NICHTKAUSALES CEPSTRALES SPRACHMODELL

Robert Vích

Institut für Photonik und Elektronik,

Akademie der Wissenschaften der Tschechischen Republik

vich@ufe.cz

Kurzfassung: Konventionelle cepstrale Sprachsynthese basiert auf dem minimalphasigen parametrischen Spracherzeugungsmodell mit unendlicher Impulsantwort. Die Übertragungsfunktion des minimalphasigen cepstralen Vokaltraktmodells wird aus dem gefensterten reellen Cepstrum mit Hilfe der Padé Approximation gewonnen. In diesem Fall approximiert der logarithmische Frequenzgang des Modells nur das logarithmische Betragsspektrum des zugehörigen Sprachsegments. In diesem Beitrag wird für das cepstrale Sprachmodell das komplexe Cepstrum angewendet, das auch die Phaseninformation beinhaltet. Das mischphasige Spracherzeugungsmodell wird in diesem Fall durch die Kaskadenschaltung eines kausalen und eines nichtkausalen Filters mit endlicher Impulsantwort realisiert. Das kausale Filter entspricht dem kausalen Teil, das nichtkausale Filter dem antizipativen Teil des komplexen Cepstrums. Beide Filter können separat aus den zugehörigen gefensterten Cepstrumteilen mit Hilfe der diskreten Fourier Transformation oder rekursiv konstruiert werden. Man kann auch die nichtkausale mischphasige Impulsantwort des Vokaltraktmodells direkt mit Hilfe der diskreten Fourier Transformation zu dem gefensterten komplexen Cepstrum bestimmen. Das nichtkausal synthetisierte Sprachsignal approximiert mit größerer Genauigkeit das originelle Signal im Vergleich zu der konventionellen Cepstralsynthese mit Hilfe des reellen Cepstrums, ist aber rechentechnisch mindestens doppelt so anspruchsvoll.

1 Einführung

Das cepstrale Spracherzeugungsmodell ist ein parametrisches System und kann durch das lineare zeitvariable, in Abb. 1 dargestellte Schema realisiert werden.

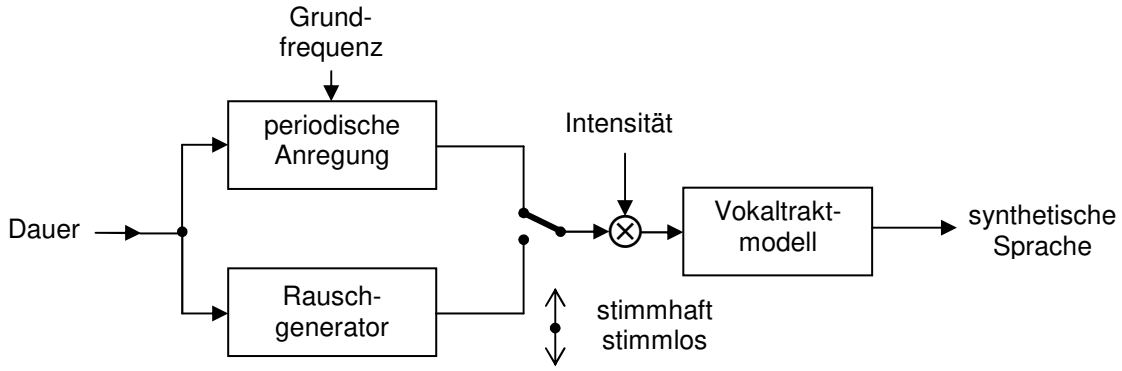

Abbildung 1 – Parametrisches Spracherzeugungsmodell.

Das Vokaltraktmodell in dem Bild wird in diesem Beitrag durch eine Übertragungsfunktion beschrieben, die aus dem Cepstrum konstruiert wird. Es kann in der Form eines Digitalfilters mit endlicher (FIR) als auch unendlicher (IIR) Impulsantwort realisiert werden. Die Intensität der Anregung des Vokaltraktmodells wird aus der Energie des Restsignals nach der inversen

Filterung des Sprachsignals mit dem cepstralen Vokaltraktmodell bestimmt. Die Grundfrequenz und Stimmhaftigkeit werden mit üblichen Methoden der Grundfrequenzbestimmung berechnet. Die Dauer entspricht der Länge der realisierten Laute.

