Bestimmung von Rohrmodellparametern aus Sprachsignalen

K. Schnell, A. Lacroix

Institut für Angewandte Physik, Johann Wolfgang Goethe-Universität
Robert-Mayer-Straße 2-4, D-60325 Frankfurt am Main, email: Lacroix@iap.uni-frankfurt.de

Einleitung

Sprachsignale lassen sich mit dem zeitdiskreten Rohrmodell mit entsprechender Anregung synthetisieren. Probleme bestehen dabei einerseits in der Wahl eines geeigneten Modells für den Sprechtrakt und andererseits in der Schätzung der Modellparameter aus dem natürlichen Sprachsignal. Nur für den einfachsten Fall eines unverzweigten Rohrmodells mit nur einem reellen Abschluß an den Lippen und der Anregung am Rohranfang lassen sich die Koeffizienten beispielsweise mit der BURG-Methode [1] schätzen. Die Parameter für erweiterte Rohrmodelle werden aus dem Sprachsignal $x(n)$ geschätzt, indem mit Hilfe eines Optimierungsverfahrens ein Fehler minimiert wird, welcher ein spektrales Abstandsmaß zwischen dem Rohrmodell und dem Sprachsignal darstellt. Da für den Erfolg des Verfahrens die Definition des Fehlers sehr entscheidend ist, wird neben einer schon existierenden Fehlerdefinition [2] eine Alternative vorgestellt.

Zeitinvariantes Rohrmodell und Fehlerdefinition

Das zeitdiskrete Rohrmodell, realisiert durch Kreuzglied-Kettenfilter, beschreibt die Ausbreitung von ebenen Schallwellen. Die Unstetigkeiten der Querschnittsflächen, welche sich an äquidistanten Rohrstellen befinden, werden durch Zweitoradaptoren beschrieben, während die uniformen Rohrelememte durch Laufzeitglieder für verlustlose Wellenausbreitung realisiert werden. Verzweigungen werden durch 3-Tor Adaptoren ermöglicht. Der Rohrabschluß an den Lippen wird durch die frequenzabhängige Lippenimpedanz von Laine [3] modelliert. Ein aus diesen Elementen bestehendes Rohrmodell besitzt Pole und Nullstellen und hat die allgemeine Übertragungsfunktion:

$$H(z)=\frac{b_0-\sum_{k=1}^{M} b_k z^{-k}}{a_0-\sum_{k=1}^{N} a_k z^{-k}}=\frac{b_0}{a_0}\cdot\frac{\prod_{k=1}^{M}\left(z-z_{0k}\right)}{\prod_{k=1}^{N}\left(z-z_{\infty k}\right)}. \qquad (1)$$

Die Polynomkoeffizienten a_k und b_k sind komplizierte Funktionen der Rohrparameter. Die Parameterschätzung soll durch Minimierung eines geeigneten Fehlers durchgeführt werden. Hierfür bietet sich das Konzept der Inversen Filterung an. Diese kann abgeleitet werden aus der Prädiktion eines Signalwertes aus den vorangegangenen Signalwerten $x(n)$ und Prädiktionsfehlerwerten. Die Fehlerleistung ε eines solchen Prädiktors kann durch das Parsevalschen Theorem im Frequenzbereich dargestellt werden:

$$\varepsilon=\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|E(\mathrm{e}^{\mathrm{j}\omega})\right|^2\mathrm{d}\omega=\frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{1-\sum_{k=1}^{N}a'_k\mathrm{e}^{-\mathrm{j}\omega k}}{1-\sum_{k=1}^{M}b'_k\mathrm{e}^{-\mathrm{j}\omega k}}\cdot X(\mathrm{e}^{\mathrm{j}\omega})\right|^2\mathrm{d}\omega \quad . (2)$$

Zu beachten ist bei dieser Fehlerdefinition, daß die absoluten Glieder im Zähler und Nennerpolynom den Wert eins haben. In $H(z)$ ist die Konstante b_0/a_0 allerdings von Rohrparametern abhängig und somit variabel. Um dies zu berücksichtigen existiert in der Fehlerdefinition für das Rohrmodell ein zusätzlicher Faktor $\left|b_0/a_0\right|^2$.

$$\varepsilon_k=\left|\frac{b_0}{a_0}\right|^2\cdot\sum_k\left|\frac{X(\mathrm{e}^{\mathrm{j}\omega_0 k})}{H(\mathrm{e}^{\mathrm{j}\omega_0 k})}\right|^2 \qquad (3)$$

Bei Nichtbeachten dieses Korrekturfaktors erhält man ungenügende Ergebnisse [4]. ω_0 stellt die Grundfrequenz dar. Der Fehler (3) wird mit Hilfe eines Gradientenverfahrens minimiert. Der Gradient kann durch Differenzen von kleinen Variationen der Rohrparameter approximiert werden.

