Grundfrequenzbestimmung unter Berücksichtigung linearer Frequenzänderungen

Peter Birkholz
Institut für Computergraphik, Fachbereich Informatik, Universität Rostock; Email: piet@informatik.uni-rostock.de

Einleitung

Ein häufig für die Grundfrequenzbestimmung verwendetes Verfahren ist die Autokorrelationsanalyse. Dabei wird jeweils für einen kurzen Sprachsignalausschnitt (Frame) die Autokorrelationsfunktion (AKF) berechnet und in dieser die Position des Maximums bestimmt, das der Grundperiode T_0 entspricht. Die Grundfrequenz $F_0 = 1/T_0$ kann dann als Schätzwert an der Stelle betrachtet werden, an der sich der Frame befindet.

Bei diesem Verfahren wird davon ausgegangen, dass sich das (stimmhafte) Sprachsignal innerhalb des Frames nur geringfügig ändert, seine Periodizität also weitestgehend beibehält. Für Sprache ist dies jedoch oft nicht der Fall, da sich neben der Signalform, die u.a. durch die Resonanzen des Vokaltrakts bestimmt wird, sowohl die Signalamplitude als auch die Grundfrequenz rasch ändern können. Das führt häufig dazu, dass die der Grundperiode entsprechende Spitze in der AKF nicht mehr eindeutig vom Grundfrequenzbestimmer (GFB) erkannt wird und dadurch Schätzfehler gemacht werden. Im vorliegenden Beitrag wird gezeigt, wie AKF-basierte Verfahren trotz relativ langer Frames (hier: 50 ms) robust auf solche Änderungen reagieren können.

Die Methode

Periodizitätsfunktion

Der Ausgangspunkt für unsere Überlegung ist die folgende instationäre Kurzzeit-AKF f_λ, bei der jeder Abtastwert von f_λ einer eigenen Normierung unterzogen wird[1]:

$$f_\lambda = \left(\sum_{i=0}^{N-1} x_i x_{i+\lambda} \right) / \sqrt{E_0 \cdot E_\lambda} \qquad \text{mit } E_\lambda = \sum_{i=0}^{N-1} x_{\lambda+i}^2 \tag{1}$$

Dabei ist N die „Integrationslänge“ und λ ($0 \le \lambda < N$) die relative Verschiebung. Die Anzahl der Abtastwerte x_i im Frame ist demnach folglich $W=2N-1$. Durch die Normierung haben Änderungen der Energie des Signals keinen Einfluss auf f_λ, d.h. Amplitudenänderungen innerhalb des Frames werden bereits berücksichtigt.

Die Abbildung 1 (b) zeigt die Funktion f_λ für den in (a) abgebildeten Frame. Der GFB hat die Aufgabe, die Spitze zu bestimmen, die der Grundperiode T_0 entspricht. Die Bestimmung der Position des absoluten Maximums von f_λ genügt hier nicht, da zunächst die Spitze bei $\lambda=0$ stets den höchsten Wert (=1) hat, und außerdem eine der Spitzen bei $2T_0$, $3T_0$, ... geringfügig höher sein könnte als die bei T_0. Um die untere Grenze für den T_0-Suchbereich wegen der Spitze bei $\lambda=0$ nicht einschränken zu müssen, benutzen wir im Folgenden statt f_λ die Funktion s_λ, die sich wie folgt berechnet:

$$s_\lambda = \begin{cases} 1 & \text{, wenn } \lambda = 0 \\ r_\lambda / \left(\frac{1}{\lambda} \sum_{i=1}^{\lambda} r_i \right) & \text{sonst} \end{cases} \tag{2}$$

Hierin ist $r_\lambda=1-f_\lambda$, was bedeutet, dass die Spitzen in f_λ lokalen Minima (Kerben) in r_λ entsprechen. Die in Gl. (2) durchgeführte Normierung wurde von Cheveigné und Kawahara[2] vorgeschlagen und hat die Eigenschaft, dass in s_λ die Kerbe bei $\lambda=0$ nicht mehr existiert, während der Rest von s_λ der Funktion r_λ sehr ähnlich ist. Diese Normierung bereitet gleichzeitig den Weg für den „Zeitverzerrungsschritt“ im nächsten Abschnitt. Die Funktionen r_λ und s_λ sind in Abbildung 1 (c) und (d) für den in (a) abgebildeten Frame dargestellt. s_λ werden wir im folgenden auch als Periodizitätsfunktion (PF) bezeichnen.