Anwendung der homomorphen Signalanalyse bei der Sprachmodellierung findet man in den Arbeiten von Oppenheim und Quatieri [1,2]. Eine Einführung in die Anwendung des minimalphasigen als auch komplexen Cepstrums in der Sprachsynthese ist in [3] mit weiteren Literaturangaben zu finden.

In diesem Beitrag wird das Prinzip der cepstralen Sprachanalyse- und Synthese nur kurz zusammengefasst und die wichtigsten Begriffe und Formeln angegeben. Aufmerksamkeit wird der Konstruktion der nichtkausalen Impulsantwort des Vokaltraktmodells mit Hilfe der rekursiven inversen Cepstral-Transformation gewidmet. Es werden Überlegungen präsentiert, die zu der Implementierung mischphasiger Sprachsynthese mit größerer Natürlichkeit führen können.

2 Cepstrale Sprachanalyse

Es sei $\{s_n\}$ das gefensterte Sprachsegment eines digitalisierten Sprachsignals der Rahmenlänge *N*. Das zugehörige *komplexe Cepstrum* $\{\hat{s}_n\}$, $-\infty \geq n \geq +\infty$ ist durch folgende Beziehungen gegeben [1,2,4]

$$\hat{s}_n = \frac{1}{2\pi T} \int_{-\pi/T}^{\pi/T} \hat{S}(e^{\mathrm{j}\omega T}) e^{\mathrm{j}\omega T n} d\omega, \tag{1a}$$

$$\hat{S}(e^{\mathrm{j}\omega T}) = \ln S(e^{\mathrm{j}\omega T}) = \ln\left|S(e^{\mathrm{j}\omega T})\right| + \mathrm{j}\arg S(e^{\mathrm{j}\omega T}), \tag{1b}$$

$$S(e^{\mathrm{j}\omega T}) = \sum_{n=0}^{N-1} s_n e^{-\mathrm{j}\omega T n}. \tag{1c}$$

Die Funktion $S(e^{\mathrm{j}\omega T})$ ist das *komplexe Spektrum* der Folge $\{s_n\}$, $T = 1/F_A$ und F_A ist die Abtastrate. Der imaginäre Teil des logarithmischen Spektrums $\hat{S}(e^{\mathrm{j}\omega T})$, d.h. die Phase $\arg S(e^{\mathrm{j}\omega T})$, wird hier in eine *ungerade kontinuierliche Funktion* von ω durch Subtraktion des linearen Trends der Funktion $\arg S(e^{\mathrm{j}\omega T})$ umgewandelt (Phasenentfaltung). Die Beziehungen (1) können als *direkte Cepstral-Transformation* in dem Frequenzbereich bezeichnet werden.

Das komplexe Cepstrum $\{\hat{s}_n\}$ ist eine *zweiseitige,* für mischphasige Signale $\{s_n\}$ eine allgemein *unsymmetrische Folge*. Der rechtsseitige Teil der Folge $\{\hat{s}_n\}$ für $0 \leq n$ ist der *kausale Teil*, der linksseitige Teil für $n < 0$ ist der *antizipative Teil* des komplexen Cepstrums.

Das Cepstrum wird mit Hilfe der *schnellen Fourier-Transformation* (FFT) der Dimension $M \geq N$ berechnet.

$$\hat{s}_{\mathrm{p},n} = \frac{1}{M} \sum_{k=0}^{M-1} \hat{S}_k e^{\mathrm{j}2\pi kn/M}, \tag{2a}$$

$$\hat{S}_k = \ln S_k = \ln\left|S_k\right| + \mathrm{j}\arg S_k, \tag{2b}$$

$$S_k = \sum_{n=0}^{M-1} s_n e^{-\mathrm{j}2\pi kn/M}. \tag{2c}$$

In Folge der Anwendung der diskreten Fourier-Transformation ist $\{\hat{s}_{\mathrm{p},n}\}$ eine *periodische Folge* mit der Periode *M,* die mit $\{\hat{s}_n\}$durch folgende Beziehung

$$\hat{s}_{\mathrm{p},n} = \sum_{i=-\infty}^{+\infty} \hat{s}_{n+iM} \tag{3}$$

verbunden ist. Diese Tatsache wird als *cepstrale Überfaltung* (cepstral aliasing) bezeichnet. Um große Fehler bei der Berechnung des Cepstrums zu vermeiden, muss die Folge $\{s_n\}$ mit Null Werten erweitert werden, so dass $M > qN$ ist, in gewissen Anwendungen bis $q = 5$. Im Weiteren werden wir auch für $\hat{s}_{\mathrm{p},n}$ die einfache Schreibweise $\hat{s}_n$ benutzen.