Zeitvariables Rohrmodell und Fehlerdefinition

Der Einfluß der schwingenden Stimmbänder kann durch einen zeitvariablen Glottiskoeffizienten im Rohrsystem modelliert werden [2]. Dieser Koeffizient beschreibt den Querrschnittssprung zwischen der Glottis und der ersten Querrschnittsfläche im Vokaltrakt. Der Koeffizient ist abhängig von der Glottisfunktion, welche durch ein Modell von Oliveira [5] modelliert wird. Durch diesen zeitvariablen Glottiskoeffizienten ist das Rohrmodell nicht mehr zeitinvariant und besitzt folglich keine Übertragungsfunktion im z-Bereich. Dieser Umstand hat Auswirkungen auf die Fehlerdefinition, da für die Berechnung von b_0/a_0 die Übertragunsfunktion $H(z)$ nicht mehr zur Verfügung steht. Durch das Theorem

$$\int_{-\pi}^{\pi}\ln\left(\left|L(\mathrm{e}^{\mathrm{j}\omega})\right|^2\right)\mathrm{d}\omega=2\pi\cdot\ln(l_0{}^2), \qquad (4)$$

welches für ein minimalphasiges System $L(z)$ gültig ist [6], besteht die Möglichkeit den benötigten Faktor aus dem Filterausgangssignal $y(n)$ zu schätzen. Dadurch erhält man die Fehlerdefinition

$$\varepsilon_1=\exp\left(\int_{-\pi}^{\pi}\ln\left(\left|Y(\mathrm{e}^{\mathrm{j}\omega})\right|^2\right)\mathrm{d}\omega\right)\cdot\int_{-\pi}^{\pi}\left|\frac{X(\mathrm{e}^{\mathrm{j}\omega})}{Y(\mathrm{e}^{\mathrm{j}\omega})}\right|^2\mathrm{d}\omega, \qquad (5)$$

bzw.

$$\varepsilon_1=\prod_k\left|Y(\mathrm{e}^{\mathrm{j}\omega_0 k})\right|^2\cdot\sum_k\left|\frac{X(\mathrm{e}^{\mathrm{j}\omega_0 k})}{Y(\mathrm{e}^{\mathrm{j}\omega_0 k})}\right|^2 \qquad (6)$$

für endliche Signalwerte. Dieser Fehler ε_1 wird für stimmhafte Laute der Grundfrequenz ω_0 auf eine Periode des Filterausgangs $y(n)$ angewendet, nachdem das Ausgangssignal sich nach etwa 10 Perioden eingeschwungen hat. Auffällig ist bei der Fehlerdefinition (6), daß der Vorfaktor durch ein Produkt dargestellt werden kann. Motiviert durch diese Tatsache wurde nach einer Fehlerdefinition gesucht, die ebenso durch ein Produkt darstellbar ist. In (6) ist eine Summe von Produkten enthalten. Deshalb wird in einer neuen Fehlerdefinition ein Produkt von Potenzen angenommen, da zwischen der Multiplikation und der Addition eine ähnlich formale Beziehung besteht wie zwischen dem Potenzieren und Multiplizieren. Evaluationen von einigen möglichen Fehlerdefinitionen führte auf:

$$\varepsilon_2=\prod_k\left|\lambda\frac{X(\mathrm{e}^{\mathrm{j}\omega_0 k})}{Y(\mathrm{e}^{\mathrm{j}\omega_0 k})}\right|^{\left|\lambda\frac{X(\mathrm{e}^{\mathrm{j}\omega_0 k})}{Y(\mathrm{e}^{\mathrm{j}\omega_0 k})}\right|} \quad \text{mit} \quad \lambda=\prod_k\left|\frac{Y(\mathrm{e}^{\mathrm{j}\omega_0 k})}{X(\mathrm{e}^{\mathrm{j}\omega_0 k})}\right|. \qquad (7)$$

Die zu (7) äquivalente Darstellung:

$$\varepsilon_2=\exp\left(\sum_k\left|\lambda\frac{X(\mathrm{e}^{\mathrm{j}\omega_0 k})}{Y(\mathrm{e}^{\mathrm{j}\omega_0 k})}\right|\cdot\ln\left(\left|\lambda\frac{X(\mathrm{e}^{\mathrm{j}\omega_0 k})}{Y(\mathrm{e}^{\mathrm{j}\omega_0 k})}\right|\right)\right), \qquad (8)$$

läßt gewisse Ähnlichkeit zu ε_1 erkennen. Anzumerken ist, daß diese Fehlerdefinition ohne einen Vorfaktor auskommt. In Analysen zeigt sich daß die Fehlerdefinition (7) ähnlich gute Resultate liefert wie ε_1. Falls bei einer Analyse die Fehlerdefinition ε_1 nicht zum Ziel führt, kann die andere günstigere Resultate liefern. Es sollen aber noch Analysen von umfangreicheren Sprachproben erfolgen, um diese Aussage zu erhärten.

