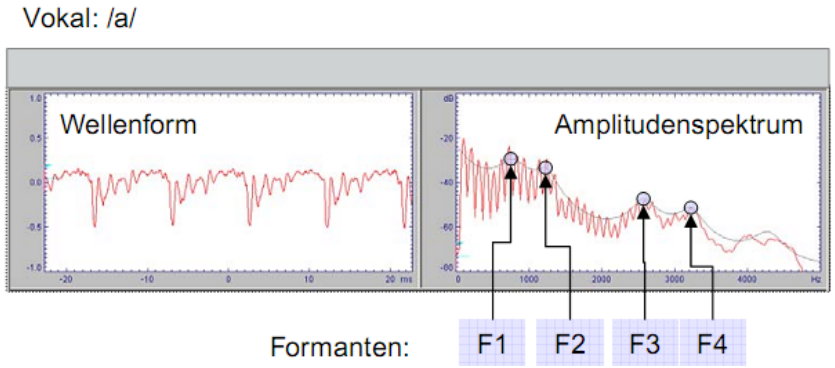
Auswirkungen von Sprachcodecs auf Formantmessungen für Sprechervergleiche

Ewald Enzinger

Österreichische Akademie der Wissenschaften
Institut für Schallforschung

DAGA 2011

- Akustisch-phonetische Sprechererkennung, z.B.
  - Grundfrequenz f0
  - Formantmittenfrequenzen
- Formanten
  - Spektrale Prominenzen/Peaks (Fant 1960)



  - Resonanzfrequenzen des Vokaltrakts (Stevens 1999)

*The **poles** [of the denominator polynomial of the vocal tract transfer function] represent the **complex natural frequencies of the vocal tract**. The imaginary parts indicate the frequencies at which oscillations would occur in the absence of excitation, and are called the **formant frequencies**.*

Adaptive Multi-Rate Codec

- Verwendet in GSM und UMTS
- Spezifiziert durch 3GPP/GSM als Mandatory Codec
- Algebraic Code Excited Linear Prediction (ACELP)
- 8 ähnliche Codecstufen mit unterschiedlichen Bitraten
  ➡ 4.75, 5.15, 5.90, 6.70, 7.40, 7.95, 10.20, 12.20 kbit/s
- DTX - Discontinuous transmission
- CNG - Comfort noise generation

Adaptive Multi-Rate Codec

- Verwendet in GSM und UMTS
- Spezifiziert durch 3GPP/GSM als Mandatory Codec
- Algebraic Code Excited Linear Prediction (ACELP)
- 8 ähnliche Codecstufen mit unterschiedlichen Bitraten
  ➡ 4.75, 5.15, 5.90, 6.70, 7.40, 7.95, 10.20, 12.20 kbit/s
- DTX - Discontinuous transmission
- CNG - Comfort noise generation

Adaptive Multi-Rate Codec

- Verwendet in GSM und UMTS
- Spezifiziert durch 3GPP/GSM als Mandatory Codec
- Algebraic Code Excited Linear Prediction (ACELP)
- 8 ähnliche Codecstufen mit unterschiedlichen Bitraten
  ➡ 4.75, 5.15, 5.90, 6.70, 7.40, 7.95, 10.20, 12.20 kbit/s
- DTX - Discontinuous transmission
- CNG - Comfort noise generation

