Bestimmung von Maskierung mit hoher zeitlicher Auflösung in Audiocodecs mittels Subband Instantanfrequenz-Analyse

Nils Koppaetzky[1], Stephan D. Ewert, Birger Kollmeier, Volker Hohmann
Institut für Physik, Medizinische Physik, 26111 Oldenburg, Deutschland
[1] *Email: nils.koppaetzky@uni-oldenburg.de*

Einleitung

Perzeptive Audiokodierung ist eine verlustbehaftete auf Irrelevanz-Reduktion basierende unumkehrbare Kompressionsmethode für Audiosignale. Ihr Prinzip ist die Reduktion der Bitrate durch Reduktion der Kodierungsgenauigkeit. Das dabei durch den Kodierungsfehler erzeugte Quantisierungsrauschen wird spektro-temporal so verteilt, dass es vom menschlichen Gehör nicht wahrgenommen wird. Zu diesem Zweck nutzen die Kodierungsalgorithmen psychoakustische Maskierungs-Modelle. Aktuelle Forschungsergebnisse zeigen, dass das menschliche Gehör sowohl eine hohe Zeit- als auch eine hohe Frequenzauflösung besitzt. Trotzdem verwenden die meisten "Standard" Audioencoder klassische spektrale Maskierungsmodelle. Diese erreichen nur unter Verlust einer hohen Zeitauflösung die benötigte hohe spektrale Auflösung. Die Anwendung eines aktuellen Gehörmodells welches die excellente Zeit-Frequenzauflösung des menschlichen auditorischen Systems simuliert, könnte die Effektivität der Datenreduktion bzw. die Audioqualität erhöhen. Zur Analyse des Effektes einer verbesserten Zeitauflösung wurde im Rahmen dieser Arbeit ein aktuelles Filterbankmodell mit Subband-Instantanfrequenz-Kontrolle getested. Dabei wurde die Fähigkeit des Modells zur Schätzung von Maskierungsschwellen unter Berücksichtigung der Tonalität und des "spread of masking" getestet und seine Performance im Zusammenhang mit einem einfachen Basiscoder bei akzeptabler Audioqualität bestimmt.

Motivation

Klassische Maskierungs-Modelle verwenden meist eine FFT-Filterbank zur spektro-temporalen Analyse des zu kodierenden Signals. Diese erreicht die benötigte hohe spektrale Auflösung zu Lasten einer schlechten Zeitauflösung. Desweiteren ist die Filterbreite einer FFT-Filterbank über das gesamte Spektrum konstant, woraus sich eine frequenzunabhängige starre Zeit-Frequenzauflösung ergibt. Im auditorischen System hingegen nehmen die Filterbreiten, und damit auch die Zeitauflösung, mit steigender Mittenfrequenz der Filter in guter Näherung logarithmisch zu. Ein stärker physiologisch motivierter Ansatz bei der Wahl der Filterbank und eine zeitbasierte Merkmalextraktion könnte die Effektivität der Datenreduktion bzw. die Audioqualität erhöhen (vgl. [2]).

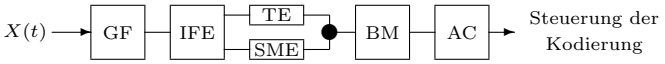

Abbildung 1: Blockschaltbild eines Filterstranges des Instantanfrequenz gesteuerten Gehörmodells bestehend aus Gammatonfilter (GF), Instantanfrequenz-Schätzung (IFE), Tonalitäts-Schätzung (TE), "spread of maskin"-Schätzung (SME), Bitmaske (BM) und Aliasing-Kompensation (AC).

Modell

Ein Filterstrang des auf Instantanfrequenz- (i.F. f_{inst} genannt) Analyse basierenden Maskierungs-Modells ist in Abbildung 1 schematisch dargestellt. Zur Simulation der auditorischen Filterung wird eine Gammatonfilterbank (GF) nach [1] verwendet. Nach der Filterung liegt am Ausgang eines jeden Filters ein schmalbandiges Signal $S'(t)$ vor, welches signaltheoretisch als Multiplikation einer Amplitude und eines oszillierenden Terms in Abhängigkeit von der Zeit $S'(t) = A(t) \cdot \exp -i\phi(t)$ dargestellt werden kann. f_{inst} ist dann definiert als

$$f_{inst}(t) = \frac{d}{dt}\frac{\phi(t)}{2 \cdot \pi} \tag{1}$$

und wird am Ausgang eines jeden Filters geschätzt (IFE in Abb. 1). Die folgenden Eigenschaften von f_{inst} können genutzt werden um Maskierungsschwellen (signal-to-masked ratio SMR) und daraus die Bitverteilung, pro 6 dB SMR wird 1 Bit für die Kodierung benötigt, zu bestimmen.