Aus dem komplexen Cepstrum $\{\hat{s}_n\}$ kann man die Approximation $\{\tilde{s}_n\}$ des Signals $\{s_n\}$ gewinnen. Es werden hier nur die Beziehungen mit der FFT angegeben.

$$\hat{S}_k = \sum_{n=0}^{M-1} \hat{s}_n e^{-\mathrm{j}2\pi kn/M} \ , \tag{4a}$$

$$S_k = \exp(\hat{S}_k) \, , \tag{4b}$$

$$\tilde{s}_n = \frac{1}{M} \sum_{k=0}^{M-1} S_k e^{\mathrm{j}2\pi kn/M} \ . \tag{4c}$$

Diese Beziehungen können wir als *inverse Cepstral-Transformation* bezeichnen.

Als Beispiel für die Cepstralanalyse nehmen wir den stationären Teil des Vokals „a“, gesprochen von einem männlichen Sprecher. Die Abtastrate ist $F_A = 8$ kHz, die Grundfrequenz $F_0 = 118$ Hz, die Rahmenlänge $N = 200$ und die Dimension der angewendeten FFT ist $M = 512$. Die Grundfrequenzperiode entspricht $M_0 = F_A / F_0 = 68$ Abtastwerten. In Abb. 2 sind das Sprachsignal, sein Amplituden- und Phasenspektrum und das komplexe Cepstrum abgebildet.

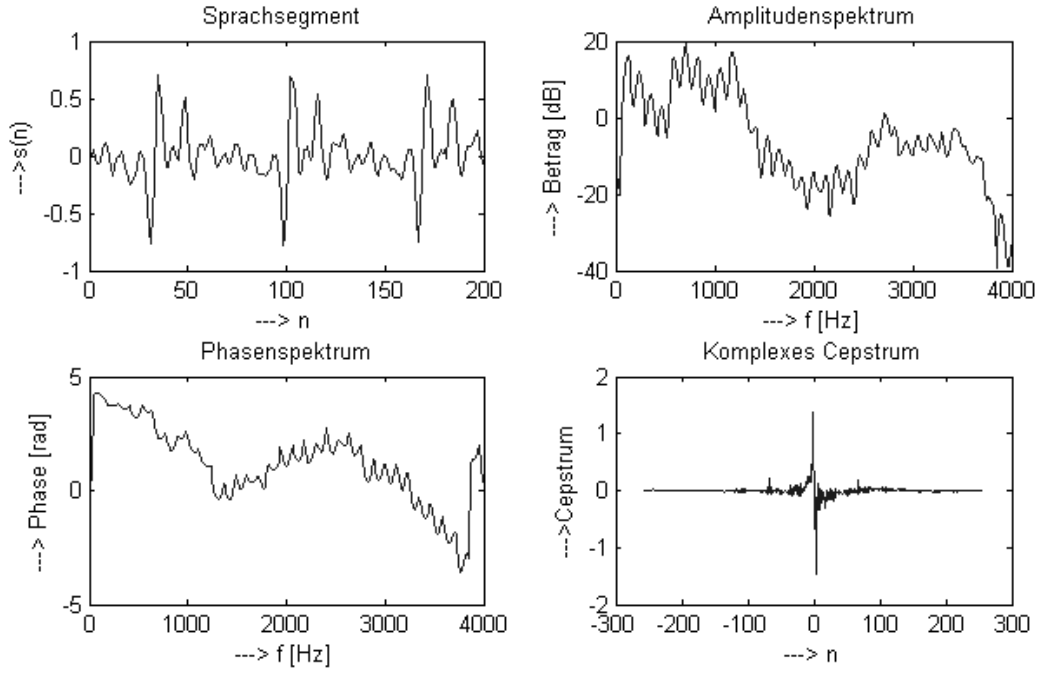

Abbildung 2 – Stationärer Teil des Lautes „a“, das Amplituden- und Phasenspektrum und das komplexe Cepstrum.