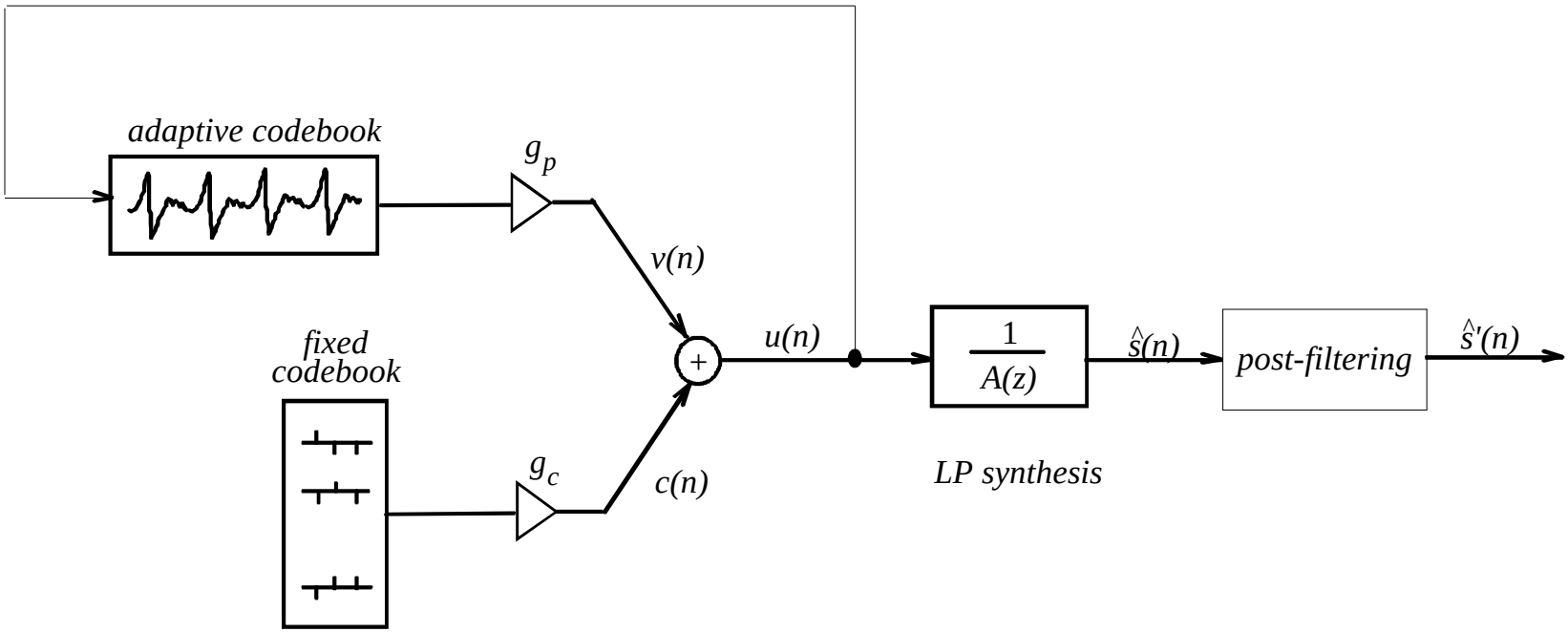
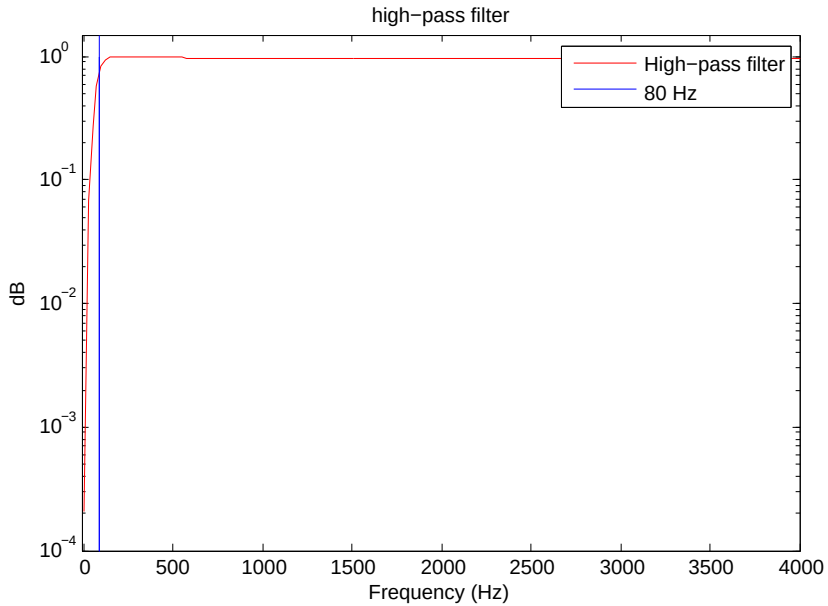


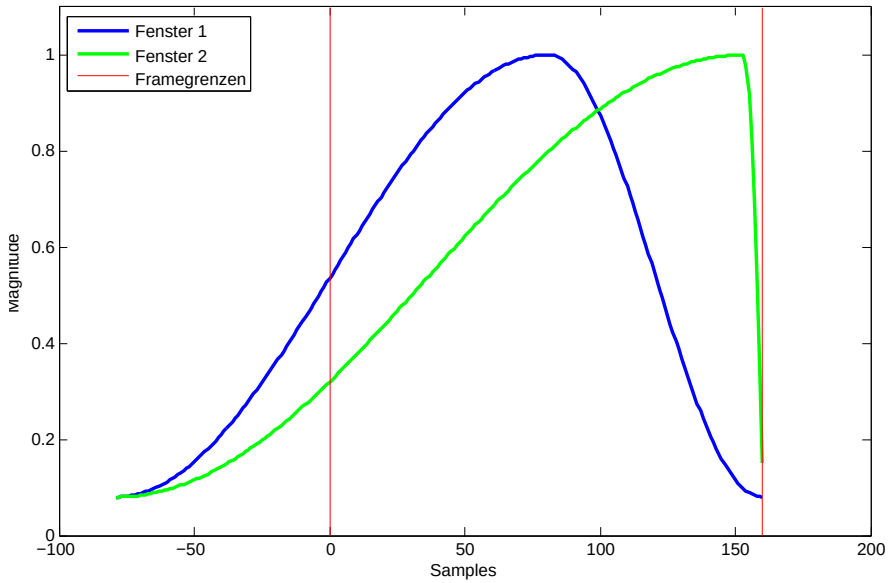
Abbildung: Vereinfachtes Blockdiagramm des CELP Synthesemodells

