# Auswirkungen von Sprachcodecs auf Formantmessungen für Sprechervergleiche

Ewald Enzinger

*Institut für Schallforschung, Österr. Akademie der Wissenschaften, Email: ewald.enzinger@oeaw.ac.at*

## Einleitung

Verschiedene Ansätze für forensische Sprechererkennung basieren auf Vergleichen von Formantmessungen zwischen Tat- und Verdächtigtenaufnahmen [5]. Neben Effekten auf Stimmparameter aufgrund der Sprechsituation [4] sind die Auswirkungen des Telefonkanals von besonderer Relevanz. Eine Studie von Byrne & Foulkes [1] betrachtete die Auswirkungen des gesamten GSM-Mobilfunkkanals einschließlich der Eigenschaften des Handsets auf Formantmessungen bei 12 Sprechern. Sie fanden, dass der erste Formant durchschnittlich um 29% höher war als bei Formantmessungen des direkt aufgenommenen Signals. Eine weitere Studie [2] konzentrierte sich auf den für GSM/UMTS-Mobilfunknetze spezifizierten Adaptive Multi-Rate (AMR) Codec, beschränkte sich jedoch größtenteils auf Aussagen zur Beeinflussung der berechneten Grundfrequenz, während bei Formanten auftretende Effekte nur Beispielhaft dargestellt wurden.

## Der AMR Sprachcodec

Der Adaptive Multi-Rate Codec sieht acht Qualitätsstufen mit verschiedenen Bitraten (4.75, 5.15, 5.90, 6.70, 7.40, 7.95, 10.20, 12.20 kbit/s) vor, zwischen denen während eines Gesprächs auf Basis der Kanalübertragungsqualität gewechselt werden kann. Er basiert auf dem Algebraic Code Excited Linear Prediction (ACELP) Prinzip (Abbildung 1). Das Signal wird nach einer Vorverarbeitung in 20 ms Frames geteilt, von dem nach Anwendung von Fensterfunktionen die Koeffizienten des Linear Prediction-Filters 10. Ordnung mittels Levinson-Durbin-Algorithmus berechnet werden. Das Signal wird in 4 Subframes geteilt, von denen Pitch Delay und Gain des Pitch-Synthesefilters (implementiert über das Adaptive Codebook) bestimmt werden. Nach Entfernen des dadurch enkodierten Beitrags zum gewichteten LP-Residuum wird der optimale Innovationsvektor (Fixed Codebook) bestimmt (Analysis-by-Synthesis). Die LP-Koeffizienten, Indices und Gains werden quantisiert, zum Empfänger übertragen und resynthetisiert.

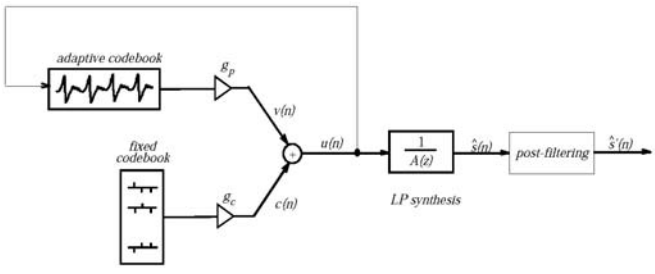


**Abbildung 1:** Vereinfachtes Blockdiagramm des CELP Synthesemodells

## Methoden

Studioaufnahmen von 27 männlichen Wiener Sprechern wurden in den einzelnen Bitraten en- und dekodiert. Bei ausgewählten Vokalsegmenten wurden automatische Formantmessungen unter Verwendung von zwei auf LPC basierenden Trackern durchgeführt, *STx* [7] (46 ms Hamming-Window, 95% overlap, 12 LP Koeff., Präemphase 0,9) sowie *SnackTk/Wavesurfer* [6] (Autokorr., 49 ms $Cos^4$-Window, 10 ms Shift, Präemphase 0,7). STx sucht dabei lokale Maxima des mittels LPC geglätteten Spektrums und bestimmt darauf aufbauend Formantspuren. SnackTk hingegen berechnet die Polstellen des durch LPC geschätzten All-Pol-Modells und bestimmt anschließend die Zuordnung zwischen Polen und Formanten über dynamische Programmierung.

