromised to such a degree that their use is inadvisable. When faced with the question of whether a specific mp3 file can be used for a particular analysis, they can only make an informed decision if data for the influence of mp3 compression on various acoustic measures and at various compression rates is available. Previous research clearly does not permit such evaluations at the moment. In order to make a contribution towards reaching this goal, we will in-
investigate measures of intonation at seven different compression rates.

4. Data and Methods

4.1. Data and Elicitation Methods

The analysis relies on recordings of a text passage read by 10 male speakers of BrE from the DyViS database [9]. The speakers were between 18 and 25 years old at the time of recording (2005-2009), which took place in a sound-treated studio.

4.2. Analysis

The recordings were compressed in Audacity with the LAME mp3 encoder [10] into mp3 files at seven compression rates (16, 32, 56, 96, 128, 256, 320 kbps) and again decompressed and saved as wav files. These were compared to the original wav recordings, saved in 44.1 kHz 32 bit quality.

Approximately two thirds of the reading passage (392 words) were segmented based on phonemic forced alignment with HTK [11] and P2FA [12]. All annotations were manually corrected in Praat [13]. Subsequently, a Praat script extracted all $f_0$ points from a Praat pitch tier object as long as they occurred within a stretch of the recording annotated as an utterance.\(^1\)

The $f_0$ data was then compared for identical segments across the mp3 and wav conditions. We define the absolute error $e_{abs}$ as $e_{abs} = |m_{j,x} - m_{j,o}|$, where $m_{j,x}$ is a measurement of segment $j$ in an mp3 file at compression rate $x$, and $m_{j,o}$ is a measurement of the identical segment in the original file. The relative error is defined as $e_{rel} = \frac{|m_{j,x} - m_{j,o}|}{m_{j,o}} \times 100$. A relative error of 20 means that measurements in the mp3 condition are on average 20 % higher or lower than in the original recording, but does not indicate the direction of the errors (i.e. show a negative or positive skew). For the $f_0$, $m_{j,x}$ was computed in three versions: (1) as the mean of all $f_0$ points in segment $x$, (2) as the maximum and (3) as the minimum of any $f_0$ point in segment $x$.

Next, mixed effects regression models were run in R with lme4 [17, 18]. The measurement error for each of the acoustic measurements (mean $f_0$, max. $f_0$, min. $f_0$, PDQ, 80% - range, mean $f_0$, median $f_0$, skewness of mean $f_0$ distribution) was used in turn as the dependent variable in a regression model, with bitrate as fixed factor and speaker as random factor. To ensure that conditions for regression models were met, we trimmed datapoints 2.5 standard deviations below or above the mean with function rnorm from the package lmerConvenienceFunctions (never more than 2% of the data; this step was skipped for measures of pitch range and level) [19]. Finally, post-hoc Tukey tests with alpha-level corrected for multiple comparisons with the glht function from package multcomp [20] were conducted.

5. Results

5.1. Fundamental Frequency ($f_0$)

As Fig. 1 shows, mean absolute error in mean $f_0$ is slightly and significantly higher in obstruents (1.5 Hz in the 320 kbps condition) than in (0.7 Hz, $z=55.5$, $p<0.001$) and sonorants (0.6 Hz, $z=49.8$, $z=4.9$, $p<0.001$). This corresponds to a difference of

---

\(^1\)The pitch tier object was derived with a 0.01 s time step, and 75 and 300 Hz as min. and max. $f_0$, respectively.
1.1, 0.6 and 0.5 %, respectively. Results are nearly identical for other compression rates up to 96 kbps. At 56 kbps, the error is somewhat larger and differs significantly from the 256 and 320 kbps conditions (z=3.1, z=3.0, p<0.05). The error is yet larger, and differs significantly from all other conditions, at 32 kbps (z>19.0, p<0.001) and 16 kbps (z>18.1, p<0.001). At 16 kbps, mean error amounts to 6.5 Hz or 5.7 % for obstruents, 1.1 Hz or 0.9 % for vowels and 0.8 Hz or 0.7 % for sonorants.

Results are similar for the mean absolute error in maximum and minimum \( f_0 \). For maximum \( f_0 \), the error is larger in obstruents (mean error 2.0 Hz in the 320 kbps condition) than in vowels (1.1 Hz, z=33.5, p<0.001), where it is in turn larger than in sonorants (0.8 Hz, z=4.3, p<0.001). This corresponds to an error of 1.5, 0.8 and 0.6 %, respectively. At compression rates between 56 and 320 kbps the error is relatively constant, while it is significantly larger at 32 and 16 kbps (at 16 kbps 9.1, 1.7, and 1.2 %, respectively, for obs, vow, and son; z>15.8, p<0.001).

For minimum \( f_0 \), the error is again larger in obstruents (mean error 1.5 Hz in the 320 kbps condition) than in vowels (0.9 Hz, z=23.4, p<0.001), where it is in turn larger than in sonorants (0.9 Hz, z=8.8, p<0.001). This corresponds to an error of 1.2, 0.8 and 0.8 %, respectively. At compression rates between 96 and 320 kbps the error is relatively constant. In the 56 kbps condition, the error amounts to 3.2, 1.4 and 1.2 %, respectively, which is significantly larger than in the 128, 256 and 320 kbps conditions (z=3.0, z=3.3, z=3.4, p<0.05). In the 32 and 16 kbps conditions, it is significantly larger than in all other conditions (at 16 kbps 3.2, 1.4, and 1.2 %, respectively, for obs, vow, and son; z>21.4, p<0.001).