3 Cepstrale Sprachsynthese

Die Übertragungsfunktion $H(z)$ eines nichtrealisierbaren und nichtkausalen digitalen Filters, dessen logarithmischer Frequenzgang das logarithmische Spektrum des Vokaltraktmodells approximiert, wird aus dem gefensterten komplexen Cepstrum konstruiert. Als Fenster verwenden wir ein symmetrisches Rechteckfenster $\{w_n\}$ der Länge $2N_0$, wo für n folgende Ungleichung $-M_0 < -N_0 \le n \le N_0 < M_0$ gilt. Das logarithmische Spektrum des Vokaltraktmodells ist dann

$$\ln S_{N_0 k} = \hat{S}_{N_0 k} = \sum_{n=-N_0}^{+N_0} \hat{s}_n e^{-\mathrm{j}2\pi nk/M} \,. \tag{5a}$$

In Abb. 3 sind das Amplitudenspektrum des Lautes „a“ und der durch cepstrale Fensterung konstruierte Vokaltraktfrequenzgang, die *"Einhüllende"* für unser Beispiel dargestellt. Das Rechteckfenster $\{w_n\}$ hat eine Länge $2N_0 = 68$.

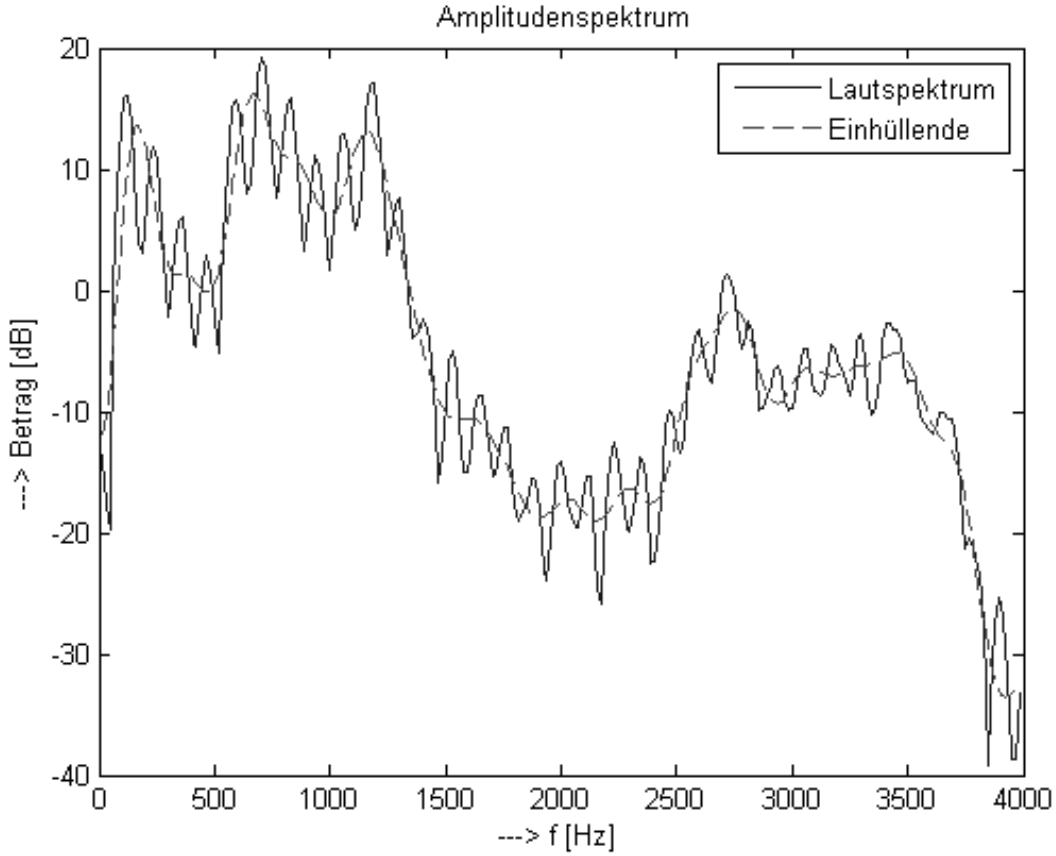

Abbildung 3 – Amplitudenspektrum und das durch cepstrale Fensterung erhaltene geglättete Spektrum des Lautes „a“.