1. Preprocessing
2. 20ms Frames
3. LP Koeffizienten
4. Algebraisches/ fixiertes Codebook
5. Post processing



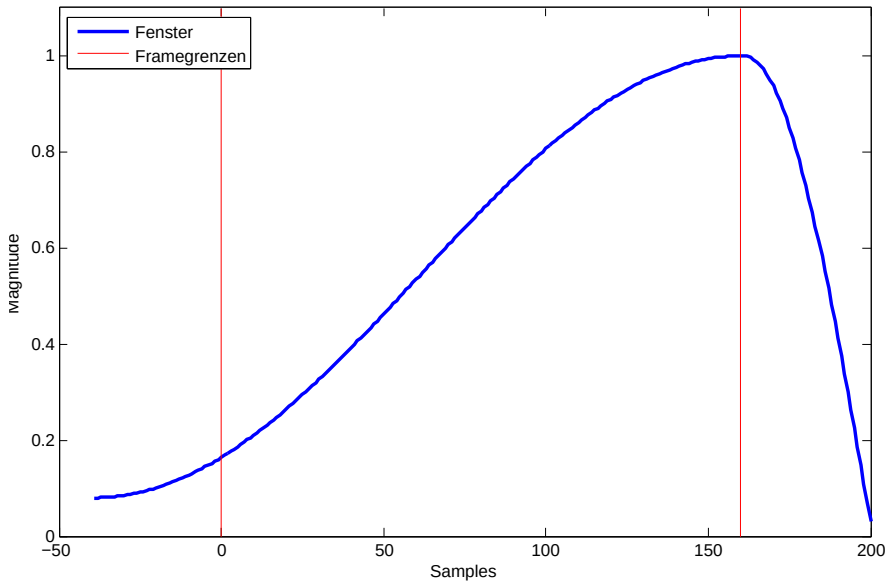
1. Preprocessing
2. 20ms Frames
3. LP Koeffizienten
4. Algebraisches/ fixiertes Codebook
5. Post processing

- 12.2 kbps: 2 überlappende Fenster



1. Preprocessing
2. 20ms Frames
3. LP Koeffizienten
4. Algebraisches/ fixiertes Codebook
5. Post processing

- andere: 1 Fenster, 5ms Lookahead



1. Preprocessing
2. 20ms Frames
3. LP Koeffizienten
4. Algebraisches/ fixiertes Codebook
5. Post processing

- 10th order Linear Prediction

$$H(z) = \frac{1}{\hat{A}(z)} = \frac{1}{1 + \sum_{i=1}^{m} \hat{a}_i z^{-i}}$$

  - 12.2 kbps: 2x pro Frame
  - andere Modi: 1x pro Frame
- Berechnung über Autokorrelation und Levinson-Durbin-Algorithmus
- Umwandlung in Line Spectral Pairs (LSP)

1. Preprocessing
2. 20ms Frames
3. LP Koeffizienten
4. Algebraisches/ fixiertes Codebook
5. Post processing

- 10th order Linear Prediction

$$H(z) = \frac{1}{\hat{A}(z)} = \frac{1}{1 + \sum_{i=1}^{m} \hat{a}_i z^{-i}}$$

  - 12.2 kbps: 2x pro Frame
  - andere Modi: 1x pro Frame
- Berechnung über Autokorrelation und Levinson-Durbin-Algorithmus
- Umwandlung in Line Spectral Pairs (LSP)

1. Preprocessing
2. 20ms Frames
3. LP Koeffizienten
4. Algebraisches/ fixiertes Codebook
5. Post processing

- Long-term (Pitch) Synthesefilter

$$\frac{1}{B(z)} = \frac{1}{1 - g_p z^{-T}}$$

- Pitch delay T und Gain factor g_p bestimmt durch Open-/Closed loop pitch search
- Fixed Codebook-Suche durch Minimierung des Fehlers zw. perzeptiv gewichteten Inputsignals und rekonstruiertem Signal *(Analysis by Synthesis)*

Adaptive Multi-Rate Codec
Verarbeitungsschritte

1. Preprocessing
2. 20ms Frames
3. LP Koeffizienten
4. Algebraisches/ fixiertes Codebook
5. Post processing

- Adaptiver Postfilter
  - Formant Postfilter
  - Tilt-Kompensations Filter
- High-pass filter

Auswirkungen auf Formantdaten

Frühere Untersuchungen

- Byrne & Foulkes (2004)
  - 12 Sprecher (6 m, 6 w)
  - Aufnahme direkt und über GSM-Netz
  - F1 um 29% höher, F2 & F3 nicht beeinträchtigt
- Guillemin & Watson (2008)
  - 8 Sprecher (5 m, 3 w)
  - Aufnahmen aus Studio-Korpus, AMR En-/Decoding
  - Keine generellen Aussagen, Auswirkungen anhand von Beispielen beschrieben

Auswirkungen auf Formantdaten

Frühere Untersuchungen

- Byrne & Foulkes (2004)
  - 12 Sprecher (6 m, 6 w)
  - Aufnahme direkt und über GSM-Netz
  - F1 um 29% höher, F2 & F3 nicht beeinträchtigt
- Guillemin & Watson (2008)
  - 8 Sprecher (5 m, 3 w)
  - Aufnahmen aus Studio-Korpus, AMR En-/Decoding
  - Keine generellen Aussagen, Auswirkungen anhand von Beispielen beschrieben

Auswirkungen auf Formantdaten
Methoden

- Studioaufnahmen von 27 männlichen Wiener Sprechern
  ➡ /a/ und /ɪ/ Vokale
- AMR En-/Dekodierung durch ANSI-C Fixed Point Referenzimplementierung (3GPP 2009)
- Formant tracking durch STx und SnackTk/Wavesurfer
  - STx: Formanten durch Peaks im LP-Spektrum (➡ *Spectral Prominences*)
  - SnackTk: Formanten von Polstellen des LPC All-Pol Filters
- Simulierter Telefonkanal (Bandpass 300-3400 Hz, POTS)
- Synthetisierte stationäre /a/ und /ɪ/ Vokale (KlattSyn)
  ➡ Polstellen des AMR-internen LPC10 All-Pol Filters

