Grundfrequenzbestimmung aus dem Modulationsspektrum

Olaf Schreiner† und Holger Quast‡
Drittes Physikalisches Institut, Universität Göttingen (beide) und
†DaimlerChrysler Forschung und Technologie / ‡Robert Bosch Forschung und Vorausentwicklung
{oschrein, holcus}@dpi.physik.uni-goettingen.de

Einleitung

In vielen Anwendungen, besonders solchen aus dem Bereich der Mensch-Maschine-Kommunikation, gewinnt die Bestimmung der Sprachgrundfrequenz immer mehr an Bedeutung, da die Grundfrequenz neben anderen Sprachbestandteilen auch Informationsträger ist. Es wird hier ein neues Verfahren zur Bestimmung der Grundfrequenz aus dem Modulationsspektrum vorgestellt. Das Verfahren orientiert sich eng an der neuronalen Verarbeitung auf dem auditorischen Cortex des menschlichen Gehörs. Die Leistungsfähigkeit des Verfahrens wurde unter verschiedenen Störgeräuschen, u.a. solchen aus dem Automobilbereich, getestet. Schließlich wurde zur Fehlerkorrektur noch eine Viterbi-Suche auf den Ergebnissen implementiert.

Modulation von Sprache

Wesentliche Anteile der menschlichen Sprache, nämlich alle Vokale und alle stimmhaften Konsonanten, weisen eine besondere zeitliche Struktur auf: Sie sind mit der Grundfrequenz F_0 moduliert.

Durch das harte Schließen der Glottisschwingung bei stimmhaften Lauten entstehen Obertöne bei Vielfachen der Grundfrequenz F_0, die mit etwa 12 dB/Oktave abfallen. Die Überlagerung der Grundfrequenz F_0 und ihrer Vielfachen kann auch als Schwebung mit der Differenzfrequenz F_0 interpretiert werden. Man erhält durch die Anwesenheit von Harmonischen also automatisch eine Amplituden-Modulation des Signals mit F_0. Mit der Amplitude $s(t)$ ist natürlich auch die momentane Signalleistung $s(t)^2$ amplitudenmoduliert. Da der Vokaltrakt als Filter betrachtet eine geringe Güte hat, d. h. seine Impulsantwort gut im Zeitbereich lokalisiert ist, bleibt die zeitliche Struktur der Glottisschwingung, die die Amplitudenmodulation beinhaltet, trotz der Filterung durch den Vokaltrakt bei allen Vokalen erhalten.

Gleichzeitig weist das menschliche Gehör einen Mechanismus auf, der gerade diese zeitliche Struktur unabhängig vom Spektrum des Lautes abbildet, nämlich auf die verschiedenen Modulationsfrequenzen. Die Wahrnehmung der Tonhöhe komplexer Töne wie Sprachsignale beruht nicht, wie zunächst angenommen, auf der Frequenzzerlegung in der Cochlea, sondern auf einer nachgeschalteten zeitlichen Analyse im Colliculus Inferior im Hirnstamm. Dort wird für jeden Frequenzkanal der Cochlea eine Art Autokorrelations-Analyse erstellt, indem das Signal mit verzögerten Versionen seiner selbst verglichen wird. Dadurch erhält man neben der Frequenzzerlegung in der Cochlea eine zweite Dimension, das Modulationsspektrum [2].

Die Stimme eines Sprechers wird im Modulationsspektrum besonders stark auf die Grundfrequenz F_0 und deren Harmonische abgebildet. Ein anderes Geräusch, das mit einer anderen Frequenz moduliert ist, wird an eine andere Stelle im Modulationsspektrum abgebildet. Insbesondere wenn ein Geräusch überhaupt keine Regelmäßigkeit in der zeitlichen Einhüllenden seiner Energie hat, sollte es auf die Modulationsfrequenz *Null* abgebildet werden. Es liegt also nahe anzunehmen, dass die Wahrnehmung der Grundfrequenz für das menschlichen Gehör eine besondere Bedeutung hat – etwa um der Stimme eines Sprechers zu folgen oder um einen Sprecher trotz Anwesenheit von Störgeräuschen verstehen zu können.

Dieser Ansatz soll hier technisch nachempfunden werden, um ihn zur Ermittelung der Grundfrequenz zu nutzen. Wenn man also eine Transformation findet, die das Signal ähnlich wie im Gehör abbildet, sollten die modulierten Anteile des Sprachsignals auf die Grundfrequenz abgebildet werden und dort ein deutliches Maximum ausprägen.

Modulationsspektrogramm

Eine Möglichkeit, eine solche Abbildung zu realisieren, ist das Fourierspektrogramm [7]. Die lineare Frequenzaufteilung der Fouriertransformation entspricht jedoch nicht der Zerlegung in der Cochlea, die eher logarithmisch verläuft. Ein weiterer Nachteil der Fourier-Filterbank ist, dass im Bereich der Grundfrequenz das Fenster kürzer ist als die Wellenlänge. Die Grundfrequenz kann daher nicht mehr korrekt aufgelöst werden.

In [6] stellten wir daher als Alternative die Wavelet-Transformation vor, die darüber hinaus den Vorteil hat, dass mit einem Transformationsschritt gleich ein komplettes Spektrogramm, also eine zweidimensionale Zeit-Frequenz-Darstellung erzeugt wird. Verwendet wird hierzu die Schnelle Kontinuierliche Wavelet-Transformation FCWT nach Dress [1] mit dem Gammaton-Wavelet, das die Impulsantwort der Basilarmembran annähert:

$$g_t(t) = a \cdot t^{n-1} \cdot \exp(-2\pi b \mathrm{ERB}(f_c)\, t) \cdot \cos(2\pi f_c t + \phi)$$

Beide Filterbänke sind geeignet, um die Grundfrequenz zu ermitteln [6]. Die einzelnen Kanäle werden dazu in Betrag und Phase getrennt. Die Absolutwerte der Kanäle werden noch einmal in Zeitrichtung in das Modulationsspektrum Fourier-transformiert. Nach einer weiteren Betragsbildung hat man die Darstellung des Sprachsignals auf dem auditorichen Cortex angenähert. In jedem Frequenzkanal findet man nun ein Modulationsspektrum, in dem sich der wesentliche Anteil der modulierten Sprachlaute um die Grundfrequenz konzentriert. Durch Aufaddieren aller Frequenzkanäle erhält man somit das Gesamt-Modulationsspektrum mit einem robusten Maximum bei der Grundfrequenz.

Modulationsspektrum

Das Modulationsspektrogramm ist jedoch sehr rechenaufwändig im Vergleich zu anderen Methoden zur Grundfrequenzbestimmung. Einen deutlichen Geschwindigkeitsvorteil erhält man, wenn man auf die Zerlegung in Frequenzkanäle verzichtet und die Berechnung des Modulationsspektrums direkt auf dem Zeitsignal ausführt.

Das Betragssignal eines komplexen Frequenzkanals wird dazu durch die Hilberteinhüllende $h(t)$ angenähert:

$$h(t) \;=\; \left| s(t) \;+\; \frac{i}{\pi} \int\limits_{\tau \neq 0} s(\tau)/(t-\tau)\,\mathrm{d}\tau \right| \;=\; |\sigma(t)|$$

In der Praxis wird das analytische Signal $\sigma(t)$ durch Berechnung der FFT (Schnelle Fouriertransformation), Nullsetzen der negativen Frequenzkomponenten und Rücktransformation durch die inverse FFT berechnet. Anschließend wird der Betrag des komplexen analytischen Signals gebildet.

Das Modulationsspektrum ergibt sich schließlich als das Leistungsspektrum der Hilberteinhüllenden und wird durch eine weitere FFT und eine abschließende Betragsbildung berechnet.

Viterbi-Suche

Selbst in unverrauschter Sprache muß die Grundfrequenz nicht immer dem höchten Peak im Modulationsspektrum entsprechen. Je nach Sprecher können auch die Vielfachen der Grundfrequenz mehr oder weniger stark ausgeprägt sein und gelegentlich einen höheren Absolutwert erreichen als der Grundfrequenzpeak. Durch additive Störgeräusche kann die Unsicherheit noch verstärkt werden.

Extrahiert man aus dem eigentlichen Pitch-Tracker statt nur einer die n besten Schätzungen für die Grundfrequenz, so erhöht sich die Wahrscheinlichkeit, darunter den wahren Wert für F_0 zu finden. Die Grundfrequenz eines realen Sprechers kann sich in geringen Zeitabsänden nicht beliebig weit verändern. Berücksichtigt man die Grundfrequenzwerte der vorangegangenen Zeitframes, ist es daher möglich auf rein statistischer Basis zu entscheiden, welcher Kandidat der wahrscheinlichste ist.

Ein solches Verfahren ist die Viterbi-Suche [4]. Die Übergangswahrscheinlichkeiten zu einem Pitch-Kandidaten im nächsten Zeitframe können anhand der mittleren vorangegangen Grundfrequenzwerte und der physiologisch maximalen Grundfrequenzänderung berechnet werden. Man erhält eine zweidimensionale Matrix von Übergangswahrscheinlichkeiten der n Kandidaten eines Zeitframes zu den n Kandidaten des nächsten. Bei m Zeitframes wäre es sehr aufwändig, alle n^m Pfade zu berechnen. Der Viterbi-Algorithmus reduziert die Kosten jedoch auf $O(n \cdot m)$ [5].

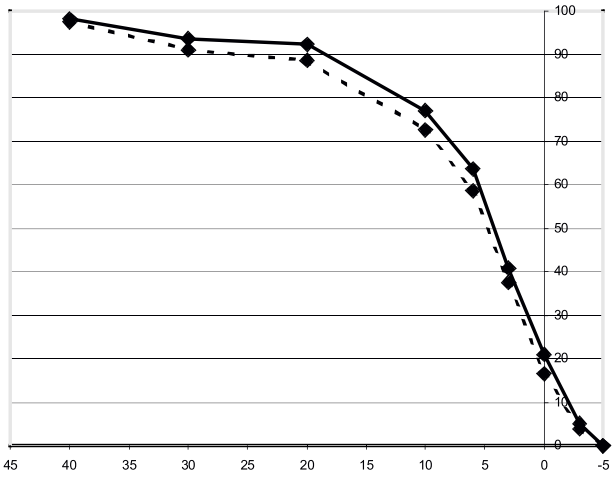

Abb. 1: Cepstrum. Pitcherkennungsraten über SNR. - - - ohne Viterbi, — mit Viterbi

Experiment und Ergebnis

Als Testmaterial wurde zu Sprachdaten der *Göttinger nonverbalen Sprachdatenbank* [3] unkorreliertes Tiefpassrauschen (KFZ-Innengeräusch) bei verschiedenen Signal-Rausch-Verhältnissen (Rauschleistung gegen mittlere Vokalleistung) addiert.

Die Vergleichswerte der Grundfrequenz wurden manuell bestimmt. Verglichen wurde der vorgestellte Algorithmus mit dem in der Spracherkennung verbreiteten Cepstrum.

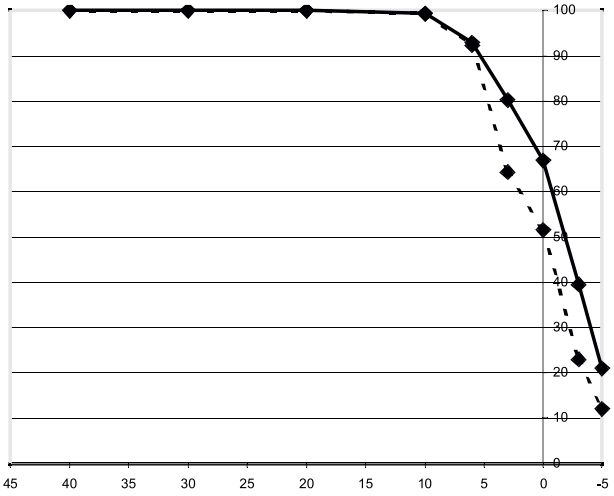

Abb. 2: Modulationsspektrum. Pitcherkennungsraten über SNR. - - - ohne Viterbi, — mit Viterbi

Die Modulationsspektrum-Analyse (MSA) ist deutlich robuster gegen Störgeräusche als die Cepstrum-Methode. Die MSA liefert bis zu einem SNR von 10 dB eine Grundfrequenzerkennungsrate von über 98%, während die Cepstrum-Methode dort bereits auf unter 80% abfällt. Bei 0 dB Störgeräusch sinkt die Erkennungsrate der Cepstrum-Methode auf unter ein Drittel der MSA-Erkennungsrate.

Literatur

[1] Dress, D. B.: *Applications of a Fast Continuous Wavelet Transform.* SPIE Proc. Wavelet Applications IV, 3078:570–580, 1997.

[2] Langner, G., C. E. Schreiner und U. W. Biebel: *Functional Implications of Frequency and Periodicity Coding in the Auditory Midbrain.* In: Palmer, A. R. et al. (Herausgeber): *Psychophysical and physiological Advances in Hearing*, Seiten 277–285. Whurr Publ. Ltd., London, 1998.

[3] Quast, H.: *Recording of Nonverbal Speech Features.* Joint Composium on Neural Networks, Caltech. INC, San Diego, 1999.

[4] Quast, H., O. Schreiner und M.R. Schroeder: *Robust Pitch Tracking in the Car Environment.* Int. Conf. on Acoustics Speech and Signal Processing, 2002.

[5] Rabiner, L.R.: *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.* In: A. Waibel, K.F Lee (Herausgeber): *Readings in Speech Recognition.* Morgan Kaufmann, San Mateo, 1990.

[6] Schreiner, O. und H.W. Strube: *Modulationsfilterung mit Fourier-Spektrogramm und Wavelet-Transformation.* Fortschritte der Akustik – DAGA, 2001.

[7] Wilmers, H.: *Hervorhebung von Signalen durch Operationen im Modulationsspektrum.* Fortschritte der Akustik – DAGA, 1998.