Man kann das logarithmische Spektrum des Vokaltraktmodells in drei Teile zerlegen

$$\ln S_{N_0 k} = \hat{S}_{N_0 k} = \hat{s}_o + \sum_{n=-N_0}^{-1} \hat{s}_n e^{-\mathrm{j}2\pi nk/M} + \sum_{n=1}^{+N_0} \hat{s}_n e^{-\mathrm{j}2\pi nk/M} = \hat{s}_o + \hat{S}_{\mathrm{A},k} + \hat{S}_{\mathrm{K},k} \,. \tag{5b}$$

$\hat{S}_{\mathrm{A},k}$ entspricht dem logarithmischen Spektrum des *antizipativen Teilcepstrums*, $\hat{S}_{\mathrm{K},k}$ dem logarithmischen Spektrum des *kausalen Teilcepstrums*.

$$\hat{S}_{\mathrm{A},k} = \sum_{n=-N_0}^{-1} \hat{s}_{\mathrm{A},n} e^{-\mathrm{j}2\pi nk/M} \,, \tag{6a}$$

$$\hat{S}_{\mathrm{K},k} = \sum_{n=1}^{+N_0} \hat{s}_{\mathrm{K},n} e^{-\mathrm{j}2\pi nk/M} \,. \tag{6b}$$

Der Parameter $\hat{s}_0$ entspricht dem Mittelwert des logarithmischen Betragsspektrums.

Die Zerlegung des Cepstrums $\{\hat{s}_n\}$in den linksseitigen und rechtsseitigen Teil wird mit Hilfe von zwei cepstralen Rechteckfenstern $\{w_{\mathrm{A},n}\}$ und $\{w_{\mathrm{K},n}\}$der Länge N_0 durchgeführt, wo $w_{\mathrm{A},0} = w_{\mathrm{K},0} = 0$, und $w_{\mathrm{A},n} = 1$ für $-N_0 \le -n < 0$ und $w_{\mathrm{K},n} = 1$ für $0 < n \le N_0$ ist.

Die Übertragungsfunktion *eines nichtrealisierbaren und nichtkausalen digitalen Filters* ist durch folgende Beziehung gegeben

$$H(z) = \exp\Big(\sum_{n=-N_0}^{+N_0} \hat{s}_n z^{-n}\Big) = e^{\hat{s}_0} H_{\mathrm{A}}(z) H_{\mathrm{K}}(z) = K H_{\mathrm{A}}(z) H_{\mathrm{K}}(z) \,, \tag{7}$$

wo $H_{\mathrm{A}}(z) = \exp\Big(\sum_{n=-N_0}^{-1} \hat{s}_{\mathrm{A},n} z^{-n}\Big)$ der *antizipative Teil* und $H_{\mathrm{K}}(z) = \exp\Big(\sum_{n=1}^{+N_0} \hat{s}_{\mathrm{K},n} z^{-n}\Big)$ der *kausale Teil* der Übertragungsfunktion $H(z)$ sind und $K = e^{\hat{s}_0}$ der Verstärkungsfaktor ist. Die *mischphasige nichtkausale Impulsantwort* $\{h_n\}$ erhalten wir mit Hilfe von

$$h_n = \frac{1}{M} \sum_{k=0}^{M-1} H(z_k) e^{\mathrm{j}2\pi kn/M} \,, \tag{8}$$

wo $z_k = e^{\mathrm{j}2\pi k/M}$. Man kann auch geteilt die *antizipative (maximalphasige) Impulsantwort* $\{h_{\mathrm{A},n}\}$ und die *kausale (minimalphasige) Impulsantwort* $\{h_{\mathrm{K},n}\}$ von $H_{\mathrm{A}}(z)$ und $H_{\mathrm{K}}(z)$ bestimmen und die gesamte Impulsantwort $\{h_n\}$ durch Faltung dieser beiden Teilantworten konstruieren.

Die kausale Impulsantwort $\{h_{\mathrm{K},n}\}$ ist mit dem zugehörigen Cepstrum $\{\hat{s}_{\mathrm{K},n}\}$ durch eine zeitvariable rekursive Beziehungen verbunden [1,4]

$$h_{\mathrm{K},n} = \hat{s}_{\mathrm{K},n} h_{\mathrm{K},0} + \sum_{m=0}^{n-1} \left(\frac{m}{n}\right) \hat{s}_{\mathrm{K},m} h_{\mathrm{K},n-m} \,, \qquad h_{\mathrm{K},0} = 1 \,. \tag{9}$$

Die antizipative Impulsantwort $\{h_{\mathrm{A},n}\}$ kann auf ähnliche Weise wie $\{h_{\mathