Vokaltraktmodellierung unter Verwendung eines Pol-Nullstellen Modells

Wolfgang Kreuzer[1], Peter Balazs[2], Damian Marelli[3]

[1] *Österreichische Akademie der Wissenschaften, Institut für Schallforschung, 1040 Wien, Email: wolfgang.kreuzer@oeaw.ac.at*

[2] *Österreichische Akademie der Wissenschaften, Institut für Schallforschung, 1040 Wien, Email: peter.balazs@oeaw.ac.at*

[3] *Univeristy of Newcastle, Sydney, Australien, Email: Damian.Marelli@newcastle.edu.au*

Einleitung

In der Literatur gibt es viele Modelle zur Schätzung der Querschnittsflächenfunktion des Vokaltrakts basierend auf Sprachsignalen. Dabei wird der Vokaltrakt durch ein Rohrmodell bestehend aus gleich langen Zylindern mit jeweils konstanten Radien modelliert [1, 2]. Mit Hilfe eines Allpol-Modells können die Radien für jeden dieser Zylinder direkt aus dem Sprachsignal bestimmt werden. Um auch Einflüsse des Nasaltrakts berücksichtigen zu können, wurde ein Modell entwickelt, dass einen zusätzlichen Trakt inkludiert [3]. Wir vergleichen dieses akustisch basierte Vokaltrakt Modell mit unserem signal-basierten Pol-Nullstellenmodell [4] und diskutieren Ergebnisse und Unterschiede. Im Folgenden bezeichnet der Begriff 'Vokaltraktmodell' das Rohrmodell aus [3], während sich der Begriff 'Pol-Nullstellenmodell' auf das Modell aus [4] bezieht.

Pol-Nullstellenmodell

Bei diesem Modell wird angenommen, dass ein abgetastetes Sprachsignal $y(t)$ durch ein Anregungssignal $u(t)$ erzeugt wird, das durch einen Sprachproduktionsfilter $g_\tau(t)$ gefiltert wird. Es wird angenommen, dass dieser Filter in der z-Ebene als rationale Funktion dargestellt werden kann, d.h.

$$G(z,\boldsymbol{\theta}) = \frac{B(z,\boldsymbol{\theta})}{A(z,\boldsymbol{\theta})} = \frac{a_0 + a_1 z^{-1} + \cdots + a_n z^{-n}}{1 + b_1 z^{-1} + \cdots + b_m z^{-m}},$$

wobei $\boldsymbol{\theta} = \{a_0, \ldots, a_n, b_1, \ldots, b_m\}$ den Vektor der Polynomkoeffizienten bezeichnet. Im Allgemeinen wird nun versucht, den Parametervektor $\boldsymbol{\theta}$ so zu bestimmen, dass $B(z,\boldsymbol{\theta})/A(z,\boldsymbol{\theta})$ die Einhüllende des Spektrums möglichst gut approximiert, d.h. der Parameter $\boldsymbol{\theta}$ wird durch das Lösen eines nichtlinearen Gleichungssystems bestimmt:

$$\boldsymbol{\theta} = \underset{\boldsymbol{\theta}'}{\operatorname{argmin}} \sum_{k=1}^{K} \left| \hat{G}(\omega_k) - \frac{B(e^{\mathrm{i}\omega_k},\boldsymbol{\theta}')}{A(e^{\mathrm{i}\omega_k},\boldsymbol{\theta}')} \right|^2,$$

wobei $\hat{G}(\omega_k)$ einen Schätzer für die Einhüllende des Spektrums und ω_k diskrete Frequenzpunkte darstellen.

Bei unserem Modell werden zusätzlich ein logarithmisches Fehlermaß angenommen und die Phase des Spektrums vernachlässigt. Beide Annahmen können perzeptiv motiviert werden. Der Parameter $\boldsymbol{\theta}$ wird deshalb durch

$$\theta = \underset{\boldsymbol{\theta}'}{\operatorname{argmin}} \sum_{k=1}^{K} \left| \log|\hat{G}(\omega_k)| - \log\left| \frac{B(e^{\mathrm{i}\omega_k},\boldsymbol{\theta}')}{A(e^{\mathrm{i}\omega_k},\boldsymbol{\theta}')} \right| \right|^2,$$

bestimmt. Diese Vorgehensweise erlaubt uns Pole und Nullstellen ohne vorhergehendes Wissen über Vielfachheit und Art (reell oder komplex) zu bestimmen, die Bestimmung ist lokal optimal und numerisch effizient, insbesonders wenn eine nicht reguläre Frequenzskala verwendet werden soll, und sie basiert auf einem perzeptiv relevanten Fehlerkriterium.

