Sequentielle Monte-Carlo-Verfahren zur Grundfrequenzerkennung überlagerter realer und synthetischer Vokale

Ronny Meyer, Johannes Nix, Volker Hohmann

Medizinische Physik,Universität Oldenburg, D-26111 Oldenburg, Email: ronny.meyer@mail.uni-oldenburg.de

Zusammenfassung

Ein Verfahren zur Grundfrequenzbestimmung und -verfolgung von zwei überlagerten Sprechern auf der Grundlage eines Sequentiellen-Monte-Carlo-Filters (SMC-Filter) als Grundlage zur merkmalskombinierten Computational Auditory Scene Analysis wird vorgeschlagen. Es wird das Synchrogramm-Verfahren zur Messung der zwei wahren Grundfrequenzen beschrieben. Es werden Ergebnisse des Algorithmus für die Überlagerung realer und synthetischer Vokale gezeigt.

Einleitung

Psychoakustische Messungen zeigen, dass Menschen eine Vielzahl unterschiedlicher Signalmerkmale, wie z.B. gemeinsamer Anstieg der Amplitude in verschiedenen Frequenzbändern, Kohärente Amplituden- und Frequenzmodulation in verschiedenen Frequenzbändern, Richtungsinformation, Harmonische Struktur (Grundfrequenz) zur akustischen Szenenanalyse nutzen [1]. Des Weiteren scheint das Gehör Modelle der Dynamik der Quellen zu nutzen. Algorithmen, die beide Eigenschaften berücksichtigen, fehlen bisher oder sind nur rudimentär entwickelt. Eine Möglichkeit, sowohl die Kombination der verschiedenen Signalmerkmale, als auch den Einbezug der Quellendynamik zu realisieren, bieten die Sequentiellen-Monte-Carlo-Filter (SMC-Filter)[2]. Die dieser Arbeit zugrunde liegende Hypothese ist, dass das Gehör die Grundfrequenz als primären Parameter zur Verbindung spektraler Anteile nutzt. Klapuri gibt in [3] eine Übersicht vorhandener Algorithmen zur Grundfrequenzschätzung mehrerer Sprecher. Bisherige Ansätze sind nicht in allen Situationen erfolgreich. Durch die Hinzunahme weiterer Signalmerkmale bei der Grundfrequenzschätzung, insbesondere der Schalleinfallsrichtung soll die Schätzung zukünftig verbessert werden.

Methode

Sequentielle Monte-Carlo-Filter

Wir betrachten das System Σ_2, dass aus zwei Sprachsignalen besteht. Die Beschreibung des Systems erfolgt im zweidimensionalen Zustandsraum $\mathcal{R}$ der aus allen möglichen Kombination von Grundfrequenzen besteht. Als Zustand x des Systems wird das Paar der Grundfrequenzen f^1, f^2 der beiden Sprecher gewählt. Zu jedem Zeitpunkt k befindet sich das System Σ_2 in einem Punkt $x_k = (f_k^1, f_k^2)$ des Zustandsraumes. Dieser Zustand ist nicht direkt, sondern nur über Messungen z_k zugänglich. Eine zweckmässige Beschreibung der zeitliche Entwicklung erfolgt im Zustandsraum durch die Wahrscheinlichkeitsdichtefunktionen $p(x_k \mid x_{k-1})$ (*Systemdynamik*), die die zeitiche Übergangswahrscheinlichkeit der Zustände beschreibt und $p(z_k \mid x_k)$ (*Messdynamik*), die die Wahrscheinlichkeit für eine Messung z_k angibt, wenn sich das System im Zustand x_k befindet. Alles, was ein Beobachter über den Zustand des Systems zur Zeit k durch Messungen $z_{1:k}$ bis zur Zeit k wissen kann, ist in der Wahrscheinlichkeitsdichte $p(x_k \mid z_{1:k})$ enthalten. Die Grundidee der SMC-Methoden ist die diskrete Approximation dieser Wahrscheinlichkeitsdichte im Zustandsraum durch eine Menge von N Partikeln $x^i{}_k,\ i = 1, \ldots, N$ [2]

$$p(x_k \mid z_{1:k}) \approx \sum_{i=1}^{N} w_k^i \delta(x_k - x_k^i). \tag{1}$$

Synchrogramm

Für den SMC-Filter wird ein Beobachtung z benötigt. Als Basis dafür wird das Synchrogramm $S(T,t)$ verwendet. Das Synchrogramm gibt die periodische Intensität eines Signals als Funktion der Periode T und Zeit t an. Das Synchrogramm $S(T,t)$ zu einem Zeitschritt wird als Synchrum $S(T)$ bezeichnet. Ziel des Synchrogramms ist es, durch mehrfache periodensynchrone Mittelung von Signalkomponenten Störanteile des Signals zu unterdrücken. Sei $x(n) = y_1(n) + y_2(n)$ ein diskretes Signal, das eine Überlagerung zweier periodischer Signale ist. Das Signalmodell des Synchrums geht davon aus, dass sich die einzelnen harmonischen Anteile, sowie alle Rauschanteile der beiden Signale $y_i;\ i = 1, 2$ additiv überlagern. Um die in dem Signal $x(n)$ enthaltenen Perioden zu gewinnen, wird in einem Abtastpunkt t für jede Periode $T_i \in \{T_1, \ldots, T_{N_T}\}$ die Zeitfunktion

$$v_{t,T_i}(n) = \Big(\sum_{k=0}^{\frac{m}{2}-1} x(t+n+kT_i-1) + \sum_{k=1}^{\frac{m}{2}} x(t+n-kT_i-1)\Big) \tag{2}$$

berechnet. Die Funktion (2) ist das Mittel über m um t zentrierte Signalabschnitte der Länge T_i. Von der Funktion (2) wird die Leistung berechnet

$$S(T_i,t) = \frac{1}{T_i} \sum_{n=1}^{N=T_i} (v_{t,T_i}(n))^2. \tag{3}$$

Das Synchrum zeigt deutliche Maxima an den im Signal enthaltenden Grundperioden, aber auch an deren Vielfachen. Dies macht die Bestimmung der Grundperioden schwierig. Um die wahren Grundperioden zu finden, wird

ein von Parsons vorgeschlagenes Verfahren [4] für das Synchrogramm adaptiert. Dazu wird ein gewichtetes Histogramm $R_n(T)$ über die 25% größten Synchrumwerte, sowie die zugehörigen Vielfachen und Teiler der Perioden gebildet. Als erste Koordinate der Messung z_k wir dann gewählt $z_k^1 = 1/\hat{f}^1 = \hat{T}_0^1 = \arg\max R_n(T)$ Alle Vielfachen und Teiler dieser Grundfrequenz werden aus dem Histogramm gelöscht; man erhält das Resthistogramm $R_{res,n}(T)$. Als zweite Koordinate der Messung z_k wird dann gewählt $z_k^2 = 1/\hat{f}^2 = \hat{T}_0^2 = \arg\max R_{res,n}(T)$.

Ergebnisse

Um Oktavvertauschungen auszuschließen, wird der Zustandsraum $\mathcal{R}$ in den Frequenzbereich der ersten Oktave (440-880 Hz) zyklisch abgebildet. Dieser Zustandsraum wird mit $\mathcal{R}_{okt}$ bezeichnet. Die Systemdynamik für einen Sprecher wird mittels des Grundfrequenzerkenners YIN [5] aus einer Datenbasis, die 14 Stunden gesprochener Sprache verschiedener Sprecher umfasst erstellt. Abbildung 1 zeigt die empirisch bestimmte Systemdynamik für den Zustandsraum $\mathcal{R}_{okt}$. Gezeigt ist das logarithmierte zweidimensionale Histogramm mit je 60 Kategorien für die *a priori* Grundfrequenz $f_{okt,k}$ zur Zeit k und für die *a posteriori* Grundfrequenz $f_{okt,k+1}$ zur Zeit k+1. Es zeigt sich, dass der zeitliche Übergang der Grundfrequenz nicht beliebig ist und nur sehr selten sprunghaft verläuft. Die Systemdynamik ist nicht gaußsch. Abbildung 2 (oben) zeigt das Ergebnis der

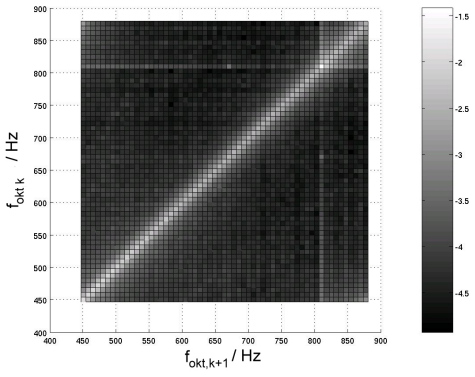

Abbildung 1: Logarithmiertes zweidimensionales Histogramm mit je 60 Kategorien für die *a priori* Grundfrequenz $f_{okt,k}$ zur Zeit k und für die *a posteriori* Grundfrequenz $f_{okt,k+1}$ zur Zeit k+1 für den Zustandsraum $\mathcal{R}_{okt}$.

SMC-Schätzung für die Überlagerung zweier synthetischer Vokale. Die wahre Grundfrequenz f_{okt}^1 steigt von 560 Hz auf 800 Hz; f_{okt}^2 bleibt konstant 480 Hz. Die Schätzung durch den SMC-Filter zeigt erwartungstreue Ergebnisse. Abbildung 2 (unten) zeigt das Ergebnis der Grundfrequenzschätzung mittels SMC-Filter für die Überlagerung realer Vokale. Die Signale habe die Grundfrequenz $f_{okt}^1 \approx 688$ Hz und $f_{okt}^2 \approx 824$ Hz. Die Schätzung durch den SMC-Filter zeigt auch hier erwartungstreue Ergebnisse.

Schlussfolgerungen und Ausblick

Der vorliegende Algorithmus schätzt die Grundfrequenz einzelner gemischter realer und synthetischer Vokale erwartungstreu im Zustandsraum $\mathcal{R}_{okt}$. Durch Erweiterung des Zustandsraumes durch räumliche Koordinaten, wie z.B. Azimut und Elevation, lässt sich eine Integration der räumlichen Richtungsinformation in den vorliegenden SMC-Filter bewerkstelligen, die in folgenden Arbeiten durchgeführt werden wird.

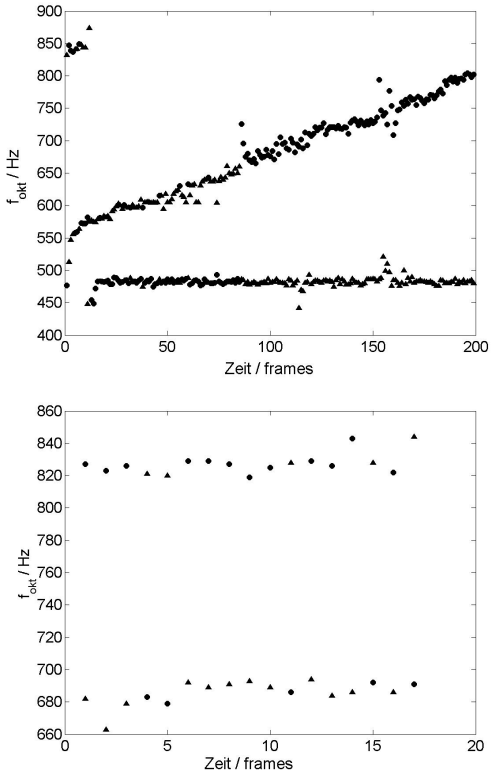

Abbildung 2: Ergebnis der Grundfrequenzschätzung mittels SMC-Partikelfilter für den Zustandsraum $\mathcal{R}_{okt}$ für die Überlagerung zweier synthetischer Vokale (oben) und zweier realer Vokale (unten).

Literatur

[1] A.S. Bregman, „Auditory scene analysis: Hearing in complex environments," in *Thinking in Sound*, S. McAdams and E. Bigand. Clarendon Press, 1993, 10-36.

[2] Sanjeev Arulampalam, Simon Maskell Neil Gordon and Tim Clapp„A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking", in *IEEE Transactions on Signal Processing*, vol.50, no. 2, feb. 2002, 174-188.

[3] Anssi Klapuri, Signal Processing Methods for the Automatic Transcription of Music, Tampere University of Technology, April 2004. `http://www.cs.tut.fi/sgn/arg/klap/phd/klap_phd.pdf`

[4] T.W. Parsons, Separation of speech from interfering speech by means of harmonic selection, JASA, vol. 60,911–8, 1976.

[5] A. de Cheveigné and H. Kawahara,YIN, a fundamental frequency estimator for speech and music, JASA, vol. 111,1917–1930,2002.