- Der Mittelwert von f_{inst} zeigt die Frequenz höchster Energie im Signal an.
- Die Varianz von f_{inst} kann als Maß der Tonalität eines Signals genutzt werden.

Der SMR kann aus f_{inst} über die in Abbildung 2 dargestellten Funktionen extrahiert werden. Dabei bedeutet ein kleiner SMR die Möglichkeit zur Kodierung mit einer geringen Bitrate. Abb. 2 l. zeigt den für die Kodierung benutzen SMR in Abhängigkeit der auf die Filterbreite normierten Varianz von f_{inst}. Dabei zeigt eine Varianz von 0 ein vollständig tonales Signal an, welches einen SMR von 36 dB benötigt, während ein vollständig rauschhaftes Signal nur einen SMR von 6 dB benötigt (Asymmetrie der Maskierung) und durch den Fußpunkt der Funktion dargestellt. Dazwischen wird ein linearer Zusammenhang angenommen.
Abb. 2 r. zeigt den Zusammenhang zwischen dem auf die Filterbreite normierten Mittelwert über 5 ms von f_{inst} und dem für die Kodierung benötigten SMR. Hier-

bei zeigt eine starke Abweichung des Mittelwertes von f_{inst} von der Filtermittenfrequenz ein maskierendes Signal in einem Nachbarfilter an ("spread of masking"). Dabei bedeutet die obere Kante des Trapezes ein unmaskiertes (36 dB SMR), die Flanken ein teilweise maskiertes und der Bereich außerhalb des Trapezes ein vollständig maskiertes (0 dB SMR) Signal innerhalb des Filters. Die geschätzten SMR-Werte werden in einer Bitmaske gespeichert die nach einer, auf den Encoder abgestimmten, Aliasing Kompensation (AC) die Kodierung steuert.
Die optimale Lage der Fußpunkte bzw. Flanken der Funktionen muss empirisch bestimmt werden.

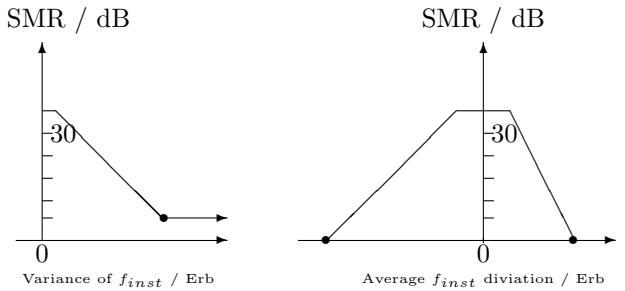

Abbildung 2: f_{inst} zu SMR Funktionen: Varianz von f_{inst} zu SMR Funktion als Maß für die Tonalität eines Signals (l.) zur Ausnutzung der Asymmetrie der Maskierung; Abweichung von f_{inst} von der Mittenfrequenz des zu Grunde liegenden Filters als Maß für "spread of masking" (r.);

Methode

Die empirische Bestimmung der optimalen Lage der Fußpunkte bzw. der Flanken der Funktionen aus Abbildung 2 erfolgt über die Messung des sogenannten Objective Difference Grade (ODG) und der Bitrate für verschiedene Fußpunkte bzw. Flankenlagen. Der ODG wird über eine Software geschätzt (vgl.) und ist ein Maß für die perzeptive Qualität von Audiosignale welche in 5 Stufen von 0 (unwahrnehmbare Verzerrungen) bis -4 (stark störende Verzerrungen) eingeteilt ist. Die ermittelten Bitraten und ODGs wurden für verschiedene Fußpunkte, bzw. Flankenlagen gegeneinander aufgetragen. In dieser Messreihe wurde zunächst der optimale Fußpunkt der Funktion in Abbildung 2.l bestimmt, um von diesem Ausgehend die optimale Lage der linke Flanke der rechten Funktion in Abbilung 2.r zu bestimmen. Für die rechte Flanke wurde ein Fußpunkt von 10 Erb festgesetzt, da hochfrequente Signale niederfrequente kaum maskieren.

Ergebnisse

Das f_{inst} basierte Maskierungs-Modell kann mit den in Abbildung 2 dargestellten f_{inst} / SMR Funktionen sowohl zur Schätzung des SMR der Asymmetrie der Maskierung als auch für die Schätzung des SMR des "spread of masking" in Audiosignalen zum Zweck der perzeptiven Audiokodierung genutzt werden. Abbildung 3 zeigt die ODGs/Bitraten Paare für verschiedene Fußpunkte, bzw. Flankenlagen, für die in Abschnitt beschriebene Messung. Beide Graphen weisen einen Bereich flacher Steigung auf. In diesen Bereichen ist die Reduktion der Bitrate bei moderatem Verlust der perzeptiven Qualität

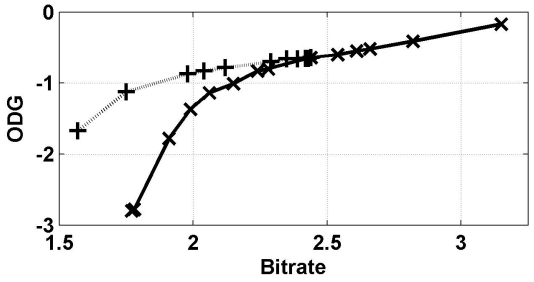

Abbildung 3: ODG über Bitrate für Variation des Fusspunktes der Tonalitätsfunktion in Abb.2.l (durchgezogene Linie) und Variation der Lage der linken Flanke in Abb.2.r bei optimalem Fusspunkt der Tonalitätsfunktion (gestrichelte Linie). Verwendetes Singal: Die ersten 15 Sekunden aus "La valse d'Amelie (Orchester Version)" aus dem Soundtrack zu "Die Fabelhafte Welt der Amélie" (2001).

möglich. Dies zeigt das es sich um valide Schätzungen des SMR und um valide psychoakustisch basierte Maße handelt. Eine Anwendung beider Maße sorgt für eine höhere Kompression bei gleicher Audioqualität. Die Festlegung der Fußpunkte und Flanken der f_{inst} / SMR Funktionen hat gezeigt das diese Signalabhängig sind. Insgesamt bietet die f_{inst} Analyse die Möglichkeit der Merkmalextraktion bei gleichzeitig hoher Zeit- und Frequenzauflösung. Bei perzeptiv nicht störenden Verzerrungen konnte das Modell im Zusammenhang mit einem rudimentären, in MATLAB implementierten Encoder Bitraten von durchschnittlich 2.58 B/Sample (123,84 kB/Sec bei 48 kHz) mono über ein Set von 6 unterschiedlichsten Testsignalen erzielen.

Diskussion

Obwohl gezeigt wurde, dass sich das f_{inst} basierte Modell zur Schätzung von SMR-Schwellen in perzeptiven Audio-Encodern eignet konnte auf Grund des rudimentären Test-Encoders keine aussagekräftige quantitative Vergleichsmessung zwischen ihm und Modellen der Standard-Encoder erfolgen. Die wesentlich unausgereiftere Signalverarbeitung des Test-Encoders macht einen solchen direkten Performancevergleich und damit das Aufzeigen von vermuteten Vorteilen des f_{inst} basierten Modells unmöglich.
Fazit: Für einen quantitativen Performancevergleich des f_{inst} basierten Maskierungs-Modells mit den Modellen der Standard-Encoder unter Ausschöpfung aller vermuteten und hier teilweise gezeigten Vorteile die das Modell bietet, müsste es in einen Standard-Encoder integriert werden.

Literatur

[1] Hohmann, V.: Frequency analysis and synthesis using a Gammatone filterbank. Acta Acustica united with Acustica 88 (2002), 433-442

[2] Hohmann, V. and Kollmeier, B.: A nonlinear auditory filterbank controlled by sub-band instantaneous frequency estimates. Hearing - From Sensory Processing to Perception, Springer, 2007