Vokaltraktmodell

Das in [3] beschriebene Vokaltraktmodell besteht aus drei segmentierten Rohrmodellen[1], die Pharynx, Nasaltrakt, und Vokaltrakt beschreiben, und einer 3-Stufen Kreuzung, die die Trakte verbindet. Das Modell ist speziell auf Nasale abgestimmt, d.h. es wird angenommen, dass der Oraltrakt geschlossen und der Nasaltrakt offen sind. Aus diesem Modell ergibt sich eine rationale Transferfunktion $\frac{\hat{B}(z,\boldsymbol{\mu})}{\hat{A}(z,\boldsymbol{\mu})}$ in Abhängigkeit von Reflexionskoeffizienten μ_n, die den Übergang vom n–1-ten Segment zum n Segment beschreiben

$$\mu_n = \frac{S_{n-1} - S_n}{S_{n-1} + S_n},$$

wobei S_n die Querschnittsfläche des n-ten Segments bezeichnet. Die Kopplung der 3 Trakte wird zusätzlich noch durch das Flächenverhältnis σ zwischen geschlossenem und offenem Trakt (Oral- und Nasaltrakt) beeinflußt. Um das Vokaltrakt-Modell besser an reelle Signale anpassen zu können, haben Lim und Lee[3] zusätzlich im geschlossenen Trakt am Ende einen Reflexionskoeffizienten $\tilde{\mu_0}$ eingeführt, der Verluste im Schallfluss im Oraltrakt modellieren soll.

Erste Ergebnisse

Im ersten Schritt sollen aus einem synthetischen Signal mit Hilfe des Pol-Nullstellen Modells die rationelle Filterfunktion $B(z,\boldsymbol{\theta})/A(z,\boldsymbol{\theta})$ bestimmt und mit der Funktion $\hat{B}(z,\boldsymbol{\mu})/\hat{A}(z,\boldsymbol{\mu})$ aus dem Vokaltraktmodell verglichen werden. Das Signal wurde vorher durch das Vokaltraktmodell aus vorgegebenen Reflexionskoeffizienten $\boldsymbol{\mu}$ generiert. Für eine genaue Beschreibung der Reflexionskoeffizienten $\boldsymbol{\mu}$ siehe Tabelle 1, Spalte 1, Lim und Lee[3] ermittelten diese Reflexionskoeffizienten basierend auf Features eines Sprachsignals /m/, das mit 10 kHz abgetastet wurde.

Als Startpolynome für den nichtlinearen Gleichungslöser des Pol-Nullstellenmodells wurden Polynome verwendet,

[1] siehe z.B. [2] für eine Formulierung eines segmentierten Rohrmodells.

die ebenfalls aus Reflexionskoeffizienten x_0 gewonnen wurden (siehe die 2. Spalte in Tabelle 1). x_0 wurde so gewählt, dass die Randbedingungen der einzelnen Trakte realistisch sind, ansonsten wurden gleiche Querschnittsflächen innerhalb der jeweiligen Rohrmodelle angenommen.

Der direkte Vergleich der Koeffizienten der Polynome $B(z, \boldsymbol{\theta})$ und $\hat{B}(z, \boldsymbol{\mu})$, bzw. der Polynome $A(z, \boldsymbol{\theta})$ und $\hat{A}(z, \boldsymbol{\mu})$ lieferte eine Übereinstimmung im Rundungsfehlerbereich. Das deutet darauf hin, dass das signal-basierte Modell Filterfunktionen bestimmen kann, die eine physikalische Bedeutung haben.