An analysis of the mean absolute error in mean \( f_0 \) for specific phonemes (across all compression rates) shows that the error is much greater for certain phonemes than for others. Specifically, glottal stops have a mean absolute error in mean \( f_0 \) of 4.8 Hz, the palatal fricative /ʃ/ of 4.5 Hz, voiceless plosives (/p/, /t/, /k/) of 3.0, 2.9 and 3.5 Hz, respectively. By contrast, other phonemes have smaller errors, such as voiced plosives (/b/, /d/, /ɡ/), with an error 2.2, 1.6 and 2.3 Hz, dental fricatives (/vl/, /vd/) with an error of 1.9 and 1.4 Hz, respectively, and other fricatives (/ʃ/, /ʃ/, /z/) with an error of 2.0, 1.8 and 2.0 Hz, respectively.

5.2. Pitch Range and Level

Mean PDQ, a measure of pitch range, was 0.1656 in the wav condition. At compression rates of 320 to 32 kbps (see Fig. 2a), mean PDQ deviated on average from the wav condition by between 0.0013 or 0.8 % (320 kbps condition) and 0.0029 or 1.7 % (32 kbps, z<0.26, p=1). Only the 16 kbps condition differed significantly from the other compression rates, with a mean error of 0.0394 or 17.6 % (z>5.7, p<0.001).

Mean 80% -range (see Fig. 2b), another measure of pitch range, deviated by up to 0.2 Hz or between 0.4 % in the conditions between 32 and 320 kbps, with no significant differences between these conditions (z<0.2, p=1). At a compression rate of 16 kbps, the measurement error was significantly higher than in the other conditions and amounted to 2.0 Hz or 4.5 % (z>6.3, p<0.001).

Mean \( f_0 \), a measure of pitch level, show very small deviations in the conditions between 320 and 32 kbps (see Fig. 2c), with a mean error of up to 0.1 Hz or 0.1 %, with no significant differences between these conditions (z<0.3, p=1). At 16 kbps, the measurement error is significantly greater and amounts to 1.8 Hz or 1.5 % (z>6.7, p<0.001). Results for median \( f_0 \) are essentially the same, with a measurement error of less than 0.1 Hz and up to 0.1 % in the conditions between 320 and 32 kbps, and a significantly higher error 0.3 Hz/0.4 % at 16 kbps (z>5.4, p<0.001).

The skewness of the distribution of \( f_0 \) measurements shows much greater measurement errors, as Fig. 2e shows. Mean skewness in the wav condition is 1.27. At compression rates of 320 to 96 kbps, the error amounts to 0.082 to 0.105, or 6.9 to 7.6 %, and increases to 0.232 or 16.2 % at 56 kbps, and up to 1.110 or 44.8 % at 16 kbps. Only the latter condition differs significantly from the others (z>6.3, p<0.001).
6. Discussion and Conclusion

This study investigated to what extent acoustic measurements suprasegmental features ($f_0$, pitch range, pitch level) are influenced by the reduction in acoustic information caused by mp3 compression at seven different bitrates (compression strengths). This was evaluated by calculating the difference (“error”) between measurements of mp3 compressed speech data compared to the uncompressed original.

Similar to the findings of [1], greater errors were found for more extreme compression rates. This is intuitively plausible, since employing a more extreme compression rate means that more acoustic information must be discarded in the compression process. However, it is also reassuring for the application of acoustic measures to mp3 compressed data, since it suggests that mp3 compression tends to first discard the kind of acoustic information that is less essential for acoustic measurements taken by speech scientists. The latter point, we contend, is non-trivial, since mp3 compression was not specifically designed for this task. More specifically, the results suggest that where measurement errors differ substantially across compression rates, compression at the most extreme levels of 16 and 32 kbps was more likely to differ significantly in the magnitude of the error from the other conditions.

Looking now at specific acoustic features, the analysis of $f_0$ measurements suggests that compression rates between 56 and 320 kbps show relatively small mean errors of 2 % or less, with median errors well below 0.5 %. The error in measurements of pitch range and level tends to remain well below 1 % at compression rates between 56 and 320 kbps, and below 2 % at 32 kbps. By contrast, pitch range (but not pitch level) measurements at 16 kbps show major deviations. The only exception to the conclusion than pitch range and level measurements show small errors at most compression rates concerns skewness. Even between 96 and 320 kbps, the error ranges between 6 and 8 %. The severity of this problem might perhaps be reduced by excluding extreme outliers, to which skewness is presumably particularly susceptible.

To conclude, the acoustic measures analysed in the present paper appear to remain reliable for audio data in the mp3 format compressed at bit rates between 56 and 320 kbps. However, whenever potential differences between groups or conditions are evaluated, these should be interpreted against the background of the error ranges reported here. In addition, other acoustic measures that underwent compression may differ considerably from the measures investigated above. Future research will examine the accuracy of vowel quality measurements (F1-F4) at various compression rates in mp3 audio files.

7. References


Author/s:
Fuchs, R; Maxwell, O

Title:
The effects of mp3 compression on acoustic measurements of fundamental frequency and pitch range

Date:
2016-01-01

Citation:
Fuchs, R; Maxwell, O, The effects of mp3 compression on acoustic measurements of fundamental frequency and pitch range, Proceedings of the International Conference on Speech Prosody, 2016, 2016-January pp. 523 - 527

Persistent Link:
http://hdl.handle.net/11343/197938

File Description:
Published version