Um den Effekt einer GSM-/UMTS-zu-Festnetz-Übertragung zu untersuchen, wurden weiters die Formanten nach zusätzlicher Telefonkanalsimulation (effektiver Bandpass 300-3400 Hz, POTS) berechnet. Für den Vergleich zwischen Werten der Originalaufnahmen und der kodierten Signale wurden zudem Daten von 6 Sprechern manuell korrigiert. Des Weiteren wurden kurze /ɑ/ und /ɪ/ Vokalsegmente mittels des Klatt-Synthesizers erzeugt und derselben Prozedur unterzogen. Tabelle 1 zeigt die verwendeten Parameter.

**Tabelle 1:** Parameter für synthetisierte Vokale

| | f0 (Hz) | F1 (Hz) | F2 (Hz) | F3 (Hz) |
|---|---|---|---|---|
| /ɑ/ | 124 | 730 | 1090 | 2440 |
| /ɪ/ | 136 | 270 | 2290 | 3010 |

## Effekte bei Studioaufnahmen

Abbildung 2 zeigt Streudiagramme der manuell korrigierten Formantmessungen in /a/-Vokalen der Originalaufnahmen und der Differenz zwischen diesen und von STx automatisch berechneten, nicht korrigierten Werten nach En- und Dekodierung, sowie mittels iterated re-weighted least squares (IWLS) berechneten Regression.

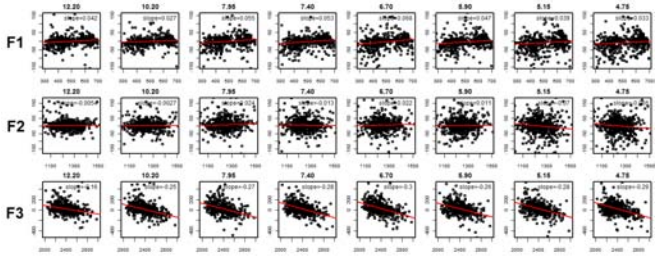


**Abbildung 2:** Formanten in /a/-Vokalen. Studioaufnahme (X-Achse) vs. Abweichung durch Codec in Hz (Y-Achse).

Für F1 und F2 von /a/-Vokalen ergaben sich leichte frequenzabhängige Verschiebungen, Abweichungen bei F3 zeigen jedoch eine negative Frequenzabhängigkeit sowie allgemein eine höhere Varianz. Bei /ɪ/-Vokalen zeigt sich ein ähnliches Bild, jedoch weist F2 zusätzlich eine starke negative Frequenzabhängigkeit auf. Diese Effekte lassen sich bei allen Bitraten erkennen. Schwierigkeiten treten bei automatischen Formanttrackern auf, vor allem durch falsche Formantzuordnung bei F2 und F3 aufgrund der verringerten Amplitude der spektralen Peaks im höheren Frequenzbereich. Abbildung 3 zeigt die für alle Codecstufen manuell korrigierten Werte von F1-3 von /a/-Vokalen von 6 Sprechern. Zwischen den Bitraten zeigen sich nur geringe Unterschiede, etwa ist F1 leicht höher als in der Studioaufnahme.