Es muss jedoch erwähnt werden, dass die Methode in [4] zwar lokale Optimalität garantiert, sie aber deshalb von den Startwerten des nichtlinearen Gleichungslösers abhängt. Für Startwerte x_0, die zum Beispiel zufällig bestimmt werden, kommt es deshalb zu Problemen mit der Eindeutigkeit und damit zu Unterschieden zwischen $A(z, \boldsymbol{\theta})$ und $\hat{A}(z, \boldsymbol{\mu})$, bzw. $B(z, \boldsymbol{\theta})$ und $\hat{B}(z, \boldsymbol{\mu})$. Bild 1 zeigt den Unterschied zwischen den Pol- und Nullstellen des Vokaltraktmodells (blaue Kreuze, bzw. blaue Kreise) und den Pol- und Nullstellen des Pol-Nullstellenmodells (rote Kreuze und Kreise). Die Übereinstimmung der komplexen Pol- und Nullstellen, die getrennt voneinander auftreten, ist an sich sehr zufriedenstellend. Probleme mit der Eindeutigkeit gibt es jedoch, wenn Pol- und Nullstelle knapp beieinander liegen bzw. ident sind, und sie sich deshalb in der rationalen Funktion wegkürzen. Das ist jedoch ein grundlegendes konzeptionelles Problem. Zusätzlich sehen wir in Bild 1 einen Unterschied in den reellen Pol- und Nullstellen. Es wird jedoch in der Literatur (siehe z.B. [5]) darauf hingewiesen, dass Pole, die durch die Anregung entstehen im Allgemeinen reell sind, bzw. kleine Peaks produzieren. Außerdem spielen reelle Pole und Nullstellen bei charakteristischen Merkmalen der Stimme keine Rolle. Die dritte Spalte in Tabelle 1 zeigt die Reflexionskoeffizienten, die aus diesen Pol-Nullstellen Polynomen durch das Vokaltraktmodell berechnet wurden. Vor allem im Kopplungskoeffizienten σ kommt es zu einer großen Abweichungen, die dann auch die Koeffizienten für den Oraltrakt beeinflussen.

Für eine weitere Entwicklung des Pol-Nullstellen Modells, bzw. des Vokaltraktmodells bedeutet das, daß bei den nichtlinearen Gleichungslösern zusätzliche a-priori Kriterien wie zum Beispiel kleinere Gewichtung von reellen Pol und Nullstellen berücksichtigt werden sollten.

Tabelle 1: Reflexionskoeffizienten aus [3] (Spalte 1) und Startwerte x_0 zur Bestimmung des Pol-Nullstellenmodells (Spalte 2). $\tilde{\mu}$ bezeichnet Reflexionskoeffizienten die dem Oraltrakt zugeordnet werden können.

	Original	x_0	Fall 2
μ_1	-0.97381396	-0.9	-0.9783
μ_2	0.36926744	0.5	0.2305
μ_3	0.43586850	0.0	0.0114
μ_4	0.13602392	0.0	0.0443
μ_5	0.39757567	0.0	0.4516
μ_6	0.64828503	0.5	0.4421
μ_7	-0.10183206	0.0	-0.0212
μ_8	0.43124188	0.0	0.3201
μ_9	-0.35648535	0.0	-0.2975
μ_{10}	-0.0543100	0.0	-0.1124
$\tilde{\mu_0}$	-0.28604275	-0.5	-0.2933
$\tilde{\mu_1}$	-0.091589150	0.0	-0.3293
$\tilde{\mu_2}$	-0.06728477	0.0	0.2640
$\tilde{\mu_3}$	-0.32191045	0.0	-0.3203
$\tilde{\mu_4}$	-0.069770125	0.0	-0.2933
σ	0.12168441	0.5	0.0480

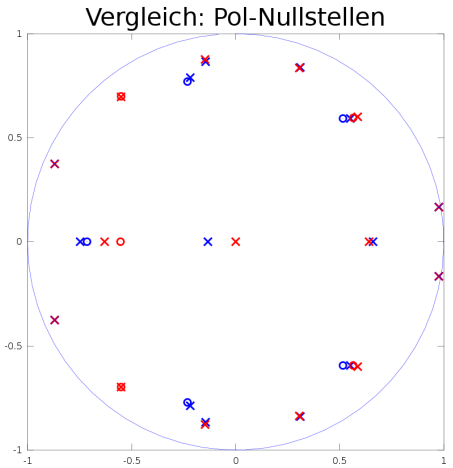

Abbildung 1: Pole ('x') und Nullstelle ('o') des Vokaltraktmodells (blau) und des Pol-Nullstellenmodells (rot) für zufälliges Startwerte x_0 bei der Berechnung des Pol-Nullstellenmodells

Literatur

[1] Fant, G.: Acoustic theory of speech production: With Calculation based on X-Ray Studies of Russian Articulations Mouten de Gruyter, The Hague, 1970

[2] Wakita, H.: Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms, IEEE Transactions on Audio and Electroacoustics 5 (1973), 417–427

[3] Lim, I.-Z., Lee B. G.: Lossy Pole-Zero Modeling for Speech Signals, IEEE Transactions on Speech and Audio Processing 4 (1996), 81–88

[4] Marelli, D., Balazs, P.: On Pole-Zero Model Estimation Methods Minimizing a Logarithmic Criterion for Speech Analysis, IEEE Transactions on Audio, Speech and Language Processing 18/2 (2010), 237–248

[5] Atal, B.S., Hanauer S.L.: Speech Analysis and Synthesis by Linear Prediction of the Speech Wave, The Journal of the Acoustical Society of America 50 (1971), 637–655