Speech Prosody 2016
31 May - 3 June 2016, Boston, USA

The Effects of mp3 Compression on Acoustic Measurements of Fundamental Frequency and Pitch Range

Robert Fuchs[1] *and Olga Maxwell*[2]

[1]Englisches Seminar, Westfälische Wilhelms-Universität Münster, Germany
[2]La Trobe University and Melbourne University, Australia
robert.fuchs@uni-muenster.de, omaxwell@unimelb.edu.au

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 determine whether this is the case, we compressed 10 recordings of read speech in different compression rates (16-320 kbps), and reconverted them to wav in order to examine the effect of compression on commonly used suprasegmental measures of fundamental frequency (f_0), pitch range and level.

Results suggest that at compression rates between 56 and 320 kbps, measures of f_0 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 f_0 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, intonation

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 in optimal quality, i.e. 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 a view 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. However, it is not known in how far the fact that all recordings are stored in the mp3 format influences acoustic measurements. The quality of such audio recordings might not be adequate for fine-grained acoustic analysis [1].

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 several measurements of fundamental frequency and pitch range and level at various compression rates in mp3 audio files. Fundamental frequency and related prosodic measures are of particular interest, given cross-linguistic and cross-dialectal differences in the realisation of f_0 and its timing and scaling properties that can be used to signal sentence modality or information structure [2–4].

2. Audio Compression

The MPEG1 Layer 3 compression format, commonly known as mp3, is widely used to store large amounts of data. This allows online access and opens up the possibility of sharing large amounts of data in speech corpora. Unlike in lossless speech compression algorithms, mp3 compression does not have a specific source model but instead is based on psychoacoustic coding schemes, whereby the encoder relies on the characteristics of the human auditory system to compress the audio, and removes parts of the signal that are perceptually irrelevant (or less relevant compared to the remaining information) [5–7]. At moderate compression rates, listeners often cannot distinguish between compressed and uncompressed speech [6–8]. Compression can be performed at different bitrates, a number of bits per second measured in kilobits per second (kbps).

The process of mp3 compression consists of the following blocks:

1. the audio is decomposed into subsampled spectral components (time to frequency mapping) through the filter bank;
2. the signal is computed following the perceptual or psychoacoustic model where the spectral components are quantised and coded keeping the noise below the masking threshold;
3. the information is accumulated and processed by the bitstream formatter into the coded stream [6].

The quality of the audio depends on basic parameters such as bitrate and the sophistication of encoders. Generally, a higher quality of the recording is achieved during compression at higher bitrates. However, there is still a lack of information about the degree of distortion of the audio through mp3 compression [7]. The masking effect or allowed noise, for example, raises the noise threshold. This allows compression by reducing the effective dynamic range of the signal. If the encoder is not able to encode the audio at the level of allowed noise determined

10.21437/SpeechProsody.2016-107

by the bitrate, it may lead to a loss of bandwidth. Compression at low bitrates and lower sampling frequencies can result in an audio sounding as if it was recorded twice and overlayed, an effect sometimes called double-speak [6].

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 [1, 5, 8–13]. All studies agree that acoustic measurements are affected to some degree, but draw varying conclusions. For example, [5] found that Ogg Vorbis compression at 40 and 80 kpbs and mp3 compression 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 f_0, 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 f_0, 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. Even after such jumps were removed from the data, higher mean errors in F1, F2, F3 and CoG were still found for the low bitrate compression. On this basis, [5] suggested that the use of audio data derived from mp3 files is relatively unproblematic.

By contrast, [9] found that compressed audio data yielded measurements where F1 was raised, F2 was raised for front vowels and lowered for back vowels, and F3 and F4 were also altered. This led 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. More evidence speaking against the use of (mp3) compressed audio data was presented in [13]. This study 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 potentially 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, an informed decision can be made 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 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 [14]. 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 [15] 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 through phonemic forced alignment with HTK [16] and P2FA [17]. All annotations were manually corrected in Praat [18]. 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] During the analysis, it became clear that short periods of silence had been added to the beginning of the audio data during the en-/decoding process.[2] This was taken into account while the measurements were aligned with the transcription.

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 $e_{rel} = |\frac{m_{j,x} - m_{j,o}}{m_{j,o}}| * 100$. 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 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.

