Spectral implications of codec compression on voiceless fricatives

Krestina V. Christensen

Aarhus University & University of York
kvc@cc.au.dk & kvc507@york.ac.uk

ABSTRACT

This study investigates voiceless fricatives that have been subject to codec compression, as codec compressed speech is commonplace in people's lives, and knowledge is still limited on how the compression affects the acoustics of consonants, here [s, f, fʲ, θ, ʃ] specifically. The paper presents a study of read speech from 30 male speakers of English, compressed with three codecs (AMR, MP3 and Opus) at a single bit rate per codec. For each fricative, CoG, SD, and skewness were measured and compared between the baseline uncompressed PCM-WAV and the codec compressed versions. The findings indicate significant lowering of CoG and SD following codec compression, with segment and codec dependent tendencies. Skewness is likewise affected, but less so than the other measures. This has implications for phonetics when e.g. using codec compressed speech as data, but especially in socio- and forensic phonetics with possible diffusion of sound changes, and speaker comparisons.

Keywords: Codec compression, spectral measures, fricatives, acoustics

1. INTRODUCTION

It is well known that a number of technical effects are apparent in analogue speech transmission, e.g. the limited bandwidth in telephone transmission [1]. However, digital transmission introduces a number of additional variables including codec compression. The codecs are collections of algorithms. These are designed to identify and translate (encode) speech into a numerical representation, which can be sent over a network and finally decode it back to an audible pressure wave [2].

Three aspects in the transmission process and codec compression with AMR-WB and Opus (i.e. speech-specific codecs) are key: a) the digital conversion; b) data carrying capacity and quality, and c) Voice Activity Detection (VAD) and its included voicing parameters. MP3 is a perceptual codec not designed to identify speech, and does not include the VAD. Now, the digital conversion or digitisation of the analogue signal is the translation of the pressure wave into a numerical representation. This digitisation process samples the analogue speech signal at regular intervals, determined by the sampling frequency, and represents the amplitude of the signal at each sample using a pre-determined set of linearly spaced amplitude values [3]. Secondly, the level of acoustic detail (i.e. quality) is determined by the data carrying capacity and the bit rate (i.e. the number of bits that can be sent over a digital network per second). This in turn is also expressed in the available bandwidth for the transmission. All of the investigated codecs are *lossy*, which entails that not all the information in the original pressure wave is captured in the codec-compressed signal [4, 5, 6]. The choice of information to include in the numerical representation is based on the redundancy of speech and identification of regularity (i.e. periodicity/ voicing), as the overall goal is to detect speech and exclude noise (i.e. irregular sound waves) [7, 8]. Thus, codec compression is potentially problematic for voiceless fricatives due to their aperiodic and noise-like acoustic composition. If fricatives are indeed altered by the codec compression this renders the use of digitally transmitted or encoded speech problematic in a range of fields such as sociolinguistics, forensic phonetics and phonetic analysis in general. Therefore, in the absence of relevant research, this study aims to establish a baseline for the spectral implications of codec compression on voiceless fricatives. It does so by investigating [s, f, fʲ, θ, ʃ] as produced by 30 male native speakers of English recorded in a phonetically balanced reading. [fʲ] is included as a separate phoneme as it is used as a distinct label by the forced aligner e.g. in the word e.g. in initial position of the word *furiously*. It is predicted that Centre of Gravity (CoG) and Standard Deviation (SD) will be lowered for all the segments following codec compression partially due to the limited signal bandwidth. However, the lowering is expected to be more prominent for [f], [fʲ], and [θ] considering the relatively lower intensity of the friction than in the sibilants. Skewness is expected to show effects of the codec compression, but with a less clear patterning than CoG and SD. In addition, codec dependent tendencies are expected due to the technical differences between the codecs.

2. METHODOLOGY

2.1. Corpus and participants

The *You Came to Die?!* corpus [9] was used for this study. It included 30 male native speakers of English aged between 18 and 41 reading an approximately 10 minute phonetically balanced passage recorded in studio quality. They spoke five different accents of English with six speakers of each accent: Australian (AUS), New Zealand (NZL), London (LON), Newcastle (NCL), and York (YRK). Female speakers are reported to be more affected by codec compression than male speakers [10], and as the aim is a baseline, male speakers were chosen.

2.2. Segmentation

The files were forced aligned with the Montreal Forced Aligner (MFA) with a customised version of the English (UK) MFA dictionary [11]. The files were then manually corrected in Praat textgrids [12]. These corrections included boundary corrections, where the alignment included e.g. non-speech elements, voicing or wrong transcription.

2.3 Sound files and codec compression

The original 44.1 kHz recordings were down-sampled to 16 kHz. This was done to single out the effects of the codec compression rather than the limitations in bandwidth. The 16 kHz WAV files were codec compressed with AMR-WB [13], MP3 [14], and Opus [15]. A different bit rate was selected for each codec, which represented a typical or average to low quality for each [16, 17, 6]. This was 12.65 kbps for AMR-WB [13], 32 kbps for MP3 [14], and 24 kbps for Opus [15]. This provides a baseline for further research. Spectral analysis was conducted on the 16 kHz WAV files and the three codec compressed files to determine the extent of any changes to the frequency profile of noise-like signals, with particular focus on attenuation at higher frequencies.

2.4 Data extraction

The recordings elicited the fricatives [s, f, fʲ, θ, ʃ] in varying segmental contexts and initial, medial, and final position. This yielded a final baseline dataset with 5104 tokens of [ʃ], 5156 of [θ], 13668 of [f], 960 of [fʲ], and 29336 of [s], and the same number of tokens in each of the three codec compressed versions.

The spectral measures of CoG, SD, and skewness were obtained from the central 20ms frame using multi-taper analysis [18] implemented in MATLAB [19]. Multi-taper analysis requires no additional windowing, and no pre-emphasis was applied. During preliminary examination of the data, it was found that a number of segments had CoG values below 1 kHz, which is unexpected for these fricatives [20]. These segments were excluded from the main dataset along with their equivalent segments in the other codec compressed or baseline files. The 1 kHz threshold was set as the exclusion criterion because a CoG value less than 1 kHz might reflect a strong influence of the mains hum (at around 50 Hz), or a substantial reduction in intensity. An audio file of white noise was also subject to the same encoding processes as the speech samples to estimate the actual upper-frequency limit of the speech encoding.

2.5 Statistical analysis

To provide an overall summary of the spectral measures for each segment, the mean values and difference between these values (i.e. baseline vs. codec compression) for all three spectral measures (i.e. CoG, SD and skewness) in the baseline and each codec compression were calculated.

To examine the influence and interaction of the variables a statistical analysis was performed in R [21, 22] using mixed effects modelling [23] and ANOVAs [23]. The mixed effects models were lmer based with CoG, SD, and skewness as dependent variables.The maximal and optimal model, which was used for the analysis was the following:

Best fitted model = lmer (Spectral measure ~ format * segmentLabel + pre + post + (1|speaker) + (1|word) + (1|duration), data = data)

This model was achieved by comparing a range of models for each spectral measure and codec based on a baseline model predicted to be most accurate from the data. The model was tested with and without a number of variables for each codec and spectral measure with varying fixed and random effects as well as slopes and interactions to ensure the accuracy of the model,. This included the following independent variables: Format (6 levels (2 per model): baseline 16 kHz, and one of the following AMR-WB, MP3, or Opus), Segment (5 levels: [s, f, fj, θ, ʃ]) Speaker (30 levels: individual speakers), word position (with three levels: initial, medial, and final), preceding sound labelled *pre* (with 21 levels) and following sound labelled *post* (with 60 levels). The evaluation of fit was done by inspection of residuals, deviance and Akaike Information Criterion (AIC) [25]. Finally, Tukey adjusted post-hoc tests were done using Emmeans to extract the significance values for the interactions from the models as well as plot the linear predictions [26]. Thus, this resulted in 4 models in total, one per measure.

3. RESULTS

Similar trends are found across codecs; however, the magnitude of these trends is to some extent codec and segment specific. Figure 1 shows spectrographic representations that illustrate how the codec compression limits the intensity and upper frequency limit to varying degrees in the word *fluff* as produced by one York speaker.

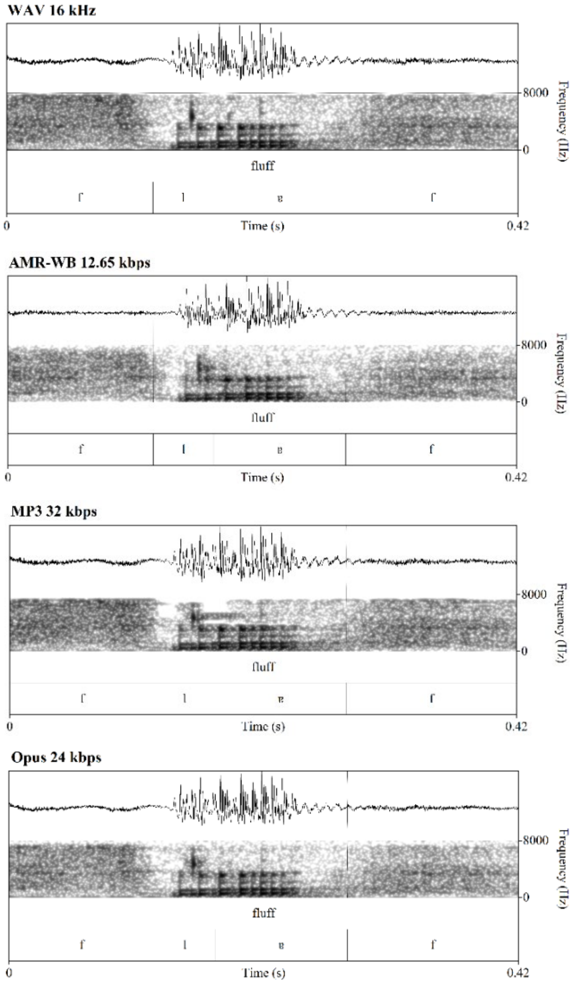

Figure 1: *fluff* produced by one York accented speaker in WAV and all codec compressions.

The white noise analysis indicated that the codecs did not have the same upper frequency limit (i.e. the point where the codec-compressed files do not follow the flat frequency profile of the white noise). For AMR-WB the estimated cut-off frequency i.e. the point when the frequencies clearly start sloping downwards was 5600 Hz, for MP3 7100 Hz, and for Opus 7000 Hz (Figure 2 below). Summary statistics were extracted for each spectral moment in each codec compression. The differences in mean values between the original and codec compressed recordings are illustrated in Table 2, with the arrows indicating the directionality of the average change.

seg	codec	bitrate (kbps)	CoG (Hz)	SD (Hz)	Skew.	Kurt.
[f]	AMR	12.65	306↓	192↓	0.06↓	0.13↓
[f]	MP3	32	128↓	119↓	0.09↓	0.19↓
[f]	Opus	24	415↓	206↓	-0.10↑	-0.35↑
[fʲ]	AMR	12.65	219↓	137↓	0.11↓	0.32↓
[fʲ]	MP3	32	105↓	112↓	0.12↓	0.28↓
[fʲ]	Opus	24	342↓	195↓	-0.02↑	-0.31↑
[s]	AMR	12.65	349↓	50↓	0.13↓	0.07 ↓
[s]	MP3	32	169↓	70↓	0.27↓	-0.19↑
[s]	Opus	24	300↓	65↓	0.33↓	-0.72↑
[ʃ]	AMR	12.65	11↓	-10↑	0.04↓	0.67↓
[ʃ]	MP3	32	18↓	37↓	0.25↓	1.17↓
[ʃ]	Opus	24	121↓	120↓	0.24↓	-0.31↑
[θ]	AMR	12.65	423↓	219↓	-0.02↑	0.08↓
[θ]	MP3	32	180↓	132↓	0.06↓	0.14↓
[θ]	Opus	24	498↓	197↓	-018↑	-0.36↑

Table 2: differences in mean values between baseline (WAV) and codec compression in Hz. Arrows indicate the direction of the change.

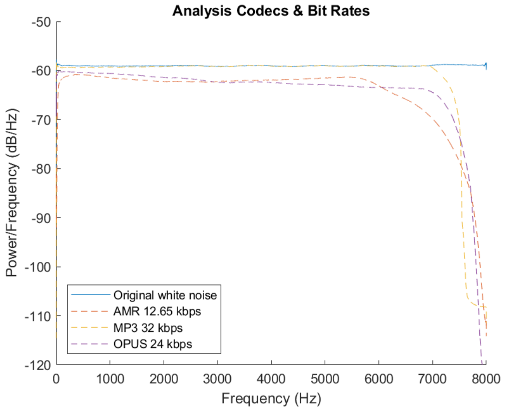

Figure 2: Frequency spectra of original white noise signal and codec compressed versions.

3.1 AMR

The AMR-WB compression generally lowered the spectral measures for all the segments based on the mean values and significantly so [p<.0001]. This is the case for all fricatives apart from [ʃ], where the biggest change is seen for CoG with a lowering of 19 Hz. For CoG the changes vary from between 218 Hz [fʲ] to 422 Hz [θ] from WAV to AMR-WB. The changes in SD vary from 137 Hz [fʲ] to 218 Hz [θ]. The changes to skewness are limited, but significant apart from [ʃ]. The biggest change for skewness was 0.13 found for [s]. The AMR-WB compression makes [f] and [θ] more alike in terms of CoG and skewness.

3.2 MP3

The MP3 compression lowers all spectral values for all segments. This is significant for most measures, with p<.0001, although not for [ʃ]. [ʃ] stays almost unaffected by the MP3 compression, with the biggest change being for SD at 36 Hz. For the remaining segments, the biggest changes in CoG is found for [s], with a lowering of 349 Hz. [f], [θ], and [fʲ] are lowered with between 104 Hz and 179 Hz. For SD [s] is similar to [ʃ], with less than 100 Hz change, whereas the other segments show changes between 112 Hz and 131 Hz. Skewness shows smaller changes with the biggest change observed for [s] at 0.27. The lowering found for both [fʲ] and [θ] are significant (p<0.05). Overall, the relation between the segments' spectral values remain stable.

3.3 Opus

The Opus compression lowers CoG and SD for all segments [p<.0001], but shows a mixed pattern for skewness. Opus is the only codec with an effect greater than 100 Hz on [ʃ] for both CoG and SD. The most substantial lowering of CoG is found for [θ], with 497 Hz, whereas the smallest change is found for [s] with 300 Hz. [f] is lowered by 415 Hz and [fʲ] by 342 Hz. In comparison, the effect on SD is smaller, with a minimum change for [s] at 63 Hz and the remaining segments changing between 119 Hz ([ʃ]) and 205 Hz ([f]). Skewness is lowered for [ʃ] (0.24) and [s] (0.33) [p<.0001]. The remaining tokens are increased, [f] and [θ] significantly with ~0.1 [p<.0001]. [fʲ] was not significantly affected. Overall, [f] and [θ] become almost identical in terms of CoG and SD, and more similar for skewness.

3.4 Rejected tokens

The number of tokens rejected were 61 pairs for WAV-AMR-WB, 18 pairs for WAV-MP3, and 31 pairs for WAV-Opus. None of these tokens were [fʲ] and [ʃ]. Analysis of these tokens is beyond the scope of the current paper. However, most rejected tokens in the AMR-WB were found to have their CoG value above the 1 kHz cut-off in the corresponding WAV file. In Opus and MP3, the CoG mean values for the rejected tokens were all below the 1 kHz cut-off in the WAV baseline apart from [f] in MP3. The MP3 compression shows mixed segment dependent patterns, whereas the Opus codec appears to lower the tokens below 1 kHz similarly to the tokens in the main dataset.

4. DISCUSSION

The study aimed to investigate the acoustic effects of codec compression with AMR-WB, MP3 and Opus on the voiceless fricatives [f], [fʲ], [θ], [s], [ʃ].

It is clear from the white noise analysis that the codecs introduce different cut-offs. This is potentially correlated with the specific bit rates that were applied (and will be investigated further in future). The lowest limit is found for AMR, which also has the most tokens changing CoG from above to below 1 kHz following compression. However, the biggest effects on spectral measures (excluding rejected tokens) is found for Opus, which has a higher upper frequency limit. This, suggests an effect of the available signal bandwidth, but also that the technical elements e.g VAD are influential as well [4]. Thus, the results suggest that the spectral changes are to some extent codec and segment dependent and must be accounted for individually.

As predicted, [f] and [θ] were more affected by the compression than [s] and [ʃ]. However, both [fʲ] and especially [ʃ] presented less to no effect of the codec compression. It is interesting to note that [f] and [θ] become more alike in both AMR-WB and Opus. These results have implications for any acoustic or auditory analysis of codec compressed speech in a number of scenarios.

Firstly, these results follow previous research on digitally transmitted speech in acoustic phonetics as it also cautions its use as primary data [27, 28], unless the goal of the research is, as here, to investigate the technical implications. This is due to the significant effects found for almost every spectral measure in each codec compression.

Secondly, the changes in spectral information potentially make sounds less distinct or not directly comparable to a higher quality recording. Thus, in forensic phonetics, the analysts need to be aware of these effects when analysing and interpreting codec compressed recordings. Whether this has an auditory and/or perceptual effect is to be assessed. If such an effect is found, it means bias could affect the forensic phonetician in their conclusions. Moreover, the amount of codec compressed speech today have potential implications fir sound change and diffusion of variants stemming from codec compression. In a socio-phonetic perspective, when judging the effects, it is important to bear in mind that the baseline WAV here is already down-sampled to 16 kHz. Finally, this study used completely controlled conditions on male speakers without any live transmission or background noise. With these factors added and using female speakers, the effects are likely to be more prominent e.g. when the codec identifies the fricatives as noise and eliminates them from transmission.

5. REFERENCES

[1] Künzel, H. J. 2001. “Beware of the ‘telephone effect’: the influence of telephone transmission on the measurement of formant frequencies”. *Forensic Linguistics* 8, 1, 80–99.

[2] Zölzer, Udo. 2008. *Digital Audio Signal Processing*. U.K.: Wiley

[3] Ladefoged, P. 1996. *Elements of acoustic phonetics.* Chicago: University of Chicago Press.

[4] 3GPP. 2022. “Technical Specification Group Services and System Aspects; Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; Voice Activity Detector (VAD)”. *3GPP TS 26.194*. (version 16.0.0). France: Valbonne

[5] Marc Gayer, Markus Lohwasser and Manfred Lutzky. 2004. "Implementing MPEG Advanced Audio Coding and Layer-3 encoders on 32-bit and 16-bit fixed-point processors". *Revision 1.11*. Germany: Fraunhofer Institute for Integrated Circuits IIS.

[6] Valin, JM., Vos K., & Terriberry T. 2012. “Definition of the Opus Audio Codec” *report*. retrieved 06 August 2022 https://www.rfc-editor.org/info/rfc6716

[7] Guillemin, Bernard J., & Watson, Catherine I. 2006. “Impact of the GSM AMR Speech Codec on Formant Information Important to Forensic Speaker Identification”. *In 11th Australian International Conference on Speech Science & Technology*, 483–88. New Zealand: Auckland University.

[8] Herre, Jürgen & Dick, Sascha. 2019. “Psychoacoustic Models for Perceptual Audio Coding—A Tutorial Review”, *Applied Sciences*. 9 (14). Doi: 10.3390/app9142854

[9] Best, C., Shaw, J., Docherty, G., Evans, B., Foulkes, P., & Hay, J. 2012-2015. *The You Came to Die?! corpus*. ARC Discovery Project DP120104596

[10] Siegert, Ingo, and Oliver Niebuhr. 2021. ‘Speech Signal Compression Deteriorates Acoustic Cues to Perceived Speaker Charisma’. In Tagungsband Der 32. Konferenz, 1–10.

[11] McAuliffe, Michael, Michaela Socolof, Elias Stengel-Eskin, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. *Montreal Forced Aligner* [Computer program]. (Version 1.0.0), retrieved May 2017 from http://montrealcorpustools.github.io/Montreal-Forced-Aligner/

[12] Boersma, P., Weenink, D. 1992-2021. *Praat: doing phonetics by computer*. 6.0.25

[13] 3GPP, 2007-03. “Technical Specification Group Services and System Aspects, ANSI-C code for the Adaptive Multi Rate - Wideband (AMR-WB) speech codec*”. TS 26.173*. (version 7.0.0). France: Valbonne

[14] FFmpeg. 2022. “FFmpeg” *64-bit static Windows 4.4.1-essentials_build.* Retrieved October 2022.

[15] Xiph.Org Foundation. 2022. “opusenc/opusdec”. *opus-tools 0.2 using libopus 1.3.1*. Retrieved October 2022.

[16] 3GPP. 2022. “Performance characterization of the Adaptive Multi-Rate Wideband (AMR-WB) speech codec”. *TR 26.976*. (version 17.0.0). France: Valbonne

[17] Triton. 2022. “Choosing audio bitrate settings”. *https://tritondigitalcommunity.force.com*. Accessed 19th of December 2022.

[18] D. Thomson. 1982. “Spectrum estimation and harmonic analysis,” *Proceedings of IEEE*, Vol. 70, No. 9

[19] MATLAB. 2010. (version 7.10.0 (R2010a)). Natick, Massachusetts: The MathWorks Inc.

[20] Shadle, Christine Helen. 1986. “The acoustics of fricative consonants” *The Journal of the Acoustical Society of America*. 79 (2). Doi: 10.1121/1.393552

[21] R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

[22] RStudio Team. 2019. RStudio Team (2019). *RStudio: Integrated Development for R*., (version 1.2.5033). Boston, MA: RStudio Inc. http://www.rstudio.com/

[23] Kuznetsova, A., PB Brockhoff, and RHB Christensen. 2017. “LmerTest Package: Tests in Linear Mixed Effects Models”. *Journal of Statistical Software* 82 (13): 1–26. https://doi.org/10.18637/jss.v082.i13

[24] Girden, E. R.. 1992. *ANOVA: Repeated measures*. Sage’

[25] H. Akaike. 1974 "A new look at the statistical model identification", *IEEE Transactions on Automatic Control*, 19(6): 716-723. doi: 10.1109/TAC.1974.1100705

[26] Lenth, Russell. 2020. *Emmeans*: *Estimated Marginal Means, Aka Least-Squares Means* (version 1.4.6). R package https://CRAN.R-project.org/package=emmeans

[27] Siegert, Ingo & Niebuhr, Oliver “Speech Signal Compression Deteriorates Acoustic Cues to Perceived Speaker Charisma” *in Tagungsband der 32. Konferenz on Elektronische Sprachsignalverarbeitung.* Germany: Berlin

[28] Leeman, Adrian et. al. 2020. “Linguistic fieldwork in a pandemic: Supervised data collection combining smartphone recordings and videoconferencing”. *Linguistics Vanguard.* 6 (s3). Doi: 10.1515/lingvan-2020-0061