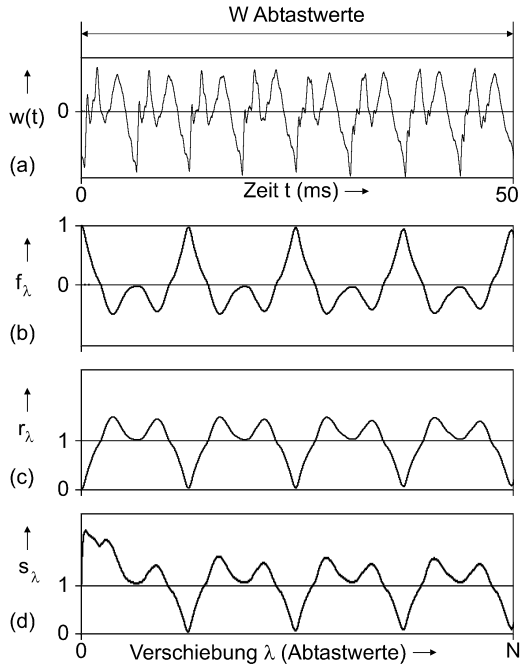

Abbildung 1: Stufen für der Berechnung der Periodizitätsfunktion in (d) für den in (a) gezeigten Frame.

Um mit Hilfe der PF die Grundperiodenkerbe zu bestimmen, gehen wir wie folgt vor. Zuerst wird der Zeitpunkt τ_{min} (in Sekunden) des absoluten Minimums von s_λ ermittelt. Da τ_{min} jedoch nicht notwendigerweise T_0 entsprechen muss, sondern auch ein ganzzahliges Vielfaches von T_0 sein kann, bestimmen wir alle lokalen Minima *vor* τ_{min}, deren Funktionswerte denjenigen an der Stelle τ_{min} maximal um einen kleinen Schwellwert (der Wert 0.1 hat sich als günstig erwiesen) übersteigen. Wenn solche lokalen Minima existieren, dann wird das mit der kleinsten Verschiebung für T_0 ausgewählt. Ansonsten setzen wir $T_0 = \tau_{min}$.

Zeitverzerrung

Wir wollen nun die Auswirkungen von Grundfrequenzänderungen auf die PF betrachten. Auf der linken Seite in Abbildung 2 sind dazu drei Frames dargestellt. Das Signal im Frame (a) hat eine konstante F_0, während sich die F_0 im Frame (b) linear mit einer Rate von –4 Oktaven pro Sekunde (oct/s) ändert und in (c) mit 10 oct/s. Rechts daneben sind die entsprechenden Periodizitätsfunktionen dargestellt. Während die PF in (a) erwartungsgemäß etwa gleichtiefe Kerben bei Vielfachen von T_0 aufweist, weichen die Funktionen in (b) und (c) deutlich davon ab, was eine sichere Identifikation der Grundperiode erschwert. Wir schlagen deshalb vor,

nicht wie üblich davon auszugehen, dass die F_0 innerhalb eines Frames konstant ist, sondern vielmehr eine lineare Funktion der Zeit.

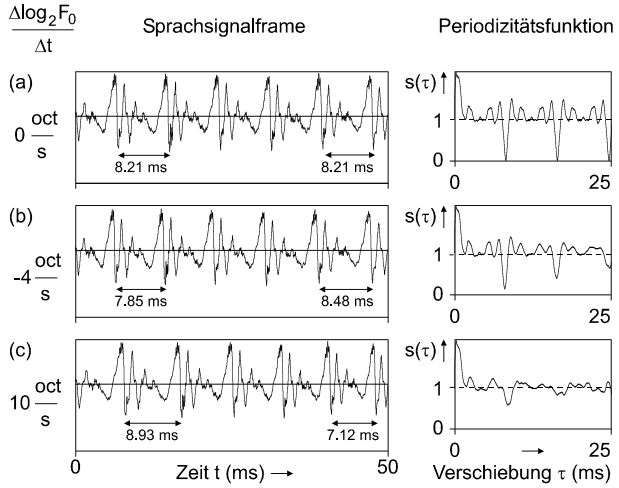

Abbildung 2: Auswirkungen linearer Grundfrequenzänderungen auf die Periodizitätsfunktion.

Für einen positiven Anstieg der Grundfrequenz bedeutet das z.B., dass die Periodenlängen am Ende des Frames kürzer sind als am Anfang (siehe Abbildung 2 (c)). Wenn uns der Anstieg der Funktion $F_0(t)$ nun bekannt wäre, dann ließe sich die Zeitachse des Frames derart verzerren, dass wieder alle Perioden innerhalb des Frames die gleiche Länge besitzen. Die PF des resultierenden periodischen Frames würde sich dann wieder optimal für die Bestimmung von T_0 eignen (siehe Abbildung 2 (a)).

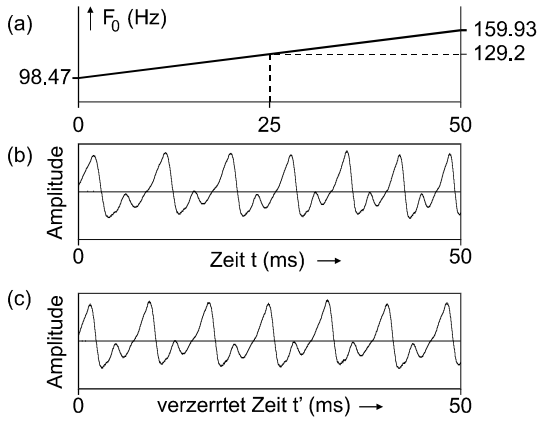

Abbildung 3: Nichtlineare Zeitverzerrung am Beispiel.

Als Beispiel ist in Abbildung 3 (b) ein Frame dargestellt, dessen F_0 linear von 98.47 Hz auf 159.93 Hz ansteigt. Darunter ist der Frame über der verzerrten Zeitachse dargestellt, bei der F_0 mit 129.2 Hz konstant ist. Der verzerrte Frame lässt sich nur dann berechnen, wenn die Änderung der F_0 im Originalframe bekannt ist. Lassen Sie uns also für einen Moment annehmen, dass die F_0-Änderung R in oct/s gegeben ist, d.h.

$$R = \log_2\left(\frac{F_0(L)}{F_0(0)}\right) / L\,, \tag{3}$$

wobei L (=50 ms) die Länge des Frames ist und $F_0(0)$ und $F_0(L)$ die Grundfrequenzen am Anfang und am Ende des Frames sind. Das Frequenzverhältnis $F_0(L)/F_0(0)$ lässt sich dann auch als $r=F_0(L)/F_0(0)=2^{R\cdot L}$ schreiben. Unter der Voraussetzung, dass die Phasenwinkel der Grundschwingung im Originalsignal und im verzerrten Signal zu jedem Zeitpunkt übereinstimmen, lässt sich die folgende Beziehung zwischen t (originale Zeitachse) und t' (verzerrte Zeitachse) ableiten:

$$t = \frac{L}{1-r} + \frac{L}{r-1}\sqrt{1+\frac{t'}{L}(r^2-1)}\,. \tag{4}$$

Die Abtastwerte des verzerrten Signals zu den Zeitpunkten $t'=nT_s$ (n ist der Abtastindex und T_s die Abtastperiode) lassen sich also an den entsprechenden Zeitpunkten t des Originalsignals ablesen[3]. Leider ist uns der Grundfrequenzanstieg R (und damit auch r) aber in der Regel nicht bekannt, so dass wir ihn nur abschätzen können. Wenn wir richtig geschätzt haben, dann ist der verzerrtet Frame wieder weitestgehend periodisch und das absolute Minimum der entsprechenden PF hat einen sehr niedrigen Wert. Wenn unsere Schätzung falsch war, dann ist der Frame auch nach der Verzerrung nicht periodisch und das absolute Minimum der PF hat einen höheren Wert. Um trotz rapider Frequenzänderungen einen robusten T_0–Schätzwert zu erhalten, kann man also für verschiedene zu erwartenden Werte für R die Zeitachse verzerren und mit dem jeweils verzerrten Frame eine PF berechnen. Von diesen Periodizitätsfunktionen wählt man diejenige aus, deren absolutes Minimum am niedrigsten von allen ist und benutzt sie für die Schätzung von T_0, wie es im letzten Abschnitt beschrieben wurde. Für Sprachsignale kommen z.B. folgende Werte für R in Frage: -10, -9, -8, ...,-1, 0, +1, ..., +10 oct/s.

Evaluation und Schlußfolgerungen

Das beschriebene Verfahren zur GFB wurde sowohl auf „normalen“ als auch emotional gesprochenen Äußerungen[4] getestet. Bei emotionaler Sprechweise treten häufig starke Schwankungen in der F_0 und der Intensität auf, die wir als typische Fehlerquellen herkömmlicher Grundfrequenzbestimmer identifiziert haben. Dies trifft auch für „normale“ Sprechweise zu, auch wenn dort z.B. starke F_0–Schwankungen weniger häufig sind. Das vorgestellte Verfahren zeigt nach ersten Messungen durchgängig positive Ergebnisse bei diesen Problemfällen. Auf dem beschriebenen Verfahren aufbauend werden wir als nächstes eine stimmhaft/stimmlos-Unterscheidung in den GFB integrieren und die Robustheit gegenüber Hintergrundstörungen untersuchen.

Danksagung

Wir danken Herr Prof. Sendlmeier von der TU-Berlin für die Bereitstellung der Emotionssprachdatenbank als Testmaterial für den beschriebenen GFB.

[1] P. Vary, U. Heute, W. Hess: „Digitale Sprachsignalverarbeitung“, S. 208, Teubner Verlag, Stuttgart, 1998

[2] A. Cheveigné, H. Kawahara : „YIN, a fundamental frequency estimator for speech and music”, J. Acoust. Soc. Am. 111 (4), S. 1917-1930, 2002

[3] Ein ähnliches Verfahren für die Verzerrung der Zeitachse wurde im Kontext einer neuen spektralen Repräsentationsform vorgeschlagen: T. Abe, T. Kobayashi, S. Imai: „The IF spectrogram: A new spectral representation“, ASVA-97, Tokyo, Japan, 1997

[4] Sendlmeier, W. (2001): „Phonetische Variation als Funktion unterschiedlicher Sprechstile.“ In: W. Hess & K. Stöber (Hrsg.), Elektronische Sprachsignalverarbeitung. 12. Konferenz, Tagungsband. w.e.b. Universitätsverlag, Dresden. S. 23 - 35