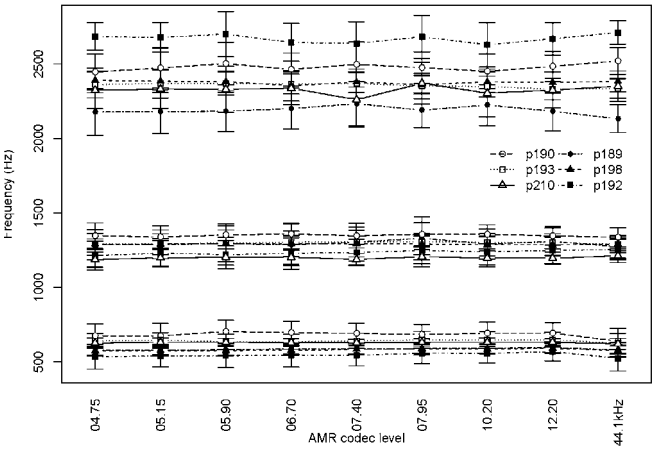


**Abbildung 3:** Manuell korrigierte Formantmessungen bei /a/-Vokalen von 6 Sprechern.

## Effekte durch Telefonkanalsimulation

Tabelle 2 zeigt den Effekt durch Kodierung sowie zusätzlich simuliertem Telefonkanal auf Formantmessungen durch STx in /ɪ/-Vokalen. Während sich geringe prozentuelle Unterschiede zwischen Mittelwerten der Original- und kodierter Aufnahme ergeben, zeigt sich eine hohe Beeinträchtigung durch die Filterung, wie Untersuchungen über Festnetztelefonie gezeigt haben [3].

**Tabelle 2:** Effekte durch AMR sowie AMR mit zusätzlichem simulierten Telefonkanal

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

## Synthetisierte Vokale

Abbildung 4 zeit Formantmessungen bei einem synthetisierten stationären /a/-Vokal durch SnackTk und STx sowie Polstellen des durch die quantisierten 10 LP-Koeffizienten gegebenen All-Pol-Filters. Während F1 und F2 bei den Trackern den Syntheseparametern großteils entsprechen, zeigen sich bei F3 höhere Variabilität, besonders bei niedrigen Bitraten. Die F1 und F2 zugeordneten Polstellen des LP-Filters zeigen teils größere Abweichungen von den Eingabeparametern, bei F3 ergeben sich hingegen genauere Werte als bei den Trackern.

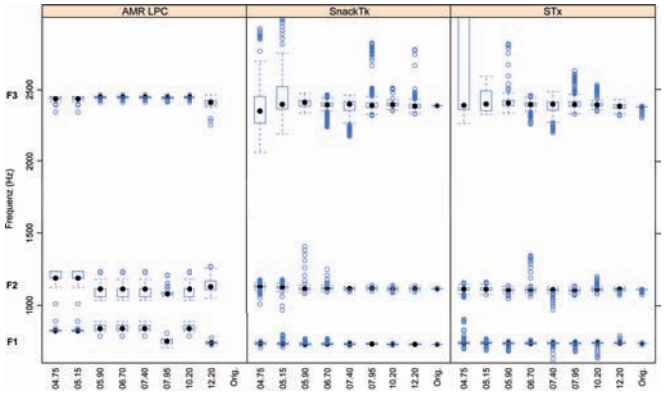


**Abbildung 4:** Formantmessungen in synthetisierten, stationären /a/-Vokal

## Diskussion

Bei den gegebenen Daten zeigen sich v.a. bei F3 durch den Codec verursachte Frequenzverschiebungen. Bei manuell korrigierten Daten zeigten sich nur geringe Unterschiede, automatische Formanttracker weisen jedoch eine höhere Fehleranfälligkeit (falsche Formantzuweisung bzw. keine Werte aufgrund geringer Amplitude) auf. Mit simuliertem Telefonkanal zeigen sich teilweise große Abweichungen, vor allem bei F1 in /ɪ/.

## Literatur

[1] Byrne, C. und Foulkes, P. The 'Mobile Phone Effect' on vowel formants. *Int. J. Speech Language and the Law*, 11(1):83–102, 2004.

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

[3] Künzel, H. J. Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies. *Forensic Linguistics*, 8(1):80–99, 2001.

[4] Rose, P. und Simmons, A. F-pattern variability in disguise and over the telephone - comparisons for forensic speaker identification. In *Proc. SST 96*, 121–126, Flinders University, Adelaide, 1996.

[5] Rose, P. Forensic speaker identification. Taylor & Francis, 2002.

[6] SnackTk, `http://www.speech.kth.se/snack/`

[7] STx, `http://www.kfs.oeaw.ac.at/`