Analyse von Vokalen

Für die Analyse von Vokalen ist das verwendete Rohrmodell unverzweigt. Der Rohrabschluß an den Lippen wird mittels der Impedanz von Laine modelliert. Diese Impedanz ist abhängig von der Mundöffnungsfläche, welche vor der Minimierung für den entsprechenden Laut fest eingestellt wird. Der Rohrabschluß an der

Glottis wird durch einen zeitvariablen Glottiskoeffizienten modelliert. Um die Analyse und Synthese von Sprachlauten zu verbessern, wird das Rohrmodell nicht mit der Glottisfunktion von Oliveira angeregt, sondern durch eine Impulsfolge, die zuvor mit reellen Polstellen gefiltert wurde. Die reellen Pole werden durch eine wiederholte adaptive Präemphase aus dem zu analysierenden Sprachsignal geschätzt. Dadurch kann der spektrale Abfall des Sprachsignals, der von der Glottisanregung und der Lippenabstrahlung bedingt ist, besser modelliert werden als durch Verwendung der Glottisfunktion von Oliveira für das Anregungssignal. Die Rohrabschlüsse an Glottis und Lippen erhalten einen zusätzlichen reellen Faktor mit Betrag kleiner eins, um Verluste zu berücksichtigen. Die Rohrparameter werden geschätzt, indem der Fehler ε_2 durch ein Gradientenverfahren mit adaptiver Schrittweite minimiert wird. Da der Wert des Fehlers (7) besonders am Anfang der Optimierung sehr groß sein kann und den Bereich der Zahlendarstellung des Computers überschreiten kann, wird zu Anfang der Minimierung die Fehlerdefinition (8) verwendet, wobei die Exponentialfunktion weggelassen werden kann. Im folgenden werden Resultate gezeigt aus der Analyse des Vokals ‘o’, welcher offen gesprochen und mit einer Abtastrate von 22 kHz aufgenommen wurde. Bild 1 zeigt Vokaltraktflächen die aus einer Periode des Sprachsignals geschätzt wurden. Bild 2 zeigt Vokaltraktflächen die durch NMR Aufnahmen von Story [7] entstanden sind.

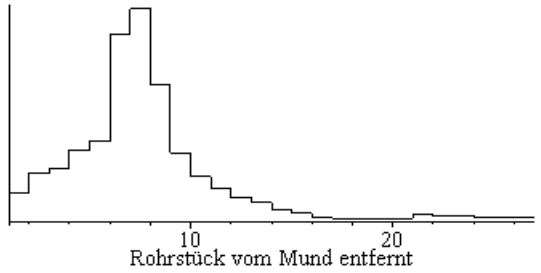

Bild 1: geschätzte Vokaltraktflächen (offenes o)

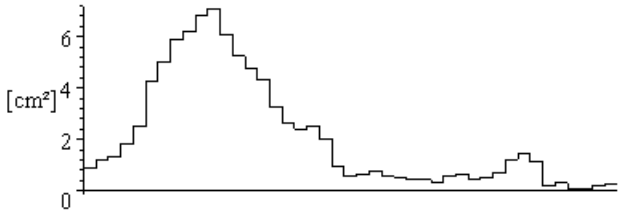

Bild 2: Vokaltraktflächen aus NMR [7] (offenes o)

Die geschätzten Flächen sind mit denen der NMR Aufnahmen von der Grobstruktur her ähnlich. Es muß dabei beachtet werden, daß die Flächen von unterschiedlichen Sprechern stammen. Dieses Resultat erhält man auch für andere Vokale.