In addition to f_0, we also consider the influence of mp3 compression on pitch range and level. Consistent with previous research, for pitch range, we use (1) the pitch dynamism quotient (pdq), defined as the standard deviation of the f_0 distribution divided by its mean in Hz [19], and (2) 80% range, defined as the difference between the 90th and 10th percentile [2]. For pitch level, we use (3) mean and (4) median [2, 20], and in addition also calculate (5) the skewness of the f_0 distribution [21], which indicates whether extreme values below or above the mean are more common. All measures of pitch range and level are calculated per speaker and based on the total number of pitch points recognised by the Praat pitch algorithm (see above). Depending on the condition, the total number of pitch points recognised for all speakers ranged from 56,031 in the uncompressed wav condition to 55,847 in the 56 kbps condition.

Next, mixed effects regression models were run in R with LME4 [22, 23]. 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 romr.fnc from the package LMERCONVENIENCEFUNCTIONS (never more than 2% of the data; this step was skipped for measures of pitch range and level) [24]. Finally, post-hoc Tukey tests with alpha-level corrected for multiple comparisons with the glht function from package MULTCOMP [25] were conducted.

[1] The pitch tier object was derived with Praat's autocorrelation method (command 'To Manipulation') and the following parameters: time step 0.01 s, min. f_0 75 Hz, max. f_0 300 Hz).

[2] This amounted to 47 ms at a 16 kbps compression rate, 25 ms at 32 kbps and 51 ms at all other compression rates.

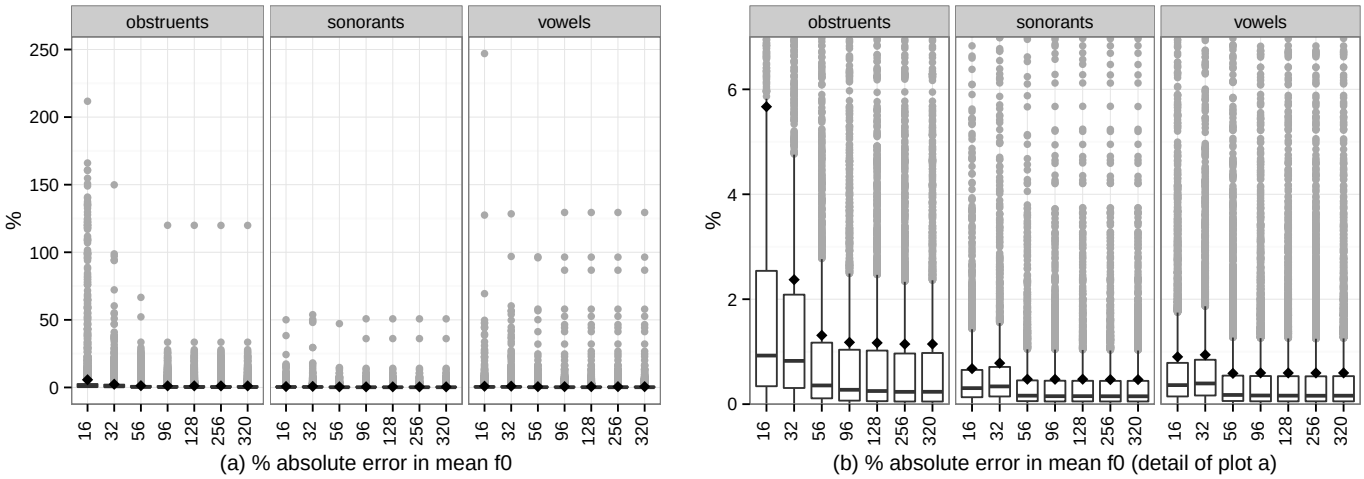

Figure 1: Error in mean f_0 in percent at seven mp3 compression rates. Panel (a) shows the full range of values with all outliers (grey dots), panel (b) shows a detail of panel (a). Diamonds indicate means, boxes extend to the 25th and 75th percentiles, horizontal lines within them indicate medians, and whiskers extend to the highest/lowest point from the box that is within 1.5 times the interquartile range.

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 vowels (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 1.1, 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%, 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%, 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,g/) with an error 2.2, 1.6 and 2.3 Hz, dental fricatives (/θ, ð/) with an error of 1.9 and 1.4 Hz, respectively, and other fricatives (/f,s,z/) with an error of 2.0, 1.8 and 2.0 Hz, respectively.[3]

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 320 and 32 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 signficantly 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

[3] f_0 measurements of phonologically voiceless phonemes are due to co-articulatory voicing carried over from adjacent voiced phonemes.

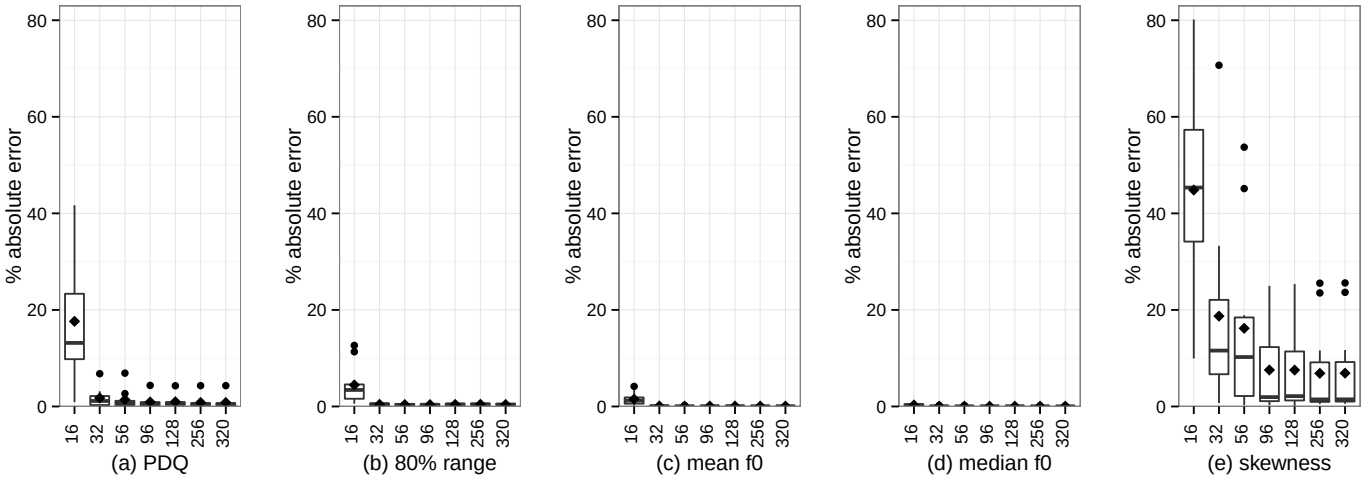

Figure 2: Error in measures of pitch range (panels a and b), pitch level (panels c and d) and skewness of the distribution of pitch values (panel e) at seven mp3 compression rates. Diamonds indicate means, boxes extend to the 25th and 75th percentiles, horizontal lines within them indicate medians, whiskers extend to the highest/lowest point from the box that is within 1.5 times the interquartile range, and points indicate outliers.

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 ($z>6.3$, $p<0.001$).

5.3. A Note Concerning the Decompression of mp3 Data

In the present study, we first **de**compressed the mp3 files before measuring f_0, but Praat can also work directly with mp3 files. A comparison of the results presented above (based on decompressed audio) with measurements derived directly from mp3 showed that the latter option does not lead to systematically better results, and in fact led to a somewhat higher error rate in several cases.

6. Discussion

This study investigated to what extent acoustic measurements of 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 [5], 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 is 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 that 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.

7. Conclusion

The acoustic measures examined in the present paper appear to remain reliable for audio data in the mp3 format compressed at bit rates between 56 and 320 kpbs, and in many cases also at lower bitrates. The findings indicate that mp3 compressed data is viable for the analysis of f_0, with the possible exception of the skewness of the f_0 distribution. However, whenever potential differences between groups or conditions are evaluated, these should be interpreted against the background of the error ranges reported here. Moreover, it is a separate question in how far the results of automatic pitch tracking algorithms match manually corrected measurements of f_0 [26, 27].

For other acoustic measures, we also assume that the effect of compression might be smaller for audio files compressed at higher bit rates, similar to the present results for measurements of f_0. However, the degree of effect for different bit rates may differ considerably from the measures above. Given the reported error differences in mean f_0 measurements for different phonemes, mp3 compression may have a greater effect on the acoustic characteristics of segments, impairing the reliability and robustness of the analysis.

8. References

[1] A. P. Vogel and A. T. Morgan, "Factors affecting the quality of sound recording for speech and voice analysis," *International journal of speech-language pathology*, vol. 11, no. 6, pp. 431–437, 2009.

[2] I. Mennen, F. Schaeffler, and G. Docherty, "Cross-language differences in fundamental frequency range: A comparison of english and german," *The Journal of the Acoustical Society of America*, vol. 131, no. 3, pp. 2249–2260, 2012.

[3] J. Fletcher, E. Grabe, and P. Warren, "Intonational variation in four dialects of english: the high rising tune," in *Prosodic Typology: The Phonology of Intonation and Phrasing*, S.-A. Jun, Ed. Oxford: Oxford University Press, 2005, pp. 390–409.

[4] E. Keane, "Phonetics vs. phonology in tamil wh-questions," in *Proceedings of Speech Prosody, Dresden*, 2006.

[5] R. J. Van Son, "Can standard analysis tools be used on decompressed speech?" in *COCOSDA 2002 Workshop of the International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques*, 2002.

[6] K. Brandenburg, "Mp3 and aac explained," in *Audio Engineering Society Conference: 17th International Conference: High-Quality Audio Coding*, 1999.

[7] J. Gonzalez, T. Cervera, and M. J. Llau, "Acoustic analysis of pathological voices compressed with mpeg system," *Journal of voice*, vol. 17, no. 2, pp. 126–139, 2003.

[8] M. Borský and P. Pollák, "Analysis and automatic recognition of compressed speech," in *Tackling the Complexity in Speech*, O. Niebuhr and R. Skarnitzl, Eds. Prague: Univerzita Karlova v Praze, Filozofická fakulta, 2015, pp. 205–222.

[9] P. De Decker and J. Nycz, "For the record: Which digital media can be used for sociophonetic analysis?" *University of Pennsylvania Working Papers in Linguistics*, vol. 17, no. 2, p. 7, 2011.

[10] B. Rozborski, "A preliminary study on the influence of sound data compression upon formant frequency distributions in vowels and their measurement," *Proceedings of ICPhS XVI, Saarbrucken*, 2007.

[11] G. de Jong, P. Newis, and J. Hunt, "The effects of repeated copying and recording media on intelligibility," *International Journal of Speech Language and the Law*, vol. 9, no. 1, pp. 58–73, 2007.

[12] E. Gold, "The effects of video and voice recorders in cellular phones on vowel formants and fundamental frequency," Master's thesis, Department of Language & Linguistic Science, University of York, 2009.

[13] Y. Zhu, R. E. Witt, J. K. MacCallum, and J. J. Jiang, "Effects of the voice over internet protocol on perturbation analysis of normal and pathological phonation," *Folia Phoniatrica et Logopaedica*, vol. 62, no. 6, pp. 288–296, 2010.

[14] F. Nolan, K. McDougall, G. de Jong, and T. Hudson, "A forensic phonetic study of dynamic sources of variability in speech: The dyvis project," in *Proceedings of the 11th Australasian International Conference on Speech Science and Technology*, P. Warren and C. Watson, Eds., 2006, pp. 13–18.

[15] A. Team. (2012) Audacity(r) version 2.0.0.

[16] S. J. Young, *The HTK Hidden Markov Model Toolkit: Design and Philosophy*, University of Cambridge: Department of Engineering Std., 1993.

[17] J. Yuan and M. Liberman, "Speaker identification on the scotus corpus," in *Proceedings of Acoustics '08*, 2008, pp. 5687–5690.

[18] P. Boersma and D. Weenink, *Praat: Doing Phonetics by Computer (Computer Program). Version 5.4.01*, Std. [Online]. Available: www.praat.org

[19] R. Hincks, "Processing the prosody of oral presentations," in *Proceedings of InSTIL/ICALL2004 Ú NLP and Speech Technologies in Advanced Language Learning Systems*, Venice, 2004.

[20] M. G. Busà and M. Urbani, "A cross linguistic analysis of pitch range in english l1 and l2," in *Online Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong, 2011, pp. 576–579.

[21] D. J. Patterson, "A linguistic approach to pitch range modelling," Ph.D. dissertation, Edinburgh University, 2000.

[22] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008. [Online]. Available: http://www.R-project.org

[23] D. Bates, M. Maechler, and B. Bolker, *lme4: Linear Mixed-Effects Models Using S4 Classes*, 2013, r package version 0.999999-2. [Online]. Available: http://CRAN.R-project.org/package=lme4

[24] A. Tremblay, D. University, J. Ransijn, and U. of Copenhagen, *LMERConvenienceFunctions: Model Selection and Post-hoc Analysis for (G)LMER Models*, 2015, r package version 2.10.

[25] T. Hothorn, F. Bretz, and P. Westfall, "Simultaneous inference in general parametric models," *Biometrical Journal*, vol. 50, no. 3, pp. 346–363, 2008.

[26] K. Murray, "A study of automatic pitch tracker doubling/halving errors," in *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, 2001.

[27] P. Martin, "Multi methods pitch tracking," in *Proceedings of Speech Prosody 2012*, Shanghai, 2012.