Impact of the GSM AMR Codec on Automatic Vowel Formant Measurement in Praat and VoiceSauce

Jitka Kaiser, and Tomáš Bořil
Institute of Phonetics
Faculty of Arts, Charles University
Prague, Czech Republic
Email: jitka.vanka@gmail.com, tomas.boril@ff.cuni.cz

Abstract—**Automatic formant measurement is generally reliable but can be affected by various factors, such as telephone transmission. As forensic speaker identification often involves comparison of direct (face-to-face) speech with a telephone recording, it is necessary to examine what effect telephony has on the speech signal. This study focuses on the impact of the AMR codec – this codec being the standard in mobile telephony – on formants. In comparison with previous studies, our study analyses the impact of both versions of the codec (narrowband and wideband) at all possible bit rates and on a large amount of data. Furthermore, the effect was examined in two processing tools – Praat and VoiceSauce. Our results revealed considerable shifts of formants when compressed by the codec and indicate that the extent of the shifts differs not only for individual formants but also for the two genders, vowel qualities and the software used.**

Keywords—**automatic extraction; formants; GSM AMR codec; speech coding; telephone transmission**

I. Introduction

Measurement of vowel formants is a common task in many areas of phonetics and a crucial tool in forensic speaker identification (FSI). The importance of formants for FSI dwells in the fact that formant frequencies reflect not only the anatomical characteristics of an individual's vocal tract – its size and shape – but also a speaker's learned articulatory patterns. As a consequence, they are widely considered to be powerful indicators of speaker identity [1], [2].

Nowadays, there are several freely-available software tools (mostly based on linear predictive coding, LPC) which allow automatic measurement of formant frequencies. Two widely used processing tools which perform formant extraction are Praat [3] and the Snack toolkit [4], the latter being implemented also in Wavesurfer [5] and in VoiceSauce (VS) [6]. Under ideal conditions, i.e., when the recordings are of studio quality, the extraction is generally reliable [7]. Nevertheless, especially in the forensic context, the quality of the speech material can be degraded in several ways, and formant tracking consequently becomes more erroneous [8].

A particular type of degradation arises as a result of telephone transmission. Previous research has discussed several effects introduced by telephone transmission on the speech signal in general and on acoustic parameters used in FSI (vowel formants and f0) in particular [9]–[12]. The effects can be divided into three different groups [13], namely environment effects (e.g., background noise), speaker effects (e.g., voice disguise, or adapting a so-called 'telephone voice') and technical effects. Focusing on the technical effects, a particularly striking one is the selective bandpass filtering of the telephone transmission.

Until recently, studies concentrated on landline telephony, which has been shown to have an impact on vowel formants [9], [12] and [14] discussed perceptual consequences of the shifts. Specifically, F1 (especially of close vowels due to its proximity to the lower cut-off) tends to be shifted upwards and F3 (especially higher F3 values, due to their proximity to the upper cut-off) downwards. F2, falling mostly within the frequency range defined by telephony, does not show a tendency to change in any direction [9], [10], cf. [14].

More recent research focuses on the new types of distortions introduced by GSM (Global Standard for Mobile Communication) telephony and on the comparison of the two networks (see [15] for a good overview). Though certain similarities concerning the impact on formants have been reported (F1 shifted upwards, F3 downwards, while F2 is less affected), the effect of mobile transmission on formants is more complex and variable as the two networks considerably differ in the way they process the signal [13].

The mobile channels require a lower transfer rate than landline network and can in addition differ significantly depending on the environment as well as vary in time. Therefore, a speech codec is used to compress the speech signal and reduce the required bit rate considerably below the 64 kbit/s typical in landline phone networks [15]. It allocates a certain number of bits for source coding (i.e., for representing the signal) and channel coding (i.e., for overcoming transmission errors) [16].

There are several codecs currently in use; the most widely used one being the GSM AMR (Adaptive Multi-Rate) codec [17], which exists in a narrowband (AMR-NB) and a wideband version (AMR-WB). What makes this codec unique is that it has no fixed relationship between source and channel coding. Instead, the codec has several different modes (8 for AMR-NB, 7 for AMR-WB), each with a different relation between source and channel coding but with a fixed combined bit rate. It can thus dynamically switch between the different source coding bit rates (4.75, 5.15, 5.90, 6.70, 7.40, 7.95, 10.20 and 12.20 kbit/s for AMR-NB; 6.60, 8.85, 12.65, 14.25, 15.85, 18.25 and 19.85 kbit/s for AMR-WB), depending on the network conditions [15] (for some post-filtering methods of telephone speech enhancement see, e.g., [18]). Importantly, the linkage between the source coding bit rate and the fidelity of the resulting speech signal is non-trivial, which has a crucial impact on the codec's frequency response, whose lower end is 100 Hz for AMR-NB, while the upper end can vary rather unpredictably between 2800 and 3600 Hz [15], [16]. As for AMR-WB, its speech bandwidth stretches from 50 to 7000 Hz, providing higher speech signal fidelity [19].

