SPEECH SIGNAL COMPRESSION DETERIORATES ACOUSTIC CUES TO PERCEIVED SPEAKER CHARISMA

Ingo Siegert[1], *Oliver Niebuhr*[2]

[1] *Institute of Information and Communication Engineering, OvG University Magdeburg, Germany*
[2] *Centre for Industrial Electronics, University of Southern Denmark, Sonderborg*
ingo.siegert@ovgu.de, olni@sdu.dk

Abstract: Previous research by the authors showed that signal compression codecs used in remote meetings and mobile communications have a substantial negative effect on perceived speaker charisma. Moreover, this effect size varied as a function of speaker gender. Following up from this previous study, we conducted a multiparametric acoustic analysis of a set of sentences elicited from male and female speakers in order to detail the effect of speech-signal compression on charisma-related acoustic-prosodic feature settings. Results show that all compression algorithms caused significant acoustic changes compared to the baseline condition. Almost all of them go in an unfavorable direction concerning speaker charisma. The six compression methods also performed differently well. While OPUS and MP3 caused the fewest negative effects, SPEEX and AMRNB resulted in the most negative effects; GSMFR took a middle position. Moreover, evidence is found for gender-specific effects in terms of both the number of negatively affected acoustic features and their type. The results are discussed with respect to their conceptual implications of perceived speaker charisma and the further development of codecs.

1 Introduction

Charisma is the perceived result of a complex cocktail of sensations, which in turn is based on a complex cocktail of communicative and non-communicative stimulus signals. Antonakis et al. [1] define charisma as "values-based, symbolic, and emotion-laden leader signaling" (p. 304), in line with the now widely accepted idea of charisma as a gradually developed, controllable skill (rather than a divine gift that one either has or does not have, see [2]). In principle, none of the signals in this definition has to be of a communicative nature. For example, non-communicative signals such as visible clothing can also convey values. However, values are based on ideas, and ideas are primarily the domain of the word [3], and thus – regarding charisma – of verbal communication signals. In contrast, emotional signals are primarily non-verbal [4], i.e. based on prosody or body language, for example, and the same probably applies to leader signals as well. Because the term "leader" in the definition by Antonakis et al. is not sufficiently specific in the eyes of Mickalsky [2], he developed the definition further. Michalsky defines charisma as a phenomenon based on three signal pillars: competence, self-confidence, and passion. Michalsky further states that competence creates trust on the part of the recipient. Self-confidence triggers motivation, and passion leads to inspiration and commitment. These pairwise properties constitute the respective basic ingredients of the above-mentioned cocktails of signals and sensations. If one disregards the fact that Michalsky's definition presumably under-emphasizes the factor of shared values, or implicitly accepts it as an essential requirement for the charismatic effect of a speaker to unfold, then Michalsky's further developed definition of charisma does better justice to the existing empirical evidence on charisma or charisma-related

attributes than that of Antonakis and colleagues. This is true insofar as Michalsky's definition implicitly places a stronger focus on the non-verbal communication signals of prosody and body language (also referred to as "delivery", [5]), which indeed have often proven to be more powerful than the word in the perception and modeling of speaker charisma [5, 6, 7]. For prosody, for example, amazing effects could be achieved, even when the prosodic signals came from machines or robots rather than from human beings. [8] and [9] applied the more or less charismatically rated prosodic profiles of Steve Jobs and Mark Zuckerberg to an otherwise identical text-to-speech synthesis output. The result was that the machine or robot that used the more charismatic prosodic profile of Steve Jobs made human interaction partners significantly more often fill out longer questionnaires, book-specific sightseeing trips, and eat healthier food. When used in a car navigation system, Steve Jobs' prosody even made drivers take detours against their better knowledge, following the system's instructions. The particularly prominent position of prosody for the sensation cocktail called charisma is also expressed in the fact that all four dimensions of prosody are involved in charisma perception - (1) pitch, (2) duration/timing, (3) loudness, (4) voice quality - and each with a variety of their corresponding acoustic parameters [10, 11, 12]. Pitch, for example, is relevant in the form of the average pitch level, as well as additionally in the form of the pitch range, the pitch variability and the pitch minimum (at the end of conclusive statements), see [13]. From the perspective of electronic speech signal processing, the intensive interweaving of acoustic-prosodic parameters and the perception of charisma is interesting from (at least) two different points of view. Firstly, not least because of the COVID-19 pandemic, the extent of digital communication has increased immensely in the past years. This applies to voice calls and voice messages on mobile phones as well as to video calls via Skype, Zoom, Teams, Blue Jeans, Big Blue Button, and other providers. The number of video calls alone has increased by about 900 % during the past 15 years; 2020 even saw an additional temporary peak growth of 2,900% due to COVID-19 shutdowns [14, 15]. Companies like Cisco assume that in 2022 about 80% of the global internet traffic is caused by video calls [15]. In video calls, body language and eye contact are often so limited and the transmitted video so blurred or dark that speakers are primarily left with prosody as their only means of creating a charismatic effect on listeners. This is all the more true of course for normal telephone calls and voice messages without a video signal. Thus, we can safely state that the importance of having a charismatic speech prosody in digital communication tools has grown considerably in the recent past and is likely to grow still faster in the future; and in this statement, we have not even taken into account new social-media professions like influencers and YouTubers. This dynamic development is both an opportunity and a challenge for electronic speech signal processing. The second, related reason why the intensive prosody-charisma interweaving is interesting from a signal-processing perspective is that no form of signal compression is conceivable that does not have direct or indirect effects on at least one of the four prosodic dimensions. In other words, there cannot be any form of modern digital communication that does not interfere in some way with the speaker's prosodic charisma triggers. This applies at least to the domain of acoustics. Whether and how compression effects of speech prosody in digital communication also extend into the domain of the perception of speaker charisma is a different question. The authors of this paper investigated this question in a previous perception experiment with German speakers and listeners [16]. Four popular audio compression codecs were tested based on a set of single sentences elicited from male and female speakers: (i) Adaptive Multi-Rate Wideband (AMRWB, 12.65 kBit/s), (ii) MP3 (16 kBit/s), (iii) OPUS (34 kBit/s), and (iv) SPEEX (3.95 kBit/s). The codecs' effects on perceived speaker charisma were compared and evaluated regarding the uncompressed reference (WAV) versions of all sentences. Results show that the three codecs MP3, OPUS, and in particular SPEEX had a significant negative impact on perceived speaker charisma. That is, [16] provided clear evidence that signal artifacts of codec compression not only concern the acoustic-prosodic cues to speaker

charisma. These artifacts also extend into the perception of prosodic speaker charisma. There was also an unexpected further finding. The negative effect of speech signal compression on perceived speaker charisma affected female speakers much more than male speakers. This was true within a codec, i.e. for how much it lowered perceived speaker charisma, as well as between the codecs, i.e. for how many codecs lowered perceived speaker charisma. In the case of the male speakers, this only concerned the SPEEX codec, but not MP3 and OPUS. The latter codec even increased the charisma of male speakers. The present paper continues this line of research of [16]. We aim to search for the acoustic-prosodic origins of the codecs' negative charisma effects in general and their gender specificity in particular. Investigating more closely whether and how the codecs affect prosody and how this relates to the perceived charisma of male and female speakers not only helps us better understand the functioning of the prosodic ingredients in the cocktail of charisma signals. It can also help program more charisma-neutral codecs or develop other compensation methods, up to and including appropriate instructions for speakers or warning feedback signals during voice or video calls. Our research questions are as follows:

(I) Does an acoustic analysis reveal systematic changes in prosodic-parameter measurements as a result of speech signal compression?

(II) Based on (I), do we find different prosodic changes for different codecs? That is, does a codec leave a specific fingerprint in the prosody of the compressed speech signal?

(III) Based on (II), is this codec-specific fingerprint in the prosody of the compressed speech signal further shaped by speaker gender?

(IV) Based on (I)-(III), do connections emerge between a codec's negative effect size on charisma perception on the one hand and the number, extent and/or type of affected prosodic parameters on the other? Can we derive predictions from these connections for how other codecs that have not yet been perceptually tested will affect speaker charisma?

Regarding (IV), we also included two further codecs in the acoustic-prosodic analysis in addition to those four used in [16]. All further details are described in the method section below.

2 Methods

2.1 Utilized Speech Stimuli

This study made use of the same speech stimuli as in [16] to maintain comparability: the Berlin Database of Emotional Speech (EMO-DB) [17]. It contains German sentences recorded by 10 professional actors (five female). Based on the sentences' emotionally neutral verbal content (e.g., "Das will sie am Mittwoch abgeben", She wants to hand that in on Wednesday), the actors realized the sentences with different emotional prosodies as well as in a neutral matter-of-fact version. The database comprises high-quality recordings in both technical and acoustic terms. In technical terms, the sentences are stored as uncompressed WAV files (mono, sampling rate 16 kHz, 16-bit quantization depth, bit-rate 256 KBit/s). The high acoustic quality achieved by studio recordings of trained speakers with clear sonorous voices is one reason why this database is seen as a benchmark dataset for various applications [18]. For our study, we selected (like [16] before) four emotionally neutral sentence realizations by two male speakers (#11 and #15) and two female speakers (#13 and #14) each.

2.2 Utilized Compression Codecs

In today's (mobile) communication systems, speech compression is heavily used, as it reduces the bandwidth for transmission, the transmission delay as well as required system memory and storage [19, 20]. A number of studies investigated the impact of compression on spectral quality and acoustic features [21, 22]. However, compression effects on parameters of emotional prosody or, more specifically, of a speaker's vocal charisma triggers, are rarely investigated. Our

study examines these compression effects for four popular mobile communication codecs and two high-quality music codecs.

Adaptive Multi-Rate Wideband (AMRWB) is a high-quality speech audio codec developed for mobile communication [23], also known as "HD Voice" and Voice over LTE (VoLTE) due to the processing of a wider speech bandwidth (50-6400/7000 Hz). The codec is based on Algebraic Code-Excited Linear Prediction (ACELP) and Linear Predictive Coding (LPC) parameters. We chose a bit-rate of 12.65 kBit/s, which is intended for pure speech signals [23].

Adaptive Multi-Rate Narrowband (AMRNB) is an audio codec specifically designed for speech coding. It operates on narrowband speech (200–3400 Hz) [24]. The codec is mainly used in GSM and UMTS applications. For compression, ACELP and LPC parameters are stored. We chose a bit-rate of 4.75 kBit/s.

GSMFR was the first digital speech coding standard used in the GSM mobile network [25]. It is based on LPC and performs quite poorly compared to its successors AMRNB and AMRWB, mainly because of its low prediction order of 8. GSMFR has a fixed bit-rate of 13 Kbit/s and samples the speech signal at a rate of 8 kHz.

MPEG-1/MPEG-2 Audio Layer III (MP3) is a well-known lossy compression codec that was actually developed for music. By identifying and discarding those parts of the original sound signal that are assumed to exceed a listener's auditory resolution ability, a perceptual coding is implemented. Besides its famous usage for music streaming, lower bitrates (16 kBit/s) are also used to encode audio dramas [26]. Moreover, MP3 has become so popular now also for data in the speech sciences that it is one of the file formats that PRAAT can process [27].

OPUS is an open-source lossy audio codec, offering a speech-oriented operation (SILK, similar to Speex) as well as a low-latency music compression mode (CELT, similar to MP3) [28]. Furthermore, OPUS allows to be operated in a hybrid mode to improve the speech intelligibility at low bit rates by enriching the synthesized signal with characteristics represented by a psychoacoustic model [29]. This hybrid mode is activated by specific bitrates like the 34 kBit/s bit-rate used in the present study.

SPEEX is an open-source lossy speech audio compression format [30]. It uses Code-Excited Linear Prediction (CELP) and is now considered obsolete. Yet, it is still used as a speech transmission codec in some speech assistants [31]. We chose the lowest SPEEX quality parameter in our study (quality=0, i.e. 3.95 kBit/s).

2.3 Acoustic Analysis

Since each of the six codecs was applied to all 16 sentences (4 sentences x 4 m/f speakers), the acoustic analysis comprised 96 compressed sentences. In addition, there were 16 uncompressed (WAV) reference sentences. This resulted in a total of 112 acoustically analyzed individual sentences. The acoustic analysis was conducted in PRAAT, based on the ProsodyPro script written by Xu [32], including the recently added bio-informational dimension (BID) measurements. A total of 12 prosodic parameters were selected, covering three phenomenological dimensions of prosody. The duration/timing dimension was omitted as phrase duration, speaking rate, etc. play hardly any role in predetermined, isolated, and read sentences. Table 1 summarizes the prosodic parameters and, in addition to a brief description, also specifies the direction of the change for which a negative effect on perceived speaker charisma can be assumed. These assumptions refer to the current state of research on acoustic-prosodic charisma cues [10, 11, 12]. In addition to the expected parameters such as pitch range and the levels of pitch and loudness, whose contribution to charisma has been known for more than a decade [10, 11], recent studies suggest that the BID measurements introduced by Xu and colleagues [33] also contribute to speaker charisma.

This relevance probably arises from the fact that BID parameters such as Hammarberg index, formant dispersion, and the spectral center of gravity (CoG), are indicative of the size and weight of a speaker (cf. the "size code" in [34]). That is, BID parameters determine charisma-related concepts like inherent authority and strength (through size) and inherent attractiveness (through weight) and, thus, indirectly also influence speaker charisma itself - not only for humans but also for robot speakers, see [35]. In addition to the 12 acoustic core parameters of charismatic prosody, we also looked at the total acoustic energy (in dB) in 15 ascending frequency bands (250 Hz bandwidth each) from 0-3,750 Hz. This additional analysis gives a more precise picture of the spectral energy distribution than the spectral-tilt estimation of the Hammarberg index. A sonorous, powerful voice and, thus, a shallow spectral tilt is immensely important for speaker charisma [12]. Based on the 15 frequency bands of 250 Hz each, we can determine exactly from which frequency band onwards which codec reduces the amount of acoustic energy in the speech signal and, thus, increases the spectral tilt and, in turn, decreases charisma.

Table 1 – The analyzed prosodic parameters and what change would negatively affect charisma.

Parameter	Description	Bad when...
f0 level	Mean value of f0 (Hz)	↘
f0 range	Difference between min and max f0 value (semitones)	↘
f0 min	Lowest f0 value in the sentence-final terminal fall (Hz)	↗
Intensity level	RMS intensity (dB)	↘
Hammarberg index	Max. energy diff. (dB) between 0-2 kHz and 2-5 kHz	↘
h1-h2	Formant-adj. amplitude (dB) difference between 1st and 2nd harmonic	↗
h1-A3	Amplitude difference (dB) between 1st harmonic and 3rd formant	↗
CoG	Spectral center of gravity (Hz), range 0-5 kHz	↗
Formant dispersion	Mean distance between adjacent formants F1-F3 (Hz)	↘
Jitter	Mean abs. diff. between consecutive periods, divided by mean period	↗
Shimmer	Mean abs. diff. betw. ampl. of consecutive periods, divided by mean ampl.	↗
HNR	Harmonics-to-noise ratio (dB)	↘
Energy distribution	Total energy in 15 frequency bands of 250 Hz each (0-3,750 Hz)	↘

3 Results

The measurements were statistically analyzed by means of a series of z-score single-sample tests. These tests relate the mean and variance of a reference sample (in our case the gender-specific values measured in the uncompressed WAV condition) to the mean of a test sample (here the gender-specific values measured in each codec condition), taking sample size into account. Alpha-error probabilities, i.e. p values, have been adjusted for multiple testing using the Holm-Bonferroni method.

Table 2 is a modification of Table 1. It summarizes separately for each gender the results of the z-score test series across all 12 core parameters. A red marking means that, compared to the WAV condition, the applied codec has shifted the respective parameter value significantly (at $p < 0.05$) in the direction of the indicated arrow and, thus, negatively influenced the prosodic foundation of perceived charisma. A green marking means the opposite, i.e. the codec effect favored the status of the respective parameter as a charisma trigger. No marking means that there is no significant parameter change due to the codec.

Table 2 shows four main results. Firstly, the significant prosodic changes due to codec compression are not equally distributed across the six codecs. While MP3 and OPUS hardly cause any significant prosodic changes in the sentences' acoustic signals, we found quite a few such changes for the AMRWB and GSMFR codecs. Still more significant changes emerged for

Table 2 – Summary of negative/positive effects of codecs on 12 acoustic-prosodic cues to perceived speaker charisma (relative to the WAV baseline); Red/green cells mark sign, m/f= male/female.

		AMRWB		AMRNB		GSMFR		MP3		OPUS		SPEEX	
Parameter	Bad when...	m	f	m	f	m	f	m	f	m	f	m	f
f0 level	↘	+							-	+			-
f0 range	↘				-					+			-
f0 min	↗		-	-	-				-				-
Intensity level	↘											-	-
Hammarberg ind.	↘	-	-	-	-		-					-	
h1-h2	↗		-	-	-		-		-	+	-	-	-
h1-A3	↗				-							-	-
CoG	↗			-	-	+	+				-	-	-
Formant disp.	↘				-	-	-					-	-
Jitter	↗			+	-	-	-		-				
Shimmer	↗		-	-	-	-	-				-	-	-
HNR	↘						-	-				-	-

AMRNB and SPEEX. For these codecs, the number of affected parameters largely exceeds the number of unaffected parameters (9 or 6 unaffected vs 15 or 19 affected parameters). That the distribution of significant prosodic changes differs significantly between codecs is also supported by a Chi-squared test (χ^2[6]=18.174, p<0.01). Secondly, we can see that the codecs did not treat the male and female speakers equally. Rather, it is mainly the female speakers whose speech prosodies were significantly changed by codec compression. More specifically, we found prosodic parameter changes in the majority of cases for the female speakers, i.e. in 60% or 34 out of 57 cases, but only in the minority of cases of the male speakers, i.e. in 40% or 23 out of 57 cases. This is also a significant difference (z[2]=3.12, p<0.01). Note moreover that this gender bias at the cost of female speakers applied in particular to the very popular MP3 and AMRWB codecs, which created 2-3 times as many compression artifacts in the female than in the male speakers' sentences. Thirdly, not all significant prosodic changes we found are presumably bad for perceived speaker charisma. Some are also suitable to increase perceived speaker charisma. However, these beneficial changes mainly concern the male speakers, i.e. in 86% or 6 out of 7 cases. They occurred mainly for the OPUS codec. Fourthly, we see in Table 2 that some prosodic parameters were, in general, more robust against effects of codec compression than others. Intensity, h1-A3, and HNR were hardly affected by codec compression, whereas h1-h2, CoG, Shimmer, and the Hammarberg index turned out to be rather fragile in that respect. Finally, Figures 1(a)-(b) illustrate for three of the six codecs plus the WAV reference condition how the spectral energy distribution develops across the 15 frequency bands. Again, we see considerable differences between male and female speakers in how their specific spectral energy distributions are altered by the codecs. While all codecs increasingly reduce the acoustic energy for higher frequency bands, this reduction is more strongly pronounced for the female than for the male speakers. Moreover, it also seems to set in earlier for the female than for the male speakers, i.e. at about 2000 Hz rather than at about 3000 Hz, where the reduction becomes most obvious and consistent across all three codecs in the male speakers' sentences (the reduction caused by SPEEX also sets in at about 2000 Hz for the male speakers). Note that the popular and widespread AMRWB codec is not excluded from this energy-reduction effect.

4 Discussion and Conclusions

As our results show, the applied six codecs affected the prosody of the test sentences to more than a minor extent. The measured changes were significant and substantial for many codecs. In addition, the number of affected parameters, as well as the direction of these effects, varied

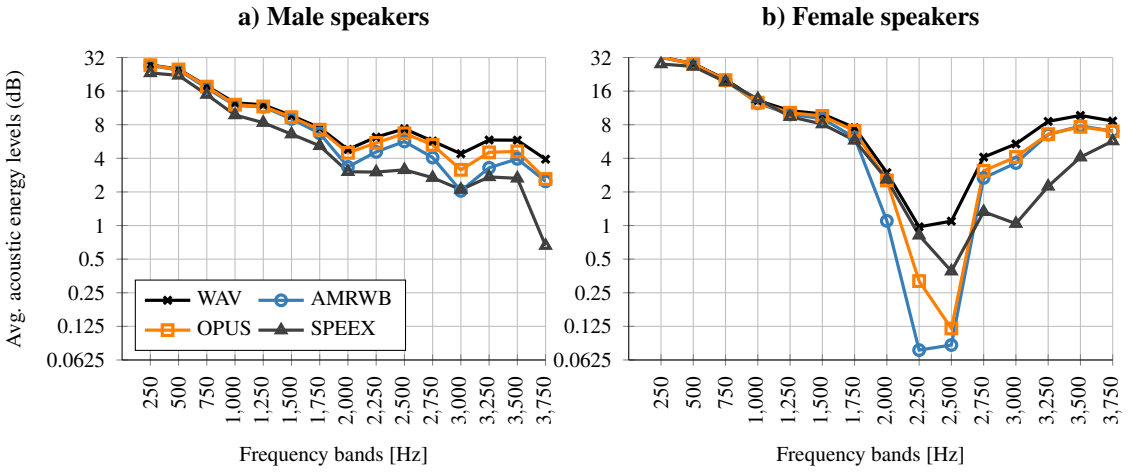

Figure 1 – Avg. acoustic energy levels across increasing frequency bands (250 Hz steps) for three codecs and the WAV reference condition, displayed separately for (a) male and (b) female speakers.

significantly as a function of both codec and speaker gender.

In their experiment with the same sentences tested here, Siegert & Niebuhr [16] found a gender bias in how the codecs affected perceived charisma. This gender bias was to the detriment of the female speakers. Compared to compressed male speech, compressed female speech was not only rated worse overall, but the worse ratings were also caused by a larger number of codecs. While male speakers received worse ratings only under SPEEX compression, female speakers were also rated worse when their speech was compressed with MP3 and OPUS. The latter codec, OPUS, even improved male speech in terms of the listener ratings.

This perceptual rating behavior in [16] fits very well with the results of the present acoustic-prosodic analysis. The SPEEX codec caused most of the prosodic changes, all unfavorable for prosodic charisma triggers and similarly strongly pronounced for male and female speakers. OPUS, on the other hand, caused positive prosodic changes for the male and negative changes for the female speakers. For the MP3 codec, we found only negative effects on charismatic prosody but restricted to the female speakers. If we combine the overall pattern of significant changes in prosodic parameters in Table 2 with the perceptual pattern of speaker ratings in [16], then the results suggest that it were primarily the codecs' changes in f0 parameters, spectral energy distribution (e.g., h1-h2 and CoG) and periodicity (jitter, shimmer) that determined the rating behavior in [16]. This is also plausible insofar as pitch and voice-quality features are considered more powerful cues to charisma than, e.g., intensity features [10].

To sum up, we can answer questions (I)-(IV) of our study positively. With regard to predicting from our results how other codecs than those included in [16] will affect perceived charisma, we assume that AMRWB and GSMFR would also cause clearly negative effects – effects that would probably again concern female speakers more than male speakers. Furthermore, note in this context that those codecs whose usage extends into the compression of music (MP3 and OPUS) performed better than those that were made of speech compression alone. In view of the fact that prosody is also referred to as "speech melody", this finding may represent an interesting perspective for the codecs' further development.

What conclusions should be drawn from these findings? First, the need for scientists and engineers to understand that the quality of a codec must not be measured in terms of word intelligibility alone. Nowadays, where entrepreneurial activities, sales, customer acquisition, leadership, and even political agendas are all handled through digital communication tools, it is also of particular importance that other, non-verbal forms and functions of the speech signal like speaker charisma are properly transmitted. Against this background, work must be carried out to ensure that the codecs do not damage the prosodic charisma cues more than necessary and,

moreover, treat female and male speakers equally. To achieve this, improving how the codecs deal with spectral parameters seems most relevant. What surprised us in this context is the relatively poor performance of the popular codecs AMRWB and MP3, which users often regard as modern state-of-the-art compression methods. The discrepancy between the consumers' trust in these codecs and their poor performance in our study increases the need for action.

In the next step, the task of phonetic research will be to test further aspects of speaker charisma in acoustics and perception and, for example, to examine codec effects on individual vowels and consonants, with an eye on gender-specific differences; because the perceived clarity of the pronunciation also determines the speaker's charisma [36]. It is also important to include spontaneous speech and duration/timing parameters in these follow-up analyses. A practical implication of our findings is that researchers need to be even more careful about collecting voice data using smartphones. This method has become more and more popular recently, also because PRAAT now processes MP3 files. If MP3 files were used, then this should be clearly communicated in the limitations of the study concerned, see also [37]. Note that, while also pointing out artifacts of MP3 compression, [38] consider MP3 files less harmful for phonetic data analysis. However, unlike in the present study, their study only relied on male speakers and thus probably underestimates the magnitude of compression artifacts.

Finally, we would like to emphasize that not all significant effects of the codecs found here underlie genuine and consistent changes in acoustic parameters. A few of these changes, especially those of f0 and formant dispersion, are due to an increased degree of measurement errors in the acoustic analysis – errors that we did not correct manually because they also have relevant perceptual manifestations. The f0 effects arose, for example, from the fact that the respective codecs generated a bleeping noise in the speech signal which – in the ears of the authors and others – actually affected the perceived intonation in a similar way as it affected the acoustic measurements. Nevertheless, the question of how prosodic codec effects must be measured and evaluated is of course also an important subject for future studies.

Acknowledgements

This research is kindly supported by the IE-Industrial Elektronik project (SFD-17-0036) which has received EU co-financing from the European Social Fund. We would furthermore like to thank Krestina Vendelbo Christensen for inspiring discussion about future research on the socio-phonetic issues of compressed speech.

References

[1] ANTONAKIS, J., N. BASTARDOZ, P. JACQUART, and B. SHAMIR: *Charisma: An ill-defined and ill-measured gift. Annual Review of Organizational Psychology and Organizational Behavior*, 3, pp. 293–319, 2016.

[2] MICHALSKY, J. and O. NIEBUHR: *Myth busted?: challenging what we think we know about charismatic speech. Acta Universitatis Carolinae*, (2), pp. 27–56, 2019. doi:10.14712/24646830.2019.17.

[3] GROBE, C.: *The power of words: Argumentative persuasion in international negotiations. European Journal of International Relations*, 16(1), pp. 5–29, 2010. doi:10.1177/1354066109343989.

[4] FILIPPI, P., S. OCKLENBURG, D. L. BOWLING, L. HEEGE, O. GÜNTÜRKÜN, A. NEWEN, and B. DE BOER: *More than words (and faces): evidence for a stroop effect of prosody in emotion word processing. Cognition and Emotion*, 31(5), pp. 879–91, 2017. doi:10.1080/02699931.2016.1177489.

[5] CASPI, A., R. BOGLER, and O. TZUMAN: *"judging a book by its cover": The dominance of delivery*

over content when perceiving charisma. *Group & Organization Management*, 44(6), pp. 1067–1098, 2019. doi:10.1177/1059601119835982.

[6] SCHERER, S., G. LAYHER, J. KANE, H. NEUMANN, and N. CAMPBELL: *An audiovisual political speech analysis incorporating eye-tracking and perception data*. In *Proc. of the LREC'12*, pp. 1114–1120. ELRA, Istanbul, Turkey, 2012.

[7] WÖRTWEIN, T., M. CHOLLET, B. SCHAUERTE, L.-P. MORENCY, R. STIEFELHAGEN, and S. SCHERER: *Multimodal public speaking performance assessment*. In *Proc. of the ACM ICMI'15*, p. 43–50. New York, NY, USA, 2015. doi:10.1145/2818346.2820762.

[8] FISCHER, K., O. NIEBUHR, L. C. JENSEN, and L. BODENHAGEN: *Speech melody matters—how robots profit from using charismatic speech*. *J. Hum.-Robot Interact.*, 9(1), 2019.

[9] NIEBUHR, O. and J. MICHALSKY: *Computer-Generated Speaker Charisma and Its Effects on Human Actions in a Car-Navigation System Experiment - or How Steve Jobs' Tone of Voice Can Take You Anywhere*. In *Proc. of Computational Science and Its Applications – ICCSA 2019*, vol. 11620 LNCS, pp. 375–390. Springer, Saint Petersburg, 2019. doi:10.1007/978-3-030-24296-1_31.

[10] ROSENBERG, A. and J. HIRSCHBERG: *Charisma perception from text and speech*. *Speech Communication*, 51(7), pp. 640–655, 2009.

[11] SIGNORELLO, R., F. DERRICO, I. POGGI, and D. DEMOLIN: *How charisma is perceived from speech: A multidimensional approach*. In *Proc. of the ASE/IEEE SOCIALCOM-PASSAT '12*, pp. 435–440. 2012. doi:10.1109/SocialCom-PASSAT.2012.68.

[12] NIEBUHR, O., R. SKARNITZL, and L. TYLEČKOVÁ: *The acoustic fingerprint of a charismatic voice - initial evidence from correlations between long-term spectral features and listener ratings*. In *Speech Prosody'18*, pp. 359–363. 2018.

[13] NIEBUHR, O. and R. SKARNITZL: *Measuring a speaker's acoustic correlates of pitch - but which? a contrastive analysis for perceived speaker charisma*. In *Proc. of the 19th International Congress of Phonetic Sciences,*, pp. 1774–1778. Melbourne, Australia, 2019.

[14] SCOTT, R.: *Must have video conferencing statistics 2020*. In *UCToday*. 2020. URL `http://bit.ly/UCToday-VideoConferencing2020`. [Online; posted 27-Jul-2020].

[15] HARGRAVE, S.: *In the grip of a climate crisis, demand for video calls is soaring*. In *Wired*. 2020. URL `http://bit.ly/Wired-VideoCalls2020`. [Online; posted 27-Jan-2020].

[16] SIEGERT, I. and O. NIEBUHR: *Women, be aware that your vocal charisma can dwindle in remote meetings*. *Frontiers in Communication*, 5, p. 7, 2021. doi:10.3389/fcomm.2019.00012.

[17] BURKHARDT, F., A. PAESCHKE, M. ROLFES, W. SENDLMEIER, and B. WEISS: *A database of german emotional speech*. In *Proc. of the INTERSPEECH*, pp. 1517–1520. 2005.

[18] SCHULLER, B., B. VLASENKO, F. EYBEN, G. RIGOLL, and A. WENDEMUTH: *Acoustic Emotion Recognition: A Benchmark Comparison of Performances*. In *Proc. of the IEEE ASRU-2009*, pp. 552–557. 2009.

[19] MARUSCHKE, M., O. JOKISCH, M. MESZAROS, F. TROJAHN, and M. HOFFMANN: *Quality assessment of two fullband audio codecs supporting real-time communication*. In *Proc. of the 18th International Conference on Speech and Computer SPECOM 2016*, pp. 571–579. 2016.

[20] SIEGERT, I., A. F. LOTZ, L. L. DUONG, and A. WENDEMUTH: *Measuring the impact of audio compression on the spectral quality of speech data*. In *Elektronische Sprachsignalverarbeitung 2016*, vol. 81 of *Studientexte zur Sprachkommunikation*, pp. 229–236. Leipzig, Germany, 2016.

[21] BYRNE, C. and P. FOULKES: *The 'mobile phone effect' on vowel formants*. *International Journal of Speech, Language and the Law*, 11(1), pp. 83–102, 2004.

[22] SIEGERT, I., A. F. LOTZ, M. MARUSCHKE, O. JOKISCH, and A. WENDEMUTH: *Emotion intelligibility within codec-compressed and reduced bandwith speech*. In *ITG-Fb. 267: Speech Communication : 12. ITG-Fachtagung Sprachkommunikation*, pp. 215–219. 2016.

[23] ITU-T: *Wideband Coding of Speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)*. REC G.722.2, 2003. URL `https://www.itu.int/rec/T-REC-G.722.2-200307-I/en`.

[24] 3GPP: *3GPP TS 26.090 - Mandatory Speech Codec speech processing functions; Adaptive Multi-Rate (AMR) speech codec; Transcoding functions*. Technical specification 26.090, 1999.

[25] ETSI: *ETSI EN 300 961 V8.1.1 (2000-11) - (GSM 06.10 version 8.1.1 Release 1999)*. REN/SMG-110610Q8R1 ETSI EN 300 961 V8.1.1 (2000-11), 2000.

[26] AHERN, S.: *Acoustical design of concert halls and theatres: a personal account*. Routledge, 2020.

[27] BOERSMA, P.: *Praat, a system for doing phonetics by computer*. *Glot Int.*, 5, pp. 341–345, 2001.

[28] VALIN, J.-M., K. VOS, and T. TERRIBERRY: *Definition of the opus audio codec*. RFC 6716, 2012. URL `http://tools.ietf.org/html/rfc6716`.

[29] VALIN, J.-M., G. MAXWELL, T. B. TERRIBERRY, and K. VOS: *The opus codec*. In *135th AES International Convention*. New York, USA, 2013.

[30] XIPH.ORG FOUNDATION: *Speex: A free codec for free speech*. 2014. URL `http://www.speex.org/software/`.

[31] CAVIGLIONE, L.: *A first look at traffic patterns of Siri*. *Transactions on Emerging Telecommunications Technologies*, 26(4), pp. 664–669, 2015. doi:10.1002/ett.2697.

[32] XU, Y.: *Prosodypro — a tool for large-scale systematic prosody analysis*. In *Proc. of the TRASP'2013*. Aix-en-Provence, France, 2013.

[33] LIU, X. and Y. XU: *Body size projection by voice quality in emotional speechEvidence from Mandarin Chinese*. In *Proc. 7th International Conference on Speech Prosody 2014*, pp. 974–977. 2014. doi:10.21437/SpeechProsody.2014-183.

[34] CHUENWATTANAPRANITHI, S., Y. XU, B. THIPAKORN, and S. MANEEWONGVATANA: *Encoding emotions in speech with the size code. a perceptual investigation*. *Phonetica*, 64(4), pp. 210–30, 2008. doi:10.1159/000192793.

[35] FISCHER, K. and N. O.: *Which voice for which robot? acoustic correlates of body size*. In *Proc. of the ACM/IEEE HRI'21*. 2021. To appear.

[36] NIEBUHR, O. and S. GONZALEZ: *Do sound segments contribute to sounding charismatic?: evidence from a case study of steve jobs' and mark zuckerberg's vowel spaces*. *International Journal of Acoustics and Vibration*, 24(2), pp. 343–355, 2019.

[37] NIEBUHR, O. and A. MICHAUD: *Speech data acquisition -: The underestimated challenge*. *Kieler Arbeiten in Linguistik und Phonetik (KALIPHO)*, 3, pp. 1–42, 2015.

[38] FUCHS, R. and O. MAXWELL: *The effects of mp3 compression on acoustic measurements of fundamental frequency and pitch range*. In *Speech Prosody 2016*, pp. 523–527. 2016. doi:10.21437/SpeechProsody.2016-107.