Auswirkungen auf Formantdaten
Methoden

- Studioaufnahmen von 27 männlichen Wiener Sprechern
  ➡ /a/ und /ɪ/ Vokale
- AMR En-/Dekodierung durch ANSI-C Fixed Point Referenzimplementierung (3GPP 2009)
- Formant tracking durch STx und SnackTk/Wavesurfer
  - STx: Formanten durch Peaks im LP-Spektrum (➡ *Spectral Prominences*)
  - SnackTk: Formanten von Polstellen des LPC All-Pol Filters
- Simulierter Telefonkanal (Bandpass 300-3400 Hz, POTS)
- Synthetisierte stationäre /a/ und /ɪ/ Vokale (KlattSyn)
  ➡ Polstellen des AMR-internen LPC10 All-Pol Filters

Auswirkungen auf Formantdaten

Methoden

- Studioaufnahmen von 27 männlichen Wiener Sprechern
  ➡ /a/ und /ɪ/ Vokale
- AMR En-/Dekodierung durch ANSI-C Fixed Point Referenzimplementierung (3GPP 2009)
- Formant tracking durch STx und SnackTk/Wavesurfer
  - STx: Formanten durch Peaks im LP-Spektrum (➡ *Spectral Prominences*)
  - SnackTk: Formanten von Polstellen des LPC All-Pol Filters
- Simulierter Telefonkanal (Bandpass 300-3400 Hz, POTS)
- Synthetisierte stationäre /a/ und /ɪ/ Vokale (KlattSyn)
  ➡ Polstellen des AMR-internen LPC10 All-Pol Filters

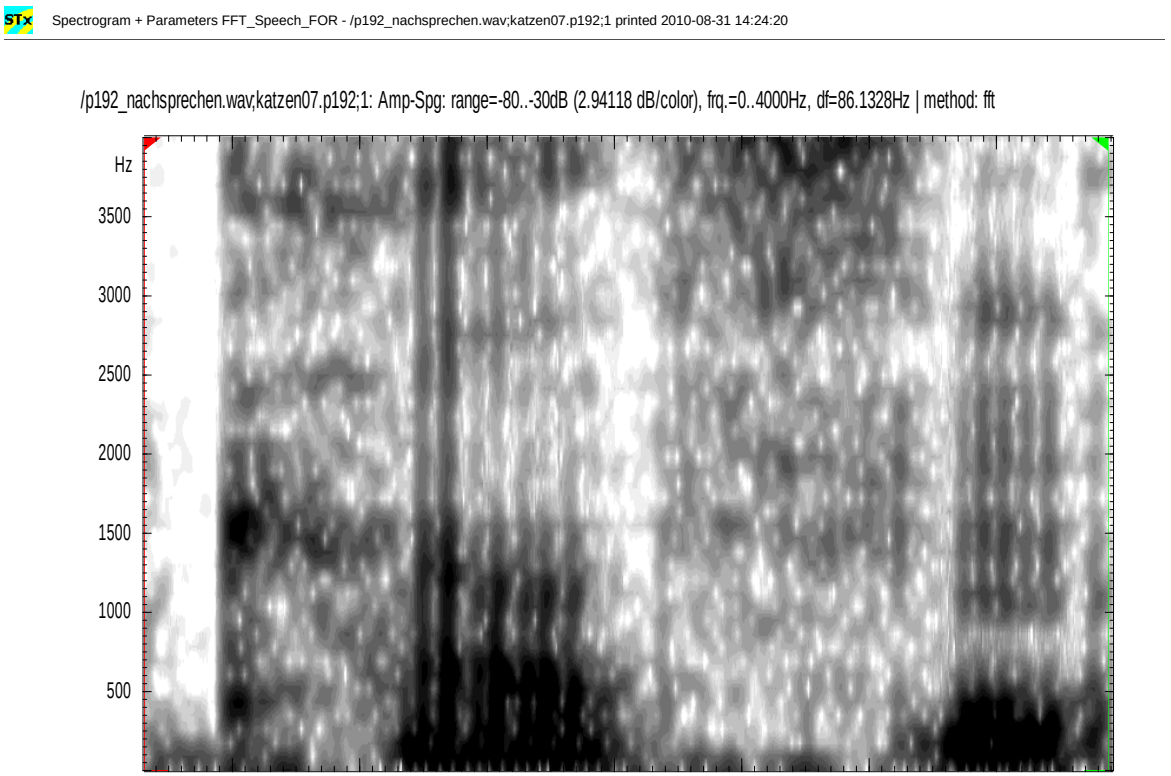
- Studioaufnahmen von 27 männlichen Wiener Sprechern
  ➡ /a/ und /ɪ/ Vokale
- AMR En-/Dekodierung durch ANSI-C Fixed Point Referenzimplementierung (3GPP 2009)
- Formant tracking durch STx und SnackTk/Wavesurfer
  - STx: Formanten durch Peaks im LP-Spektrum (➡ *Spectral Prominences*)
  - SnackTk: Formanten von Polstellen des LPC All-Pol Filters
- Simulierter Telefonkanal (Bandpass 300-3400 Hz, POTS)
- Synthetisierte stationäre /a/ und /ɪ/ Vokale (KlattSyn)
  ➡ Polstellen des AMR-internen LPC10 All-Pol Filters

Auswirkungen auf Formantdaten
Methoden

- Studioaufnahmen von 27 männlichen Wiener Sprechern
  ➡ /a/ und /ɪ/ Vokale
- AMR En-/Dekodierung durch ANSI-C Fixed Point Referenzimplementierung (3GPP 2009)
- Formant tracking durch STx und SnackTk/Wavesurfer
  - STx: Formanten durch Peaks im LP-Spektrum (➡ *Spectral Prominences*)
  - SnackTk: Formanten von Polstellen des LPC All-Pol Filters
- Simulierter Telefonkanal (Bandpass 300-3400 Hz, POTS)
- Synthetisierte stationäre /a/ und /ɪ/ Vokale (KlattSyn)
  ➡ Polstellen des AMR-internen LPC10 All-Pol Filters

Auswirkungen auf Formantdaten

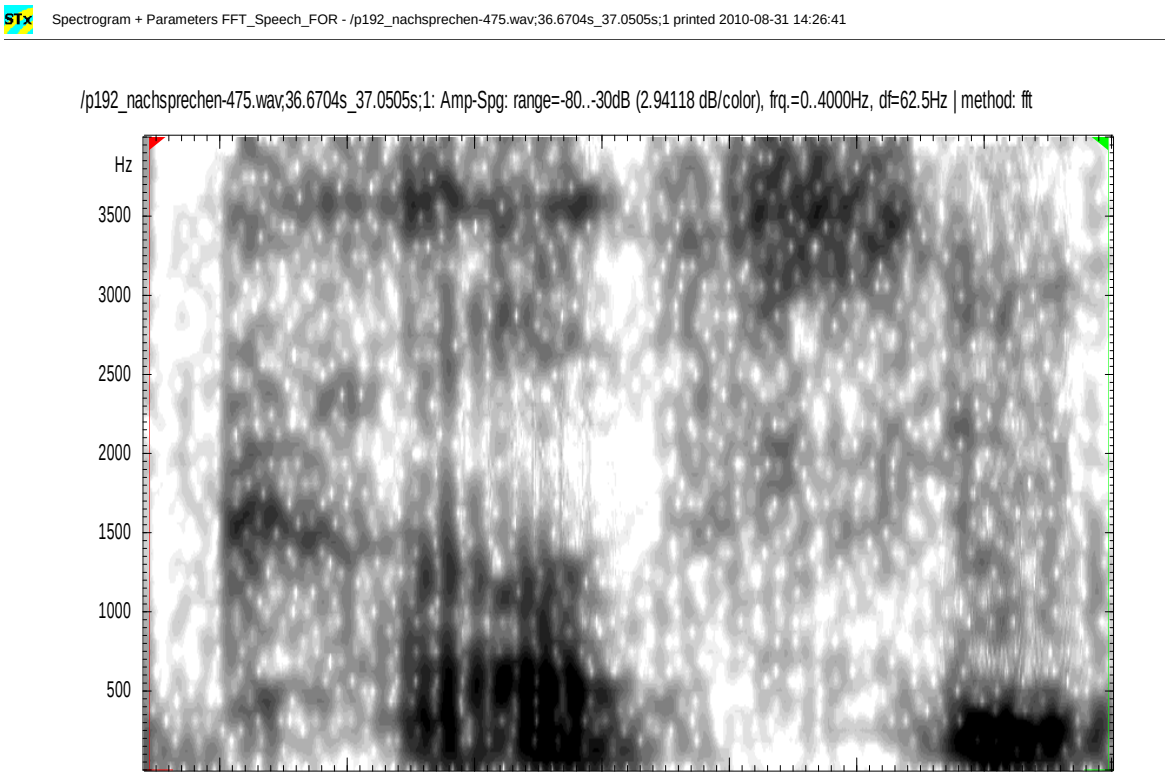
Spektrogramm von “Katzen”



PCM16 44.1kHz

Auswirkungen auf Formantdaten

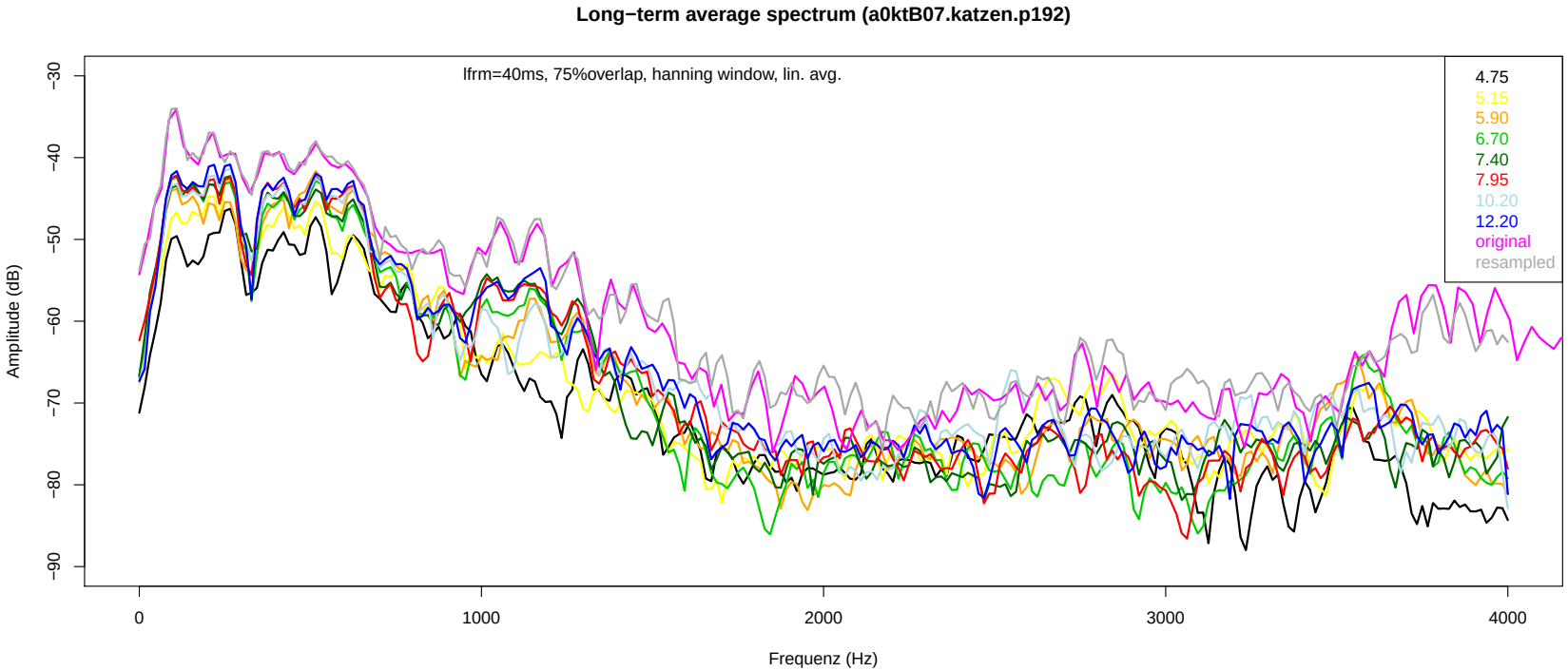
Spektrogramm von “Katzen”



AMR 4.75 kbps

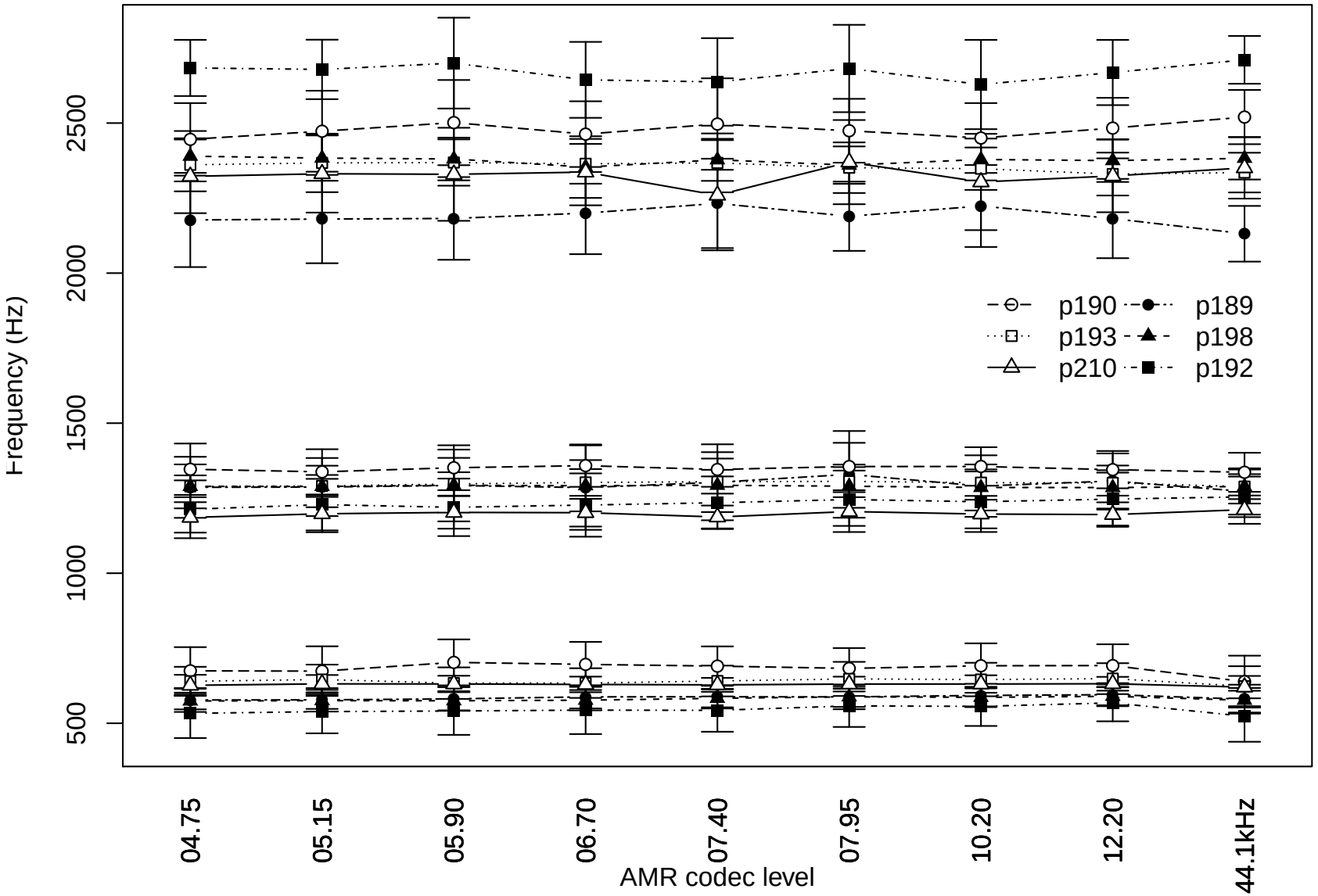
Auswirkungen auf Formantdaten

Gemitteltes Spektrum von /a/

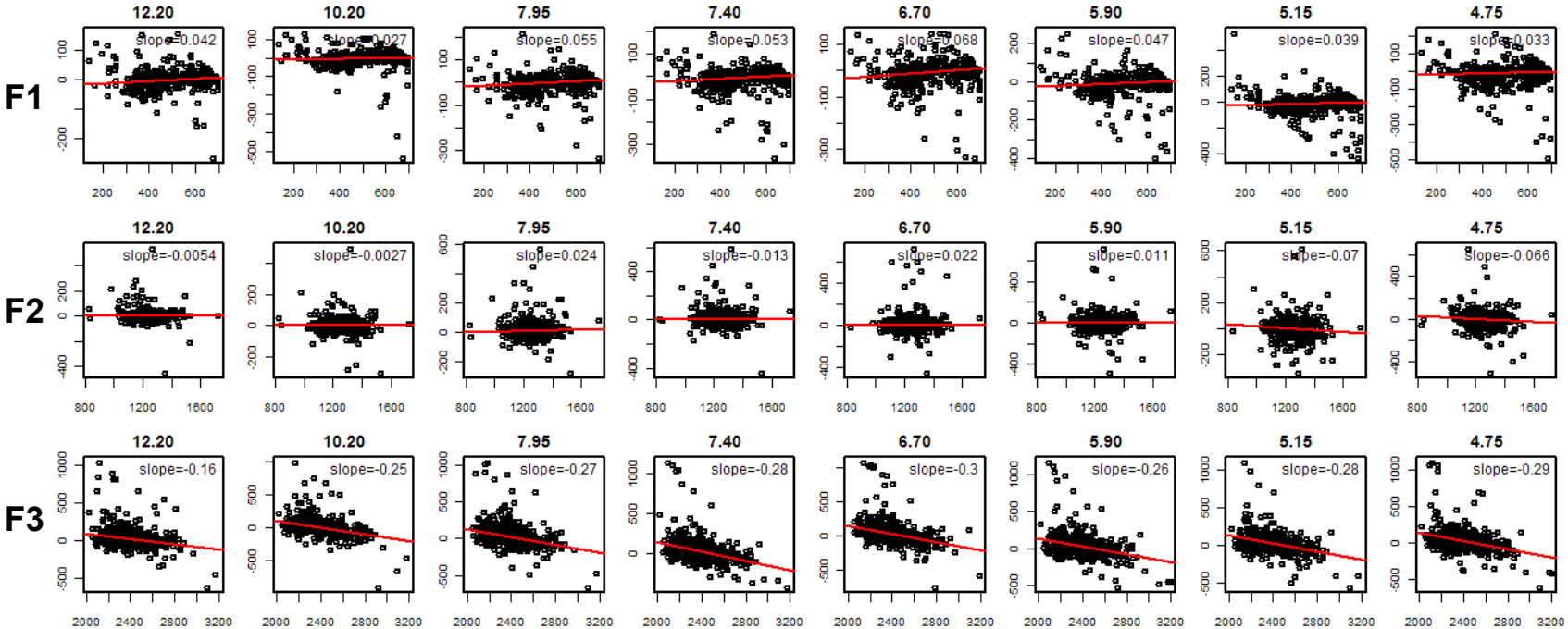


Auswirkungen auf Formantdaten

Interaktion Sprecher-Codecstufe



Auswirkungen
Formantverschiebungen in /a/-Vokalen

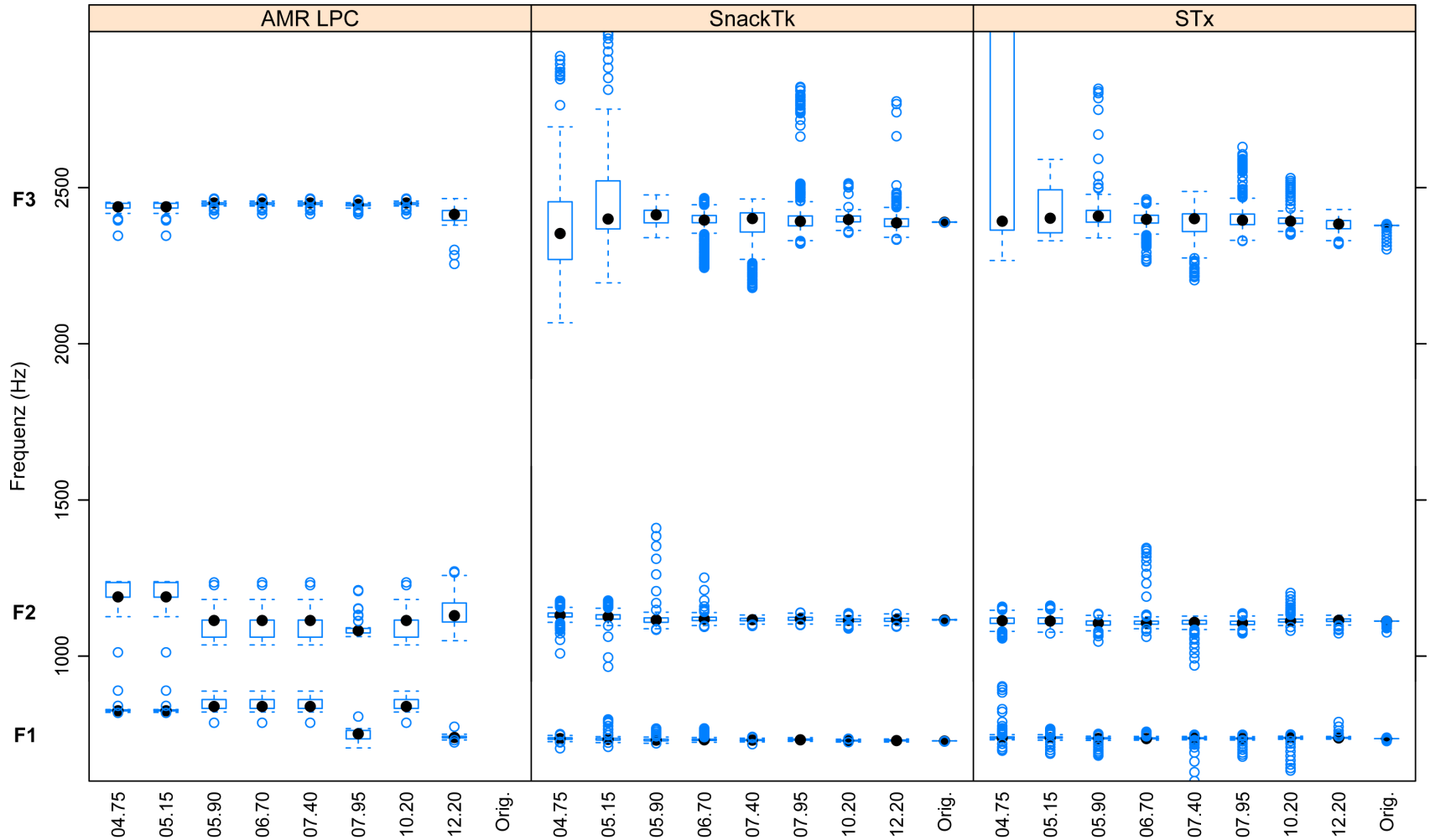


Auswirkungen

Effekte durch AMR und AMR+Bandpassfilterung

F1	Mittelw.	% diff	t-test (p)
Original	299.6		
AMR 12.20	300.2	100.2%	0.7147
BP AMR 12.20	368.3	122.9%	$<$0.001
F2	Mittelw.	% diff	t-test (p)
Original	1946.0		
AMR 12.20	1932.4	99.3%	0.2299
BP AMR 12.20	1802.8	92.6%	$<$0.001
F3	Mittelw.	% diff	t-test (p)
Original	2782.5		
AMR 12.20	2786.7	100.1%	0.7463
BP AMR 12.20	2563.8	92.1%	$<$0.001

Auswirkungen
Synthetisierte Vokale



- Nur geringe Abweichungen durch AMR-Codec an sich
  - Effekte durch zusätzlichen Bandpass erklärbar
- Höhere Fehleranfälligkeit der automatischen Tracker
  - falsche Formantzuweisung bzw. keine Werte aufgrund geringer Amplitude
- Geringer Nutzen durch Rückgriff auf AMR-interne LPC10 Koeffizienten

- Nur geringe Abweichungen durch AMR-Codec an sich
  - Effekte durch zusätzlichen Bandpass erklärbar
- Höhere Fehleranfälligkeit der automatischen Tracker
  - falsche Formantzuweisung bzw. keine Werte aufgrund geringer Amplitude
- Geringer Nutzen durch Rückgriff auf AMR-interne LPC10 Koeffizienten

- Nur geringe Abweichungen durch AMR-Codec an sich
  - Effekte durch zusätzlichen Bandpass erklärbar
- Höhere Fehleranfälligkeit der automatischen Tracker
  - falsche Formantzuweisung bzw. keine Werte aufgrund geringer Amplitude
- Geringer Nutzen durch Rückgriff auf AMR-interne LPC10 Koeffizienten

- Berücksichtigung weiblicher Sprecher (Guillemin & Watson (2006))
- VoIP Telefonie (z.B. Skype/SILK Codec)
- 4G Telefonie / Long-term evolution (LTE)
  - AMR-Wideband / G.722.2 (Bandbreitenerweiterung auf 50-7000 Hz)

- Berücksichtigung weiblicher Sprecher (Guillemin & Watson (2006))
- VoIP Telefonie (z.B. Skype/SILK Codec)
- 4G Telefonie / Long-term evolution (LTE)
  - AMR-Wideband / G.722.2 (Bandbreitenerweiterung auf 50-7000 Hz)

- Berücksichtigung weiblicher Sprecher (Guillemin & Watson (2006))
- VoIP Telefonie (z.B. Skype/SILK Codec)
- 4G Telefonie / Long-term evolution (LTE)
  - AMR-Wideband / G.722.2 (Bandbreitenerweiterung auf 50-7000 Hz)

Vielen Dank
für Ihre Aufmerksamkeit

3GPP (2009). ETSI TS 126 073 ANSI C code for the Adaptive Multi Rate (AMR) speech codec.

Byrne, C., & Foulkes, P. (2004). The 'Mobile Phone Effect' on vowel formants. *International Journal of Speech Language and the Law*, *11*(1), 83–102.

Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton.

Guillemin, B. J., & Watson, C. (2006). Impact of the GSM AMR Speech Codec on Formant Information Important to Forensic Speaker Identification. In P. Warren, & C. I. Watson (Eds.) *Proceedings of the 11th Australian International Conference on Speech Science & Technology*, (pp. 483–488).

Guillemin, B. J., & Watson, C. (2008). Impact of the GSM Mobile Phone Network on the Speech Signal–Some Preliminary Findings. *International Journal of Speech Language and the Law*, *15*(2), 193–218.

Stevens, K. N. (1999). *Acoustic Phonetics*. Cambridge, MA: MIT Press.