The effect of the AMR codec on formants – in contrast to the mobile phone network as a whole – has been examined only in a small number of studies and on a rather limited material (only a few vowel qualities in limited contexts) [15], [16], [20]. Moreover, these have, to our knowledge, focused solely on AMR-NB. They have shown that one of the effects of the compression on the speech signal is the introduction of so-called 'white islands' of low energy in the spectrogram, which affect automatic formant extraction especially in the area of F2 and F3. Some studies [20] report rather small effects of the codec on formants (with the exception of, especially higher, F3, which tends to be decreased), while others [16] suggest that all three formants are decreased by the codec and higher-frequency formants especially so. There also appear to be substantial gender differences – female voices tend to be affected significantly more than male voices, though there is considerable variability across both speakers and tokens [15], [16]. Furthermore, even though higher source coding bit rates reproduce the formant trajectory better – as can be predicted – there appears to be no clear pattern [16].

The aim of this study is to examine the impact of the GSM AMR-NR and AMR-WB speech codec (as one component of the mobile phone transmission) on automatic formant measurement in Praat and VS. The reason for employing two different types of software is that with the exception of [20], comparative studies are missing. Another contribution of our study is, we believe, the material used: while previous studies on the impact of AMR on formants [15], [16], [20] concentrated only on certain vowels in limited contexts, our study employs a large quantity of tokens of each of the five Czech short monophthongs in various contexts.

II. Method

A. Material and Subjects

The material for the present study consisted of read dialogues, which form part of the Prague Phonetic Corpus [21]. In these dialogues, students of linguistic programs (aged 20 to 25) were asked to act out, after getting familiar with the text, a series of short read dialogues, which created some degree of spontaneity, while at the same time preserving textual identity. The recordings were obtained in the sound-treated recording studio of the Institute of Phonetics in Prague at 32-kHz sampling frequency and 16-bit resolution.

For this study, we chose 5 male and 5 female speakers; attention has been paid that both lower- and higher-pitched voices are present for both genders. The recordings were segmented by Prague Labeller [22] after which the boundaries of the target segments have been adjusted manually [23].

We analysed the five short Czech monophthongs /ɪ ɛ a o u/ in autosemantic words (since vowels in autosemantic words are less prone to reductions) in various consonantal contexts – only vowels neighbouring a nasal consonant or /r l/ were excluded, as nasal formants are sometimes mistaken by the formant extractor for oral vowel formants and liquids are known to significantly perturb formant frequencies of adjacent vowels. As Czech does not have a systemic reduction in unstressed syllables, vowels in both stressed and unstressed syllables were used for the analysis.

In total, we analysed 3484 vowel tokens (from 345 to 353 per speaker), out of which were 359 items of /ɪ/, 1462 of /ɛ/, 631 of /a/, 613 of /o/, and 419 items of /u/. Each vowel token was represented by a set of sixteen mean F1–F3 values (1 studio condition and 15 codec conditions). The compressed recordings were generated by passing the studio recordings through the AMR codec 15 times, the codec being fixed at one of its source coding bit rates (see Introduction) for each pass. After the compression, the files were converted back to the .wav format (up-sampled to 32 kHz, resolution 16 bit) to match the original recordings.

B. Formant extraction and analyses

We measured the static values of F1–F3 in the central, steady-state portion of a vowel [24] in Praat and VS. Formants were extracted in both types of software from the original studio recordings as well as from the respective recordings passed through the codec at each of its 15 bit rates. For each vowel token we thus acquired a set of 16 values for F1–F3.

As for Praat, with the help of a script, F1–F3 were measured in seven equidistant points in the middle third of each vowel by means of the Burg method [3], using the following settings: 4 formants, 0 – 4 kHz for male speakers and 0 – 4.4 kHz for female speakers, window length of 25 ms, +6 dB/octave preemphasis with frequencies below 50 Hz not being enhanced. Each token was then represented by the mean value from these seven measurements.

Default settings were preserved in VS, which uses the Snack algorithm [4] for formant extraction (covariance method, preemphasis of 0.96, window length of 25 ms and frame shift of 1 ms) and relies on f0 detection by the Straight algorithm [25]. The default settings for Straight were slightly adjusted and differentiated for the two genders: for males, Min f0 was raised to 60 Hz and Max f0 lowered to 400 Hz, while for females Min f0 was raised to 100 Hz and Max f0 to 600 Hz. From the formant values extracted at 1-ms intervals, mean was computed from the middle third of each vowel.

