Measuring the Effects of Adaptive Multi-Rate (AMR) Codecs on Formant Tracker Performance

Ewald Enzinger
Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria
ewald.enzinger@oeaw.ac.at

2nd Pan-American/Iberian Meeting on Acoustics, Cancún, México, 15-19 November 2010

Austrian Academy of Sciences

1. General:

Objective: Investigate how the adaptive multi-rate (AMR) speech codec affects formant measurements obtained by automatic tools.

Motivations:
- Several approaches to forensic speaker comparison rely on formant center frequency measurements as features due to their rather straightforward interpretation as resonance frequencies of the cavities of the human vocal tract (Nolan and Grigoras, 2005; Becker et. al., 2008; Morrison, 2009).
- Telephone conversations constitute a substantial amount of forensic material, which increasingly involves wireless communication channels instead of landline transmission. The effects and limitations introduced by the Adaptive Multi-Rate (AMR) codecs used for speech transmission in GSM and UMTS networks are therefore of special interest in forensic settings.

Prior work:
- Byrne and Foulkes (2004) compared formant measurements of telephone speech recorded directly as well as transmitted over GSM. On average, F1 was 29% higher, F2 was relatively unaffected, likewise F3, except for speakers with high F3.
- Guillemin and Watson (2008) applied a software AMR implementation to studio recordings and studied effects on f0 and exemplified formant measurement degradation on the 12.20 kbps codec level.

2. Methods:

Codec effects simulation:
- AMR codec ANSI-C fixed-point reference implementation (3GPP 2009)
- Each codec bandwidth level (see Section 3) is applied individually

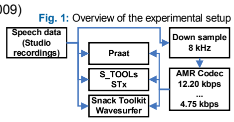

Fig. 1: Overview of the experimental setup

Speech data:
- /a/ and /i/ segments taken from studio recordings of speakers of Viennese German
- Synthesized /a/ and /i/ stationary vowel (Klatt synthesizer)

Automatic formant tracking:
- STx (46-ms frames, 95% overlap, 12 LP coeffs, hamming window, formants obtained by peak picking)
- Snack Toolkit/Wavesurfer (AC method, 49-ms frames, 10ms frame shift, 12 LP coeffs, Cos^4 window, pre-emphasis factor 0.7, formants obtained from LP polynomial roots)
- Praat (std. settings, 25-ms effective frame length, 75% overlap Gaussian-like window)

Band-pass filter:
Simulation of GSM (AMR) to land line (POTS) transmission characteristics by filtering the signal to 300-3400 Hz

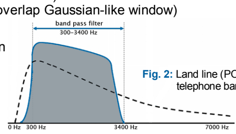

Fig. 2: Land line (POTS) telephone band pass

3. AMR Codec
- Algebraic code-excited linear prediction (ACELP)
- 8 similar modes with varying bit rates
- 4.75, 5.15, 5.90, 6.70, 7.40, 7.95, 10.20, 12.20 kbps
- Discontinuous transmission (DTX)
- Comfort noise generation (CNG)

Encoder processing steps:

Preprocessing (high-pass filtering)
(1) 20-ms frames/windowing
(2) LP coefficients / LSP conversion
(3) Open/closed loop pitch search
(4) Determine codebook indices and gains

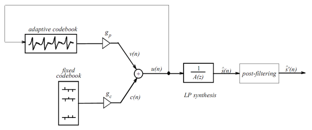

Fig. 3: Simplified block diagram of the CELP synthesis model

4. Results and discussion

Loss of spectral energy ("white islands"):

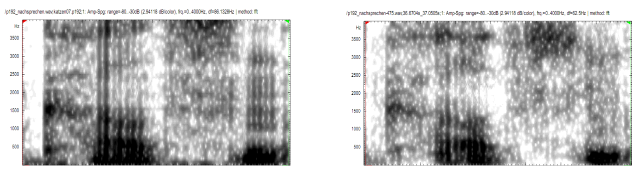

Fig. 4: Spectrogram of the word "Katzen" (cats). Left: PCM 44.1 kHz, Right: AMR 4.75 kbps

Notable loss of spectral detail at approximate F2 & F3 position can lead to wrong automatic formant tracks (especially assignment to formant slots in peak picking method/STx).

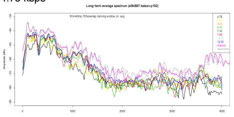

Fig. 5: Long-term average spectrum (LTAS) of /a/ in "Katzen" (cats) at different codec levels

Synthesized vowels:
Stationary /a/ and /i/ vowels of 2-s length were synthesized using the Klatt synthesizer (8 kHz samples) and subsequently band-pass filtered. Fig. 6 and 7 compare the three trackers for the original and the band-limited files. F1 of the /i/ segment (Fig. 7) is significantly affected by the band-pass filter.

	f0	F1	F2	F3
/a/	124	730	1090	2440
/i/	136	270	2290	3010

Tab. 1: Formant parameters (taken from Peterson and Barney, 1952).

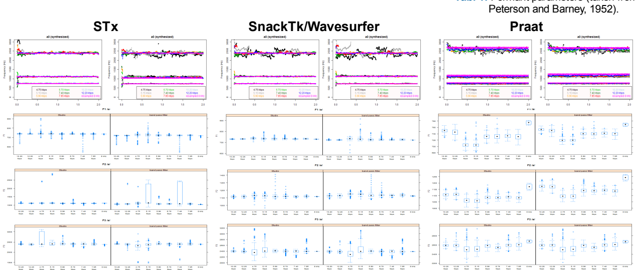

Fig. 6: Formant measurements of F1-F3 in a synthesized /a/ vowel from original/encoded and band pass filtered/encoded signals, using STx (left), SnackTk/Wavesurfer (middle) and Praat (right)

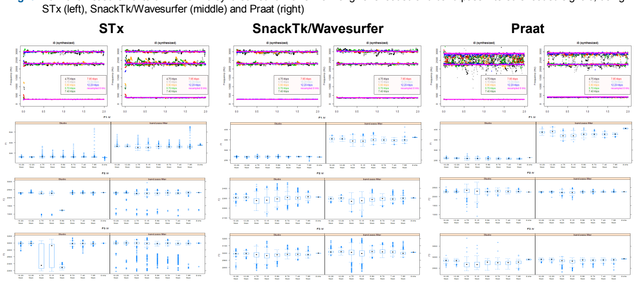

Fig. 7: Formant measurements of F1-F3 in a synthesized /i/ vowel from original/encoded and band pass filtered/encoded signals, using STx (left), SnackTk/Wavesurfer (middle) and Praat (right)

Sudio recordings: Difference between formant measurements

The scatter plots in Fig. 8 and 9 investigate frequency-dependent shifts in formants for each codec level. The measurements were obtained by STx. As can be seen, there are relatively minor differences for F1 and F2. For F3, a pattern similar to the results in Byrne and Foulkes (2004) can be observed in that especially higher frequency formant values tend to be reduced in the measurements from encoded material.

• /a/:

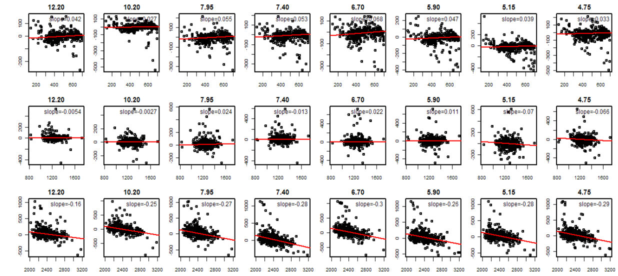

Fig. 8: Scatter plot of /a/ formant measurements. Original/studio recording (x-axis) vs. shift for different codec levels (y-axis). Values > 0 on the y-axis indicate higher values for the codec.

• /i/:

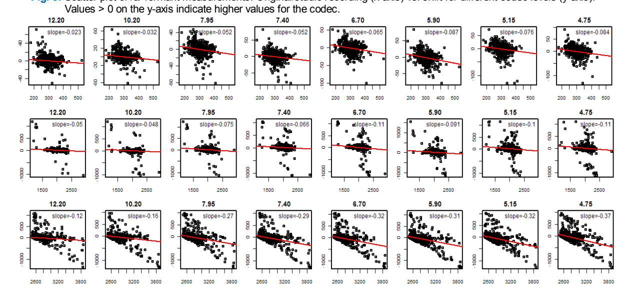

Fig. 9: Scatter plot of /i/ formant measurements. Original/studio recording (x-axis) vs. shift for different codec levels (y-axis). Values > 0 on the y-axis indicate higher values for the codec.

Differences for individual speakers:

Formant tracks obtained by STx from /a/ segments produced by six speakers were manually corrected to investigate speaker-specific codec effects. Manual editing included correcting assignment of formants and adding missing formants where they could be inferred from the spectrum.
In individual formant tracks, deviations from the formant tracks of the studio recording can frequently be observed, as outlined in Guillemin and Watson (2008). However, the distributions of the formant measurements obtained from the speakers show rather small differences.

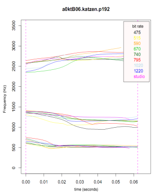

Fig. 10: Long-term average spectrum (LTAS) of /a/ in "Katzen" (cats) at different codec levels

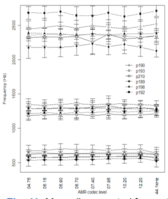

Fig. 11: Manually corrected formant measurements of /a/ segments from 6 speakers.

Effects induced by band pass fltering:

Formant measurements obtained by STx from encoded studio recordings are compared with those obtained from encoded band pass filtered recordings. This condition is of special interest if recorded telephone conversations originating from a cellular phone in the GSM/UMTS network are transmitted via land line (POTS) which results in band pass filtering.

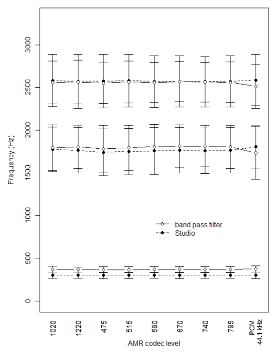

Fig. 12: Comparison of formant measurements of /i/ segments from studio recordings and band-pass filtered files.

Fig. 12 and Tab. 2 show the additional effect of band-pass filtering on the formant measurements. As can be seen, the first formant of /i/ segments is strongly affected. This is in line with results in Byrne and Foulkes (2004). The second and third formants are also affected, but to a lesser degree. The effects caused solely by the codec are relatively small.

	mean F1 (Hz)	factor	t-test (p)	std. dev.
Original rec.	299.6			38.4
AMR 12.20	300.2	100.2%	0.7147	38.0
AMR 10.20	300.1	100.2%	0.7562	37.2
AMR 7.95	299.2	99.9%	0.7882	37.6
AMR 7.40	298.7	99.7%	0.5236	37.1
AMR 6.70	298.7	99.7%	0.5439	37.4
AMR 5.90	298.8	99.7%	0.5787	37.0
AMR 5.15	298.5	99.6%	0.4359	37.8
AMR 4.75	298.7	99.7%	0.5188	37.9

	mean F1 (Hz)	factor	t-test (p)	std. dev.
Original rec.	299.6			38.4
BP AMR 12.20	368.3	122.9%	0.0000	29.0
BP AMR 10.20	370.4	123.6%	0.0000	35.0
BP AMR 7.95	367.0	122.5%	0.0000	32.3
BP AMR 7.40	367.2	122.6%	0.0000	33.8
BP AMR 6.70	366.9	122.5%	0.0000	34.0
BP AMR 5.90	366.5	122.3%	0.0000	33.0
BP AMR 5.15	365.8	122.1%	0.0000	33.3
BP AMR 4.75	365.4	121.9%	0.0000	32.7

	mean F2 (Hz)	factor	t-test (p)	std. dev.
Original rec.	1946.0			275.0
AMR 12.20	1932.4	99.3%	0.2299	292.9
AMR 10.20	1939.8	99.7%	0.5789	288.1
AMR 7.95	1944.6	99.9%	0.9013	296.2
AMR 7.40	1953.2	100.4%	0.531	297.2
AMR 6.70	1950.4	100.2%	0.7031	297.3
AMR 5.90	1957.5	100.6%	0.3156	298.8
AMR 5.15	1955.7	100.5%	0.4056	305.4
AMR 4.75	1946.8	100.0%	0.9449	313.1

	mean F2 (Hz)	factor	t-test (p)	std. dev.
Original rec.	1946.0			275.0
BP AMR 12.20	1802.8	92.6%	0.0000	251.1
BP AMR 10.20	1794.8	92.2%	0.0000	267.0
BP AMR 7.95	1806.0	92.8%	0.0000	253.7
BP AMR 7.40	1814.5	93.2%	0.0000	244.8
BP AMR 6.70	1814.5	93.2%	0.0000	247.5
BP AMR 5.90	1806.2	92.8%	0.0000	262.2
BP AMR 5.15	1795.6	92.3%	0.0000	263.6
BP AMR 4.75	1783.2	91.6%	0.0000	275.1

	mean F3 (Hz)	factor	t-test (p)	std. dev.
Original rec.	2782.5			257.5
AMR 12.20	2786.7	100.1%	0.7463	273.2
AMR 10.20	2764.3	99.3%	0.1327	248.2
AMR 7.95	2760.7	99.2%	0.073	249.5
AMR 7.40	2760.0	99.2%	0.0641	250.5
AMR 6.70	2760.8	99.2%	0.0755	248.9
AMR 5.90	2765.8	99.4%	0.1775	259.4
AMR 5.15	2766.0	99.4%	0.1832	259.8
AMR 4.75	2765.6	99.4%	0.1767	264.1

	mean F3 (Hz)	factor	t-test (p)	std. dev.
Original rec.	2782.5			257.5
BP AMR 12.20	2563.8	92.1%	0.0000	255.9
BP AMR 10.20	2560.4	92.0%	0.0000	250.5
BP AMR 7.95	2562.4	92.1%	0.0000	240.3
BP AMR 7.40	2562.9	92.1%	0.0000	233.5
BP AMR 6.70	2569.5	92.3%	0.0000	236.9
BP AMR 5.90	2559.4	92.0%	0.0000	235.7
BP AMR 5.15	2565.9	92.2%	0.0000	246.2
BP AMR 4.75	2551.4	91.7%	0.0000	236.3

Tab. 2: Comparison of mean formants of /i/ segments from orginal studio recordings and AMR encoded files as well as band pass filtered AMR encoded files. F1 measurements show a substantial increase, F2 and F3 are less affected.

4. Conclusions:

- Band pass filter (300-3400 Hz) leads to higher F1 in measurements for vowels with generally low F1
- The effects caused by the codec itself seem to be rather small compared to the band-pass effects. A small tendency for high F3 measurements to yield lower values from encoded files can be observed.
- The codec does affect automatic formant tracking in terms of wrong assignment of formants and missing values, requiring a greater amount of manual corrections.

5. References:

3GPP (**2009**). "TS 26.073 ANSI-C code for the Adaptive Multi Rate (AMR) speech codec," http://www.3gpp.org/ftp/Specs/latest/Rel-9/26_series/ (retrieved 2010-05-14).
Becker, T., Jessen, M. and Grigoras, C. (**2008**). "Forensic Speaker Verification Using Formant Features and Gaussian Mixture Models," in *Proceedings of Interspeech 2008 incorporating SST'08*, 1505-1508.
Boersma, P. and Weenink, D. (**2009**). "Praat: doing phonetics by computer (Version 5.1.05) [Computer program]," retrieved July 15, 2009, from http://www.praat.org/.
Byrne, C. and Foulkes, P. (**2004**). "The 'Mobile Phone Effect' on vowel formants," Int. J. Speech, Lang. and the Law **11**, 83-102.
Guillemin, B. J. & Watson, C. (**2008**). "Impact of the GSM Mobile Phone Network on the Speech Signal - Some Preliminary Findings," Int. J. Speech, Lang. and the Law, **15**, 193-218.
Morrison, G. S. (**2009**). "Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs," J. Acoust. Soc. Am. **125**, 2387-2397.
Nolan, F. and Grigoras, C. (**2005**). "A case for formant analysis in forensic speaker identification," Int. J. Speech, Lang. and the Law **12**, 143-173.
Peterson, G. and Barney, H. (**1952**). "Control method used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175-184.
Sjölander, K. and Beskow, J. (**2010**). "Wavesurfer - an open source speech tool [Computer program]," http://www.speech.kth.se/wavesurfer
STx (**2010**). "STx 3.9.4 [Computer program]," http://www.kfs.oeaw.ac.at/