CAN STANDARD ANALYSIS TOOLS BE USED ON DECOMPRESSED SPEECH?

R.J.J.H. van Son[*]
Institute of Phonetic Sciences/ACLC
University of Amsterdam
Herengracht 338, 1016CG Amsterdam
Rob.van.Son@hum.uva.nl

Abstract

This paper quantifies some of the effects of "lossy" audio compression on basic acoustic speech analysis procedures by comparing original audio-CD speech recordings to compressed/decompressed versions of these recordings. The differences found are benchmarked against the effects of a change of microphone. Tested are a Sony Minidisc Walkman recorder and two audio compression codecs: Ogg Vorbis 1.0rc3 and LAME 3.92 (MP3), with 3 bit rates: 40, 80 (Ogg Vorbis), and 192 kbs (MP3). These are tested against pitch and formant extraction and spectral center of gravity (i.e., first spectral moment). Audio compression added only a limited amount of "jump errors" (≤3%) to vowel pitch and formant measurements. Only small systematic effects on measurements were found that could be attributed to compression. However, rather large systematic effects resulted from a switch of microphone, mostly on the spectral center of gravity. The audio compression algorithms introduced a Root-Mean-Square (RMS) error, after removing jump errors, of less than 1 semitone to vowel mid-point pitch, formant, and CoG measurements. The effect of the microphone change on RMS error was as large, i.e., for pitch, or larger, i.e., >1.2 semitones for formants and center of gravity. Comparison of the pitch in sonorants and the spectral center of gravity measurements in continuants showed that here too, the RMS errors introduced by the audio compression were always less than 1 semitone, except for the lowest bit-rate, 40 kbs, where CoG errors exploded in vowel-like consonants and fricatives (> 2 semitones). The size of the errors shows an effect of compression factor (bit-rate). The higher bit-rate encodings always had smaller RMS errors, except for pitch measurements where there was no effect of encoding or bit-rate whatsoever. When audio compression is applied repeatedly, e.g., during recording, distribution, and archiving, the weakest link determines the total RMS error for pitch and formant measurements. However, the total RMS error of the CoG measurements is the sum of the component errors. It is concluded that Minidisc recordings and compressed speech of bit-rates from 80 kbs and up can be used for acoustic speech analysis if an increased RMS error of up to 1 semitone is acceptable. A low bit-rate encoding of 40 kbs introduces markedly larger errors in formant measurements and must be considered unsuitable for whole-spectrum measurements like the CoG. Repeatedly compressed speech is still useful for pitch and formant measurments, but whole spectrum (e.g., CoG) measurements should only be used with care.

1. Introduction

High quality "lossy" audio compression has revolutionized the distribution and storage of music. It could do the same for speech corpora. What is slowing down this revolution is uncertainty about the reliability of speech analysis tools when used on decompressed speech. Potential users fear that compression could introduce artifacts or biases in acoustic analysis results. This paper will try to quantify the effects of three popular compression algorithms with respect to basic analysis types: Pitch and Formant extraction and spectral center of gravity (CoG, first spectral moment) determination.

A range of software has become available that can compress sound recordings efficiently at very high perceptual quality. Cheap, robust, and light-weight equipment based on this software is now available for making high quality recordings. The best known device today is the Sony Minidisc. But similar high-quality devices based on MP3 or Ogg Vorbis compression standards are becoming available. As a result, many projects that record speech in natural situations have used, or plan to use, Sony Minidisc equipment to record speech. Examples are the Spoken Dutch Corpus (CGN, Oostdijk et al., 2002; c.f., http://lands.let.kun.nl/cgn/ehome.htm) and the collection of expressive speech by the Japanese JST/CREST ESP Project (c.f., Campbell, 2002a & b). The Sony Minidisc is an almost ideal device for such "field" recordings. It is cheap, small, light-weighted, and can run on standard batteries. Moreover, it can be carried around and operated by volunteers in everyday situations without the need for technical assistance.

The compression algorithm used in the Sony Minidisc, ATRAC3, is one of a class of lossy compression algorithms that remove redundant information from the sound spectrum according to a psycho-acoustic model of human hearing. Two other popular standards in this class of algorithms are MP3 (Mpeg-1 layer 3) and Ogg Vorbis. Perceptually, the decompressed audio of all these algorithms is of very high quality. Naive listeners are often unable to hear the difference between decompressed and original sounds. However, common speech analysis algorithms are *not* based on a model of human speech recognition, but on a (simplified acoustic) model of speech *production*. Therefore, it is very well possible that the compression algorithms remove spectro-temporal information that is irrelevant for speech perception, but

important for analysis. The structure of the compression algorithms is such that their effects on specific classes of sounds cannot be predicted reliably from the specifications. An evaluation needs real examples.

Current speech copora can contain ~1000 hours of speech, corresponding to ~100GB of data (at 16 kHz sampling). Such a corpus would need up to 200 CD-ROMS or 20 DVD's for distribution. Widespread distribution over the internet of such corpora is currently out of the question. The construction and distribution of these corpora would benefit enormously if speech could be recorded and distributed in a compressed format without losing its (acoustic) usefulness. Recently, Campbell (2002a and b) found that the differences between uncompressed DAT and compressed/decompressed Sony Minidisc recordings were clearly visible in the spectrum. But he concluded that the consequences for the determination of pitch, formants, and glottal parameters seemed to be limited. Gonzalez and Cervera (2001) and Gonzales et al. (in preparation) did an in-depth evaluation of the effects of MP3 compression on the LTAS and on 29 voice quality parameters, using specialist equipment, at different encoding bit-rates. They found that MP3 compression at bit-rates of 96 kbs and higher did not affect the fidelity of neither the LTAS nor the voice parameters. At bit rates of 64 kbs and lower, some of their parameters became distorted and differences became visible in the LTAS below 12 kHz. However, there is currently still little quantitative information about the effects of different compression methods on standard pitch, formant, and whole spectrum measurements from large samples of running speech.

The current paper tries to quantify the differences in measurements between original audio-CD quality recordings and compressed/decompressed speech. This study is limited to the measurement of Pitch, Formants, and spectral Center of Gravity (first spectral moment) using standard analysis algorithms from the *praat* program. Three compression algorithms were tested: a Sony Minidisc Walkman device, MP3, and Ogg Vorbis. To evaluate the importance of any observed differences, a yardstick of acceptable differences is needed. Pitch and formant extraction algorithms are infamous for their, more or less random, errors. Therefore, a considerable amount of differences is expected in the analysis of any two separate recordings of even the same utterance. For the current evaluation, the differences in measurement values that result from switching from a fixed condenser microphone to a head-mounted dynamic microphone are used as a yardstick. The choice of microphone varies widely between projects but any effect of this choice on the measured quantities can be considered a measurement error. Therefore, the difference due to microphone choice indicates the minimum uncertainty already present in the analysis. If the effects of compression are small compared to the effects of changing microphones, they can be considered acceptable for most applications.

2. Methods

2.1 Speech Material

125 segmented and labeled sentences were selected from the Dutch IFA corpus (Van Son et al., 2001). These were all segmented and labeled recordings of readings of a Dutch version of "The North wind and the Sun" and retold versions of this story. Recordings of all 8 speakers, 4 male and 4 female, in the IFAcorpus were used. The number of sentences varied per speaker as not all sentences in the original recordings were segmented and labeled. The ages of the speakers ranged from 15 to 66 years-of-age.

All sentences were recorded directly on audio-CD with a fixed HF condenser microphone (Sennheiser MKH 105) and a head-mounted dynamic microphone (Shure SM10A) in parallel (2-channel recordings). Both versions were used in this comparison, bringing the total number of files to 250. In this paper we will concentrate on the results for the HF condenser microphone. However, we will note the differences with the dynamic microphone in the text.

2.2 Compression/decompression

2.2.1 Sony Minidisc Walkman

Sentences were recorded digitally, using the optical line-in plug, from the computer to a Sony Walkman MZ-R909 which uses the ATRAC3 compression algorithm. Recordings were done in Mono at the standard (unspecified) bit-rate. Sentences were played to the Sony Walkman with short pauses between the individual files. Each sound file was preceded and followed by a 200 ms first order gamma tone of 1 or 2 kHz (after and before, respectively). These "beep" sounds were later used to automatically extract the target sentences from the recordings. As each target sentence is embedded in at least 500 ms of contextual speech, surrounding the sentences by gamma-tones will not influence the encoding of the target sentences by the recorder. As the MZ-R909 has no digital line-out (c.f., Campbell, 2002a), the Sony Minidisc Walkman recordings were played back in analogue form directly to a Philips CD-audio recorder (i.e., 44.1 kHz and 16 bit linear). This audio cd recording was then transfered to the computer in digital form. Two special *praat* scripts segmented the complete recording on the gamma-tones and aligned the fragments to the original sound files (see sections 2.3, 7 and Appendix for resources).

2.2.2. Compression software

All manipulations were done on a PC running Linux (Red Hat 7.2). All sentences were compressed and decompressed with the LAME 3.92 MP3 codec and the Ogg Vorbis 1.0rc3 codec. Programs for both codecs

were downloaded from the internet shortly before conversions (June 2002, see section 7). The aim was to determine whether standard compression software could be used safely for field recordings, storage, and distribution of recordings. Given the sheer complexity and number of options for each codec, only intuitively obvious standard options were considered. For both programs the default settings were used, except that for MP3 the predefined high quality option was used (`--preset cd` for LAME) and two sets of Ogg Vorbis encodings were used, the default and one with a low average bit-rate of 40 kbs (`-b 40`). This resulted in a 192 kbs constant bit-rate encoding for the LAME codec and two, approximately 80 kbs and 40 kbs, variable bit rate encodings for the Ogg Vorbis codec. These settings corresponded to actual compression factors of 3.46, 8.30, and 15.54, respectively (i.e., the effective bit rates were: 204, 85, and 45 kbs, respectively).

The command lines used for compression/decompression were (Linux Bash shell):
LAME 3.92 (using the `notlame` program)

```
>notlame -h -m m --preset cd --quiet <filenameIn> - \
 | notlame -h --silent --mp3input --decode - <filenameOut>
```

Ogg Vorbis 1.0rc3 (using the `oggenc` and `ogg123` programs)

80 kbs

```
>oggenc -Q <filenameIn> -o - | ogg123 -q -d wav -f <filenameOut> -
```

40 kbs

```
>oggenc -b 40 -Q <filenameIn> -o - | ogg123 -q -d wav -f <filenameOut> -
```

2.3 Alignment

To be able to compare original and decompressed recordings, it is paramount that the files are aligned as accurately as possible. Therefore, for all decompressed sentences the alignment with the original sentences was checked. Alignment was tested using *praat* by down-sampling the files to 4 kHz and subsequently performing a cross-correlation between original and decompressed files with lags from -200 ms to +200 ms. The lag of the maximal peak is considered to be the overall time-shift. This same method was used to align the Sony Minidisc Walkman recordings to the original files. For the Ogg Vorbis encoded files (both rates), the absolute lag was always less than 0.01 ms (sd=0.001) ms. For the LAME MP3 encoded speech, the absolute lag was less than 0.27 ms with a systematic average shift of 0.13 ms (sd=0.12ms). It seems that the MP3 decoder added some zero values before the real recordings (this is a documented problem). The Sony Minidisc Walkman recordings showed an absolute lag after alignment of less than 0.6 ms with a systematic average shift of 0.19 ms (sd=0.12 ms).

2.4 Analysis

All speech files (original and decompressed) were analyzed with the *praat* program (v. 4.0.16). For each file, the pitch (F_0), formants (F_1-F_3), and spectral Center of Gravity (CoG, first spectral moment) were calculated using the default settings of *praat*, emulating a "naive" user. However, the analysis window step size was set to 1 ms for better alignment (see appendix). This means that the pitch was calculated using the *autocorrelation* method (*simple* option) and formants using the *Burg* algorithm The CoG was calculated using a special script, not the *praat* command of this name. Measurements were always done at phoneme mid-points. Formants are best defined inside vowels. Therefore, formant values were only measured for vowels.

3. Results

The frequency values for F_0-F_3 and CoG cover a range from 100 Hz to over 3 kHz and their variances scale with these frequencies. To be able to compare differences in a meaningful way, all values were recalculated to semitones ($12*\mathrm{Log}_2(\mathrm{frequency})$). This way, all variances are of the same order. For Sony Minidisc and the 80 and 192 kbs encodings, there were no systematic (average) differences between the pitch and formant analysis results for the original recordings and the compressed speech (<0.04 semitones for vowels). The differences for the CoG were somewhat larger (<0.15 semitones). However, at the low, 40 kbs, bit-rate encoding, there were systematic differences in the order of 0.1 semitones for F2 and F3 and up to 0.5 semitones for CoG. A switch of microphones did induce a consistent shift of <0.2 semitones in the formant values and an even larger shift in the center of gravity values (up to 5 semitones).

3.1 Jump errors

Both pitch and formant extraction are known for the occurrence of "jump" errors where measured pitch and formant tracks contain large discontinuities. For pitch measurements, these jumps are generally octave errors, where the algorithm suddenly selects a different (sub-)harmonic as the F_0. For formant tracks, a spurious formant can be added, e.g., the F_0, or a formant can be missed. As a result, the formants are labeled incorrectly, e.g., F_2 can end up as F_1 or F_3. If the compression algorithms would introduce large numbers of jumps, this would invalidate their use. The spectral center of gravity (spectral first moment) does not show any preference for "jumps" (not shown). Figure 1 gives the number of differences larger than 9 semitones for F_0-F_3 in vowels.

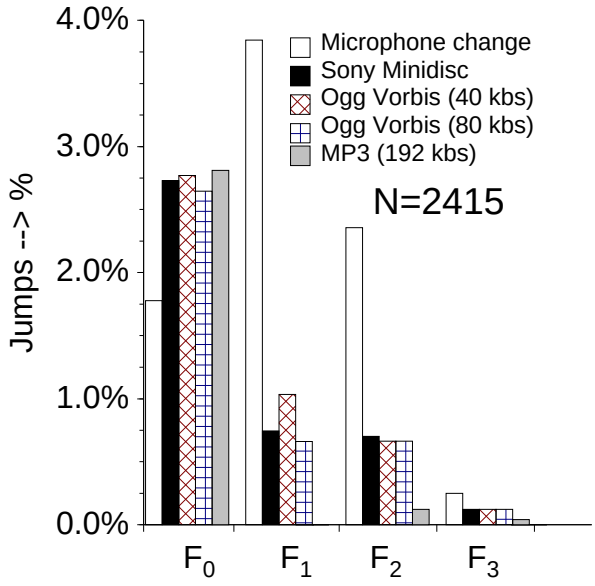

Figure 1. Number of large jumps, i.e., differences > 9 semitones, in the difference between original HF condenser microphone recordings and either dynamic microphone recordings (microphone change) or decompressed recordings. Midpoint values from N=2415 vowel realizations from 8 speakers recorded from text reading and a retold story.

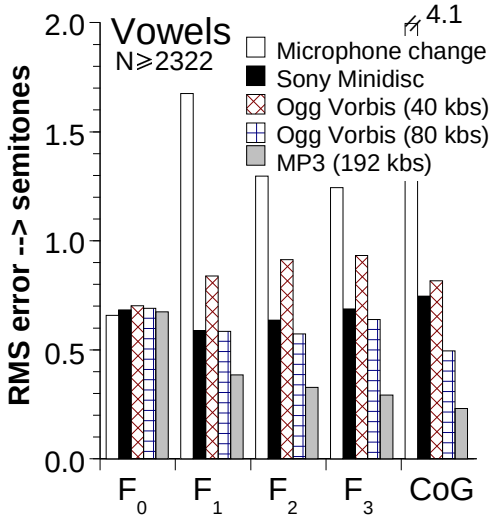

Figure 2. Root-mean-square error (in semitones) introduced to vowel measurements by a change to a dynamic microphone (microphone change) or by audio compression. Large differences (differences > 9 semitones) were excluded. Midpoint values from N=2415 vowel realizations from 8 speakers recorded from text reading and a retold story (N 2322 after removing large differences). CoG is the spectral center of gravity.

This jump size was a compromise between removing outliers that are clearly not on the same harmonic or formant track and losing too many observations. Although we do not know which file contains the erroneous jump, the original or the decompressed, a sizable number of such large differences is unwanted. It is clear that the number of these differences due to the compression is of the same order (F_0 and F_3) or considerably smaller (F_1 and F_2) than that due to a switch of microphone. Note the low number of formant "jump" differences for the high bitrate MP3 encoding.

3.2 Differences between original and decompressed vowels

Systematic differences between original and (de-)compressed speech proved to be very small for pitch and formants, and somewhat larger for the CoG (not shown). There were considerable systematic differences between the microphones. Most studies on phonetics use relative instead of absolute frequency values, and would not be impacted by (small) systematic shifts. Therefore, systematic effects will be ignored here and the standard deviation (spread) of the differences can be interpreted as the *Root-Mean-Square error* (RMS error) introduced by the compression (or microphone change).

When the differences are not corrected for jump errors (see previous section), the RMS error tends to be rather high, around 2 semitones and more. However, pitch and formant jumps are clear errors that have to be corrected or removed, often by hand, to make pitch and formant values useful. Using the pitch smoothing and formant tracking options in *praat* did not improve the situation (not shown). Therefore, it was decided to treat large (>9 semitone) differences in F_0-F_3, but *not* CoG, as outliers and remove them from the calculations. The effect of this on the sample sizes is small (<3%, see Figure 1). However, the effects on the RMS errors are large.

In Figure 2, the RMS error in vowels introduced by a microphone change and audio compression are displayed in semitones, after removing all differences larger than 9 semitones for pitch and formants (but not Center of Gravity). It is clear that 9 semitones constitute more than 10 standard deviations for these vowel measurements (RMS error = standard deviation), so we can indeed consider them as outliers. The effects of using decompressed speech instead of the original recordings introduces an RMS error of around 0.7 semitones for all but the 40 kbs encodings, which corresponds to a 4% difference in linear frequency. The 40 kbs compression introduced larger RMS errors of up to 1 semitone. Except for the F_0, the effect of switching microphones is considerably larger (RMS error > 1.2 semitones, i.e., > 7%).

3.3 Other phonemes

The previous results were obtained for vowels only. Figure 3 shows the RMS error for pitch measurements on broad sonorant classes (excluding "jump" differences). Figure 4 does the same for the CoG for all continuants

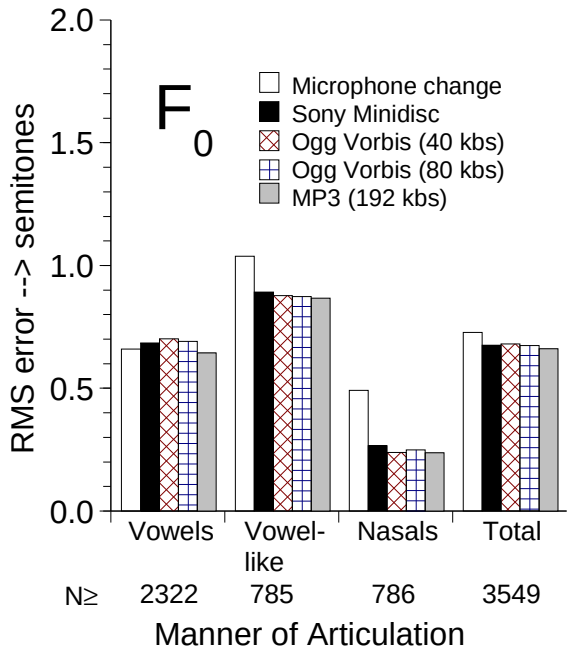

Figure 3. Root-mean-square error (in semitones) introduced to pitch measurements from sonorants by a change to a dynamic microphone (microphone change) or by audio compression. Large differences (differences > 9 semitones) were excluded. Midpoint values from in total N=4063 phoneme realizations from 8 speakers recorded from text reading and a retold story. The data for the vowels are identical to Figure 2.

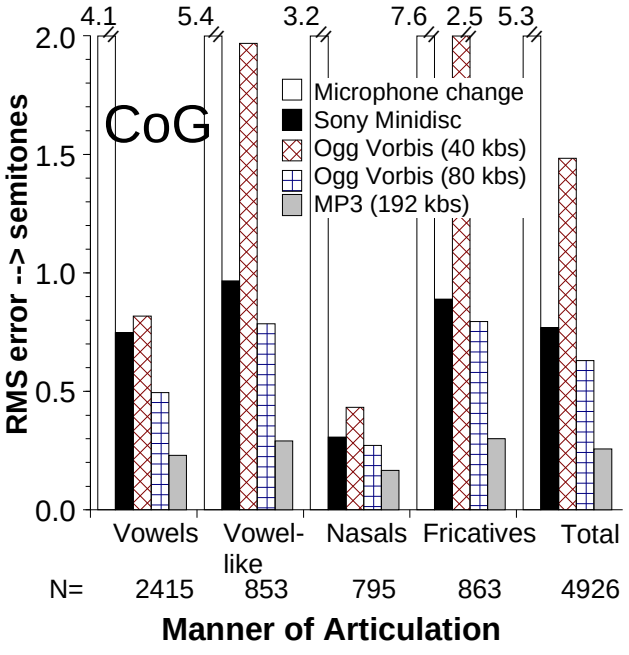

Figure 4. Root-mean-square error (in semitones) introduced to spectral Center of Gravity (first spectral moment) measurements from continuants by a change to a dynamic microphone (microphone change) or by audio compression. Midpoint values from N=4926 phoneme realizations from 8 speakers recorded from text reading and a retold story. The data for the vowels are identical to Figure 2.

(but now including *all* differences). Figure 3 clearly shows that for pitch extraction, all compression algorithms are indistinguishable and switching microphones is only marginally different. The RMS errors found in the low bit-rate (40 kbs) encoding are indistinguishable from those found in the high (192 kbs) encoding.

The picture is different for the CoG. For all phonemes studied, both switching microphones and using low bit-rate encodings introduced large errors in CoG values, i.e., >3 and up to 2.5 semitones, respectively. These large errors in the CoG indicate a large difference in spectral shape, most likely on the high frequency side. On the whole, the RMS error tends to be largest for vowel-like consonants and smallest for nasals (<0.4 semitones for both pitch and CoG).

3.4 Differences between microphones

The spectral characteristics of the HF condenser microphone were better than those of the head-mounted dynamic microphone. The dynamic microphone was considerably less sensitive at low and high frequencies. The RMS error introduced by the compression was systematically larger for the dynamic microphone, by up to 0.2 semitones, for all analysis types. Exceptions were the F_3 measurements, which were indistinguishable, and the center of gravity measurements with the Sony Minidisc which had a considerably larger error (by up to 0.7 semitones).

3.5 Compression cascades

If audio compression is going to be used in the collection and dissemination of corpora, each recording is going to be compressed/decompressed several times during its lifetime. A typical chain of compression would be: recording, transmission/distribution, and archiving. Each step is likely to use different bit-rates and codecs. Figure 5 gives the results from such a cascade of compression. It contains the RMS errors introduced by recording on Sony Minidisc followed by compression/decompression with Ogg Vorbis at 80 kbs and MP3 at 192 kbs.

For The F_0-F_3 analysis, the compression codecs behave almost ideal, the RMS error is indistinguishable from that of the weakest link in the cascade (the Minidisc). However, the CoG displays a worst case behavior. The RMS error of the CoG is the sum of the RMS errors of the three codecs used. This indicates that the codecs are correlated with respect to the CoG. Each codec moves the CoG value of a sound in the same direction as the others.

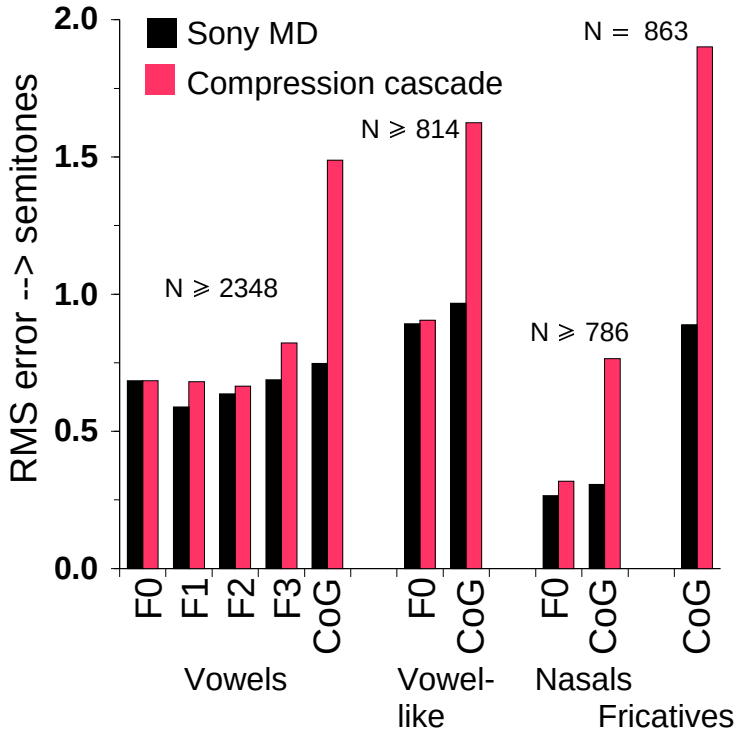

Figure 5. Root-mean-square error (in semitones) resulting from a cascade of three compression codecs: Sony Minidisc recording (ATRAC3), followed by an Ogg Vorbis (80 kbs) and an MP3 (192 kbs) compression compared to only a Sony Minidisc recording (data from figures 2-5). Large differences (differences > 9 semitones) were excluded. Midpoint values from in total N=4926 phoneme realizations from 8 speakers recorded from text reading and a retold story (numbers in graph are for the cascade). The RMS errors for F_0-F_3 are close to the RMS errors found for the weakest link, the Sony Minidisc recording. The RMS errors for the CoG are close to the sum of the RMS errors of the three codecs used.

4. Discussion and Conclusions

Note that this study was *not* intended to be a consumer test *compression benchmark* The choice of recording device or compression codec depends on practical considerations like price, availability, and technical specifications, but also on licensing terms and maintainability. For instance, MP3 codecs contain patented technology that has to be licensed, although the free LAME codec seems to be tolerated for now[§]. The LAME and Ogg Vorbis codecs were chosen for this study because they belong to the best codecs available and are both Open Source. The fact that codecs are Open Source is relevant for "grass-roots" initiatives in small language communities as these often lack the resources needed to obtain and support licensed codecs, as well as for corpus collection as it allows version management on different platforms over the lifetime of the corpus.

Before going into more detail, the limits of this paper must be stressed. This study is based on clean laboratory speech, recorded on audio-CD under quiet conditions. Furthermore, all measurements were done in the mid-points of continuant phonemes. Measurements of speech transients or using noisy recordings remain untested. In transients, e.g., vowel onsets, release bursts, or affricates, the compression algorithms will change the duration (window-length) of the compression frames and reduce the effective bit-rate for spectral modeling. This might be an explanation for the high RMS error in (dynamic) vowel-like phonemes. In a noisy environment, bits will be allocated to modeling the noises faithfully, reducing the bit-rate available for the speech. There are other, more complex effects of noise on the encoding. To conclude, both transients and noisy speech could in principle deteriorate measurements more than would be expected from this paper. This study is only a proof of principle about the use of acoustic analysis on compressed speech, it is *not* exhaustive on the limits of applicability.

The effects of the compression algorithms and the choice of microphone on the acoustic analysis procedures used can be divided into three classes of errors. First, large jump errors can be introduced in the pitch and formant extractions if the compression causes the analysis algorithms used to pick a different (sub-)harmonic or to miscount spectral peaks, respectively. Figure 1 shows that this occurs in less than 3% of all vowels (4% for the microphone switch). For many purposes, this is an acceptable number. Second, a systematic shift in the

§ MP3 decoders can be distributed free of charge, encoders have to be licensed wherever the patent is valid.

measured values can be introduced. The compression codecs introduced no shift in pitch or formant measurements and a small shift (<0.15 semitones) in CoG values. The choice of microphone proved to induce a small shift in the formant values (<0.2 semitones) but a large shift in the CoG values (up to 5 semitones). Third, any change in the spectrum of the sound will add a kind of random noise to the measurements. I will, quite arbitrarily, claim that the original recordings are "correct", and any difference is an error. Figures 2-4 show that the RMS error is consistently less than 1 semitone for pitch, formant, and CoG measurements using recordings from the HF condenser microphone and compression bit-rates of 80 kbs and more. The errors were largest for vowel-like consonants and fricatives (up to 1 semitone), smallest for nasals (<0.3 semitones), and in between for vowels (<0.7 semitones). The effect of the choice of microphone is generally much larger, and never smaller. For the low bit-rate encoding (40 kbs) it showed that there was no difference with higher bit-rates for pitch measurements. However, both formants and, especially, CoG showed larger errors as a result of the low bit-rate compression (up to 1 and 2.5 semitones, respectively).

The picture becomes more complicated when cascades of codecs are used. A very common chain of actions would be: Recording on a Minidisc, transmission/distribution using medium bit-rate compression, and archiving using high bit-rate compression. Figure 5 shows the result of simulating such a cascade. For pitch and formant measurements, the *weakest link* (Sony Minidisc in this example) determines the total RMS error. However, for CoG measurements the total RMS error is the *sum* of the component RMS errors. As the file formats are incompatible, this problem cannot be solved by a direct translation, i.e., without decompression, between encodings. However, even a partial translation, e.g., at the spectral level, would be beneficial. Such partial translaters are not available today, but they might be worthwhile for large speech archives that intend to store speech in a compressed format.

To conclude, from this study it emerges that decompressed speech that has been handled by any of the "lossy" compression algorithms discussed can indeed be used for acoustic analysis. Except for the lowest bit-rate, 40 kbs, the RMS errors introduced by the audio compression were less than 1 semitone, which corresponds to a change of less than 6%. For vowels and nasals the errors were even smaller, with RMS errors of less than 0.7 and 0.3 semitones, respectively (i.e., <4% and <2%). And even at the low bit-rate of 40 kbs, only the CoG measurements were strongly affected. Therefore, for measurements that can tolerate RMS errors of up to 1 semitone, the use of lossy compression would be acceptable. If needed, RMS errors of formant and CoG measurements can be reduced by increasing the bit-rate of the compression. Repeatedly compressed speech can still be used for pitch and formant measurements. However, it should only be used with care for whole spectrum methods like CoG.

5. Acknowledgments

I would like to thank Ton Wempe for his assistance with the Sony Minidisc Walkman recordings. Ton Wempe and David Weenink are together responsible for the design of the core *praat* scripts used to segment the Sony Minidisc Walkman recordings and subsequently align them with the original speech files. I would also like to thank Louis Pols for his comments on the text. This work has been made possible by grant 355-75-001 of the Netherlands Organization Of Research (NWO).

6. References

Campbell, N. (2002a). "Recording and storing of speech data", Proc. of the International Workshop on Resources and Tools in Field Linguistics, Las Palmas.

Campbell, N. (2002b). "Recording techniques for capturing natural every-day speech", in Proc. LREC 2002, Las Palmas.

Gonzalez, J, and Cervera, T. (2001). "The effect of MPEG audio compression on multidimensional set of voice parameters", Logopedics, Phoniatrics, and Vocology 26, 124-138.

Gonzales, J., Cervera, T., and Llau, M.J. (in preparation). "Acoustic analysis of pathological voices with MPEG system".

Oostdijk, N., Goedertier, W., van Eynde, F., Boves, L., Martens, J.P., Moortgat, M. (2002). "Experiences from the Spoken Dutch Corpus project.", in Proc. LREC 2002, Las Palmas.

Van Son, R.J.J.H., Binnenpoorte, D., van den Heuvel, H. and Pols, L.C.W. (2001). "The IFA corpus: a phonemically segmented Dutch Open Source speech database", Proc. EUROSPEECH 2001, Aalborg, Denmark, Vol. 3, 2051-2054.

7. WWW resources

IFAcorpus	http://www.fon.hum.uva.nl/IFAcorpus
praat	http://www.praat.org/
Sony Minidisc	http://www.minidisc.org/
ATRAC3	http://www.sony.co.jp/en/Products/ATRAC3/
Ogg Vorbis codec	http://www.xiph.org/
MP3	http://www.iis.fhg.de/amm/techinf/layer3/
MP3 licensing	http://www.mp3licensing.com/royalty/software.html
LAME codec	http://lame.sourceforge.net/
notlame program	http://hive.me.gu.edu.au/not_lame/
GNU project	http://www.gnu.org/

Appendix: Core *praat* 4.0 scripts

The speech files and scripts, including the perl programs that manage the bulk file conversions, can be found at http://145.18.230.99/corpus/home/Compress.html (also accessible over the official IFAcorpus site). All files are published under the GNU GPL, including the *no warranty* clause.

Core scripts used to generate the analysis files. They assume that the file to analyze is already a selected *praat* object. The result still has to be written to file.

Create pitch tier:

```
Rename... Label
select Sound Label
To Pitch... 0.001 75 600
Rename... Pitch
select Pitch Pitch
Down to PitchTier
Rename... PitchTier
```

Create formant tracks:

```
Rename... Label
select Sound Label
To Formant (burg)... 0.001 5 5500 0.025 50
Rename... Formants
Down to FormantTier
Down to TableOfReal... yes no
Remove column (index)... 6
Remove column (index)... 5
select TableOfReal Formants
```

Create spectral Center of Gravity tiers:

```
Rename... Label
select Sound Label
To Spectrogram... 0.025 16000 0.001 10 Gaussian
select Spectrogram Label
To Matrix
Rename... FE
select Spectrogram Label
Remove
select Matrix FE
Copy... E

select Matrix FE
Formula... if(row>1) then self*y+self[row-1,col] else self*y fi
To Sound (slice)... -1
select Matrix FE
Remove

select Matrix E
Formula... if(row>1) then self+self[row-1,col] else self fi
To Sound (slice)... -1
select Matrix E
Remove

select Sound FE
Rename... CoG
Formula... 12*log2(self/Sound_E[row,col])
select Sound E
Remove

select Sound CoG
```