When Praat and VS measurements differed by ≥0.5 octave in the studio recordings, they were automatically disregarded from subsequent analyses (6.1% of all F1 values, 3.1% of F2 and 0.8% of F3 values) to allow a more transparent comparison of the impact of the codec on the two processing tools.

III. Results and Discussion

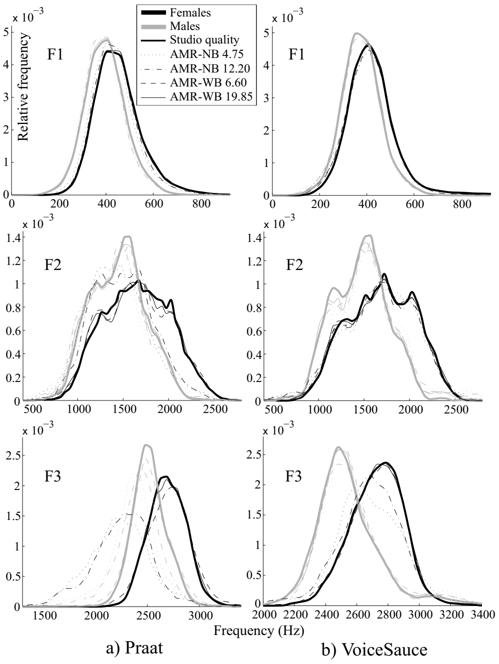
The main aim of our study was to examine the impact of both the narrowband and the wideband version of the AMR codec at all its bit rates on automatic formant measurement in Praat and VS. Figure 1 shows the impact separately for Praat (a) and VS (b), F1–F3 (in rows) as well as male and female speakers (grey and black lines, respectively) by means of relative frequency, the trend being captured by the lowest and highest bit rates of AMR-NB and AMR-WB (individual bit rates are discussed later). The impact of the codec differs considerably for the three formants: while the difference between studio and compressed recordings seems negligible for F1, it is more pronounced for F2 where a downward shift in frequency can be observed (especially for female speakers in Praat) and the biggest shifts occur for F3 (again especially in Praat for female speakers, though F3 in Praat undergoes a visible downward shift even for male speakers). These results are generally in agreement with previous studies [15], [16] which indicated that the AMR codec causes a decrease of formant frequencies – and especially so in the case of higher-frequency formants – and that the effect is more extensive for female than for male speakers. Arguably, this gender effect can be related to differences in average f0 of male and female speakers [16].

Our next aim was to examine whether the two versions of the codec differ in the extent of the effect on formant extraction. In Figure 1, the dotted and dash-dot line illustrate the lowest and highest bit rate of the AMR-NB codec, respectively, and the dashed and thin full line of the AMR-WB, respectively. It can be seen that the AMR-WB codec performs considerably better than AMR-NB, as can be predicted, and that even the lowest wideband bit rate outperforms the highest narrowband bit rate (from the figure visible especially for F3 of female speakers, in Praat even for F2, where the dashed line is aligned with the studio recordings while the dash-dot line is considerably shifted to a lower frequency band). The wider speech bandwidth therefore seems to preserve the formant patterns better than the narrower even at lower bit rates.

From the previous discussion it has also become apparent that the Praat and VS extractors are affected by the codec to a different extent. The figure reveals clear differences in this respect: the distributions appear more “compact” in VS than in Praat; the formant measurements in studio and codec conditions show higher degree of correspondence in VS, while Praat seems to be more affected by the codec.

Lastly, we wanted to examine whether the extent of the shifts differs for individual vowel qualities. These results are presented in Figure 2. As most interesting shifts were occurring for F2 and F3, the figure concentrates only on these. It presents a detailed picture of the behaviour of the formant extractor for individual vowels /ɪ ɛ a u/ (/o/ is missing in the depiction as the trend was comparable to that for /u/) at all bit rates. We can see that the effect of vowel quality and bit rate is by no means straightforward: though the overall trend is to shift formants downwards (as discussed above), looking at the individual vowels and bit rates makes the whole picture more complex. We would expect to observe the biggest shift for /ɪ/ (by virtue of its having the highest F3) but we see that a more F3 variable behaviour appears for /a/ and /ɛ/, where the shifts reach some 600 Hz for female speakers. The original nominal value therefore does not seem to be the only factor, and an interaction with vowel quality and possibly other factors (such as gender) can be expected. As for F2, an interesting effect can be observed for lower F2 studio values, /a/ and especially /u/, which tend to be shifted upward in Praat. The figure shows several interesting trends, e.g., F3 (and in some cases even F2) of female speakers being shifted by the AMR-NB codec downwards to such an extent that it lies lower than for male speakers.

The above discussed observations have been confirmed by the Kolmogorov-Smirnov test (significance level 0.05) which quantifies differences in distributions. It revealed that AMR-WB yields significant differences in distributions only in a small number of tokens when compared with studio recordings. In contrast, significant differences are considerably more numerous for AMR-NB. More pronounced shifts at lower bit rates can be observed especially for F3 of female speakers, though significant differences can be observed for all combinations (Praat or VS, male or female, individual vowel qualities) even for the highest AMR-NB bit rate. The lowest AMR-NB bit rate yields significantly different distributions even for F2; the highest one then mostly only for Praat measurements, VS being more robust. F1 has been found to be the least affected, but with the exception of the vowel /a/ for which the lowest AMR-NB causes significant differences for all combinations. As /a/ is the vowel with the highest F1, there seems to be some boundary (according to our observations around 500 Hz) above which the lowest bit rate of the AMR-NB codec starts to shift the frequency more strongly downwards.



a) Praat b) VoiceSauce

Fig. 1. Relative frequency of F1–F3 measurements (in rows) for all vowels combined in Praat and VoiceSauce (in columns) for male (grey) and female speakers (black) in studio recordings (thick full line) and recordings compressed by the lowest and highest bit rates of the AMR-NB and AMR-WB codec (dotted and dash-dot line, dashed and thin full line, respectively).

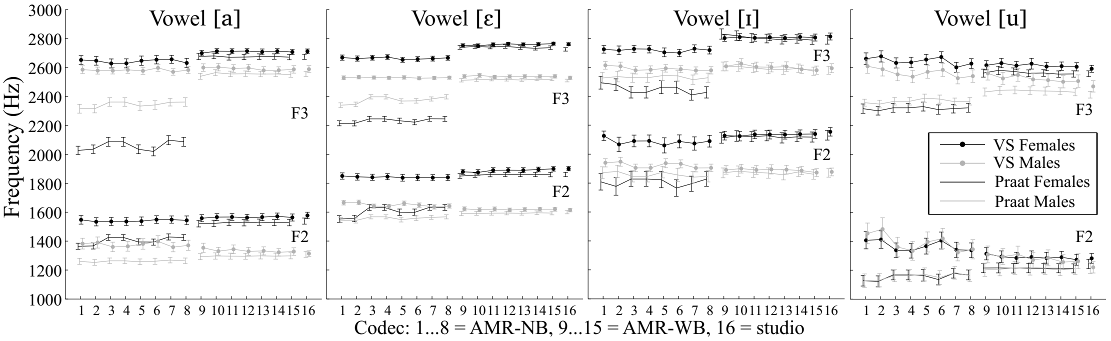


Fig. 2. Mean F2 and F3 (with 95% confidence intervals) of /a ɛ ɪ u/, averaged for 5 male (grey) and 5 female speakers (black) in Praat (simple full line) and VoiceSauce (full line with a dot) in studio recordings and recordings compressed by all bit rates.

IV. Conclusion

The results indicate that though in general a decrease in frequency can be observed, the extent of this shift differs not only for individual formants (F1 was affected the least, F3 the most), but also for the two genders: formants of female speakers undergo larger shifts than those of male speakers. In addition, differences have been found across vowels. Praat in general yielded lower values in comparison to studio recordings, VS appeared more robust to the compression. Since in studio recordings the two extractors provided comparable values, we argue that they are attuned to these ideal, studio conditions. Once the conditions are less favourable the algorithms start behaving distinctly, which has been observed especially for AMR-NB. Our study therefore further suggests that caution is necessary when working with recordings compressed by the AMR codec as automatic formant measurements tend to be affected. An interaction of several factors – formant frequency, gender, vowel quality, bit rate and the software used – can be expected.

References

[1] M. Jessen, *Phonetische und linguistische Prinzipien des forensischen Stimmenvergleichs*. Munich: Lincom, 2012.

[2] F. Nolan and C. Grigoras, "A case for formant analysis in forensic speaker identification," *Int. J. Speech Lang. La.*, vol. 12, no. 2, pp. 143–173, 2005.

[3] P. Boersma and D. Weenink, "Praat – Doing phonetics by computer (V5.3.53)." [Online]. Available: http://www.praat.org [Accessed: Jan. 18, 2014].

[4] K. Sjölander, "Snack sound toolkit (V2.2.10)," *Stockholm, KTH Royal Institute of Technology*, 2004. [Online]. Available: http://www.speech.kth.se/snack [Accessed: Jan. 18, 2014].

[5] K. Sjölander and J. Beskow, "WaveSurfer (V1.8.5)," *Stockholm, KTH Royal Institute of Technology*, 2005. [Online]. Available: http://www.speech.kth.se/wavesurfer [Accessed: Jan. 18, 2014].

[6] Y. Shue, "VoiceSauce: A program for voice analysis (V1.14)," 2013. [Online]. Available: http://www.seas.ucla.edu/spapl/voicesauce/ [Accessed: Oct. 21, 2013].

[7] G. K. Vallabha and B. Tuller, "Systematic errors in the formant analysis of steady-state vowels," *Speech Commun.,* vol. 38, pp. 141–160, 2002.

[8] C. Meinerz and H. Masthoff, "Effect of telephone-line transmission and digital audio format on formant tracking measurements," in *Proc. of the 10th IAFPA conference*, Vienna, Austria, 2011.

[9] H. J. Künzel, "Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies," *Forensic Linguist*, vol. 8, no. 1, pp. 80–99, 2001.

[10] F. Nolan, "The 'telephone effect' on formants: a response," *Forensic Linguist*, vol. 9, no. 1, pp. 74–82, 2002.

[11] H. J. Künzel, "Rejoinder to Francis Nolan's 'The 'telephone effect' on formants: a response'," *Forensic Linguist,* vol. 9, no. 1, pp. 83–86, 2002.

[12] P. J. Rose, "The technical comparison of forensic voice samples," in *Expert Evidence*, I. Freckelton and H. Selby, Eds. Sydney: Thomson Lawbook Co., 2003, ch. 99.

[13] C. Byrne and P. Foulkes, "The 'mobile phone effect' on vowel formants," *Int. J. Speech Lang. La.*, vol. 11, no. 1, pp. 83–102, 2004.

[14] S. Lawrence, F. Nolan, and K. McDougall, "Acoustic and perceptual effects of telephone transmission on vowel quality," *Int. J. Speech Lang. La.*, vol. 15, no. 2, pp. 159–190, 2008.

[15] B. J. Guillemin and C. Watson, "Impact of the GSM mobile phone network on the speech signal: some preliminary findings," *Int. J. Speech Lang. La.*, vol. 15, no. 2, pp. 193–218, 2008.

[16] B. J. Guillemin and C. Watson, "Impact of the GSM AMR speech codec on formant information important to forensic speaker identification," in *Proc. of the 11th Australian International Conference on Speech Science & Technology*, University of Auckland, New Zealand, 2006, pp. 483–488.

[17] 3GPP, "TS 26.071 AMR speech CODEC; General description," *3GPP*, 2012. [Online]. Available: http://www.3gpp.org/ftp/specs/archive/26_series/26.071/ [Accessed: Feb. 2, 2014].

[18] E. Jokinen, P. Alku, and M. Vainio, "Comparison of post-filtering methods for intelligibility enhancement of telephone speech," in *Proc. of 20th European Signal Processing Conference (EUSIPCO 2012),* Bucharest, Romania, 2012, pp. 2333–2337.

[19] ETSI, "Codecs, adaptive multi-rate wideband (AMR-WB) codec," *ETSI*. [Online]. Available: http://www.etsi.org/technologies-clusters/technologies/mobile/codecs [Accessed: Feb. 6, 2014].

[20] E. Enzinger, "Measuring the effects of adaptive multirate (AMR) codecs on formant tracker performance," *J. Acoust. Soc. Am.*, vol. 128, no. 4, pp. 2394–2394, 2010.

[21] R. Skarnitzl, "Prague phonetic corpus: status report," in *AUC Philologica 1/2009, Phonetica Pragensia, XII*, Prague: Charles University Karolinum Press, 2010, pp. 65–67.

[22] P. Pollák, J. Volín, and R. Skarnitzl, "HMM-based phonetic segmentation in Praat environment," in *Proc. of the XIIth International Conference Speech and computer – SPECOM 2007*, Moscow: MSLU, 2007, pp. 537–541.

[23] P. Machač and R. Skarnitzl, *Fonetická segmentace hlásek*. Praha: Epocha, 2009.

[24] M. Duckworth, K. McDougall, G. de Jong, and L. Shockey, "Improving the consistency of formant measurement," *Int. J. Speech Lang. La.*, vol. 18, no. 1, pp. 35–51, 2001.

[25] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous frequency based F0 extraction," *Speech Commun.*, vol. 27, 1999, pp. 187–207.