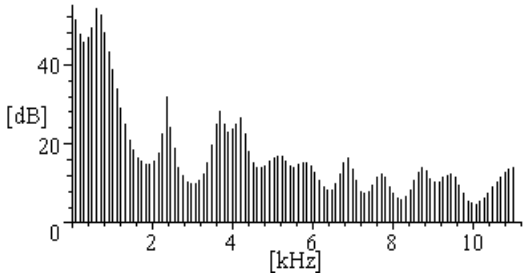

Bild 3: Betragsspektrum des Filterausgangs

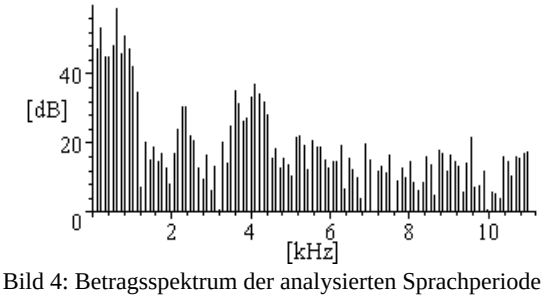

Bild 4: Betragsspektrum der analysierten Sprachperiode

Die Fehlerminimierung gleicht das Betragsspektrum des Filterausgangssignals dem Sprachspektrum an, wie in den Bildern 3 und 4 zu sehen ist. Bei der Analyse von Vokalen liefert die Fehlerdefinition ε_2 vergleichbar gute Ergebnisse zu Schätzungen, ermittelt mit der Fehlerdefinition ε_1. Daß die Fehlerdefinitionen auch unterschiedliche Ergebnisse liefern können, zeigt das folgende Beispiel einer Analyse eines vorgeschriebenen Systems. Das Ausgangssignal eines verzweigten Rohrsystems wurde analysiert, welches durch eine Impulsfolge angeregt wurde. Dieser Fall zeigt, daß die Fehlerdefinition ε_2 bessere Ergebnisse liefern kann, während die Fehlerdefinition ε_1 zu keinem befriedigenden Resultat geführt hat. Da für die Analyse das Rohrmodell die gleiche Ordnung hat wie das, das für die Erzeugung des Testsignals verwendet wurde, ist das System in der Lage, das zu analysierende Signal vollständig zu beschreiben. Wie in den Bildern 5 und 6 zu sehen ist, ist das mit Fehlerdefinition ε_2 gelungen, wohingegen mit Fehlerdefinition ε_1 der Optimierungsalgorithmus in einem lokalem Minimum verharrt.

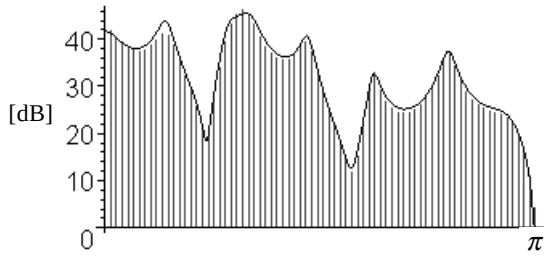

Bild 5: Minimierung von ε_2 mit 200 Iterationen;
Linienspektrum: analysierte Periode;
durchgezogene Linie: geschätzter Betragsgang $|H(z)|$

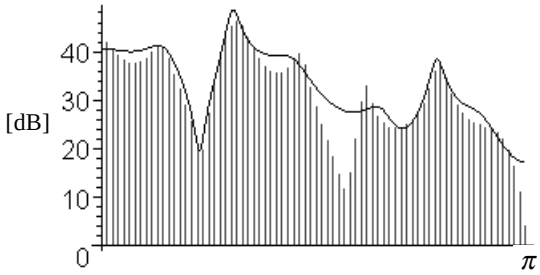

Bild 6: Wie in Bild 5, jedoch Minimierung von ε_1 (2000 Iterationen)

Zusammenfassung

Die Parameterbestimmung für zeitvariante und zeitinvariante Rohrmodelle wird mittels der Minimierung eines Fehlers vollzogen, der ein spektrales Abstandsmaß zwischen dem Rohrmodell und dem Sprachsignal darstellt. Durch eine neue Fehlerdefinition steht zu dem schon existierendem Fehler, der aus der inversen Filterung abgeleitet ist, eine Alternative zur Verfügung. An Beispielen wird die Überlegenheit der neuen Fehlerdefinition demonstriert.

Literatur

[1] Burg, J.: A New Analysis Technique for Time Series Data, *NATO Advanced Study Institute on Signal Processing*, Enschede 1968.

[2] Schnell, K.; Lacroix, A.: Parameter Estimation for Tube Models with Time Dependent Glottis Impedance, *Proc. of the second EURASIP Conference (ECMCS’99)*, Krakow (1999), CD-ROM.

[3] Laine, U.K.: Modeling of Lip Radiation Impedance in the z-domain, *Proc. ICASSP-82* pp. 1992-1995.

[4] Schnell, K.; Lacroix, A.: Erweiterte Rohrmodelle für die Sprachproduktion,Tagungsband *DAGA 1998* Zürich, pp. 384-385.

[5] Oliveira, L. C.: Estimation of Source Parameters by Frequency Analysis. *Proc. Eurospeech*, Berlin, Vol 1, 99-102, 1993.

[6] Papoulis, A.*: Probability, Random, Variables, and Stochastic Processes*, McGraw-Hill, New York, 3rd Edition 1991.

[7] Story, B.H. et al.: Vocal Tract Area Functions from Magnetic Resonance Imaging , *JASA* Vol. 100 (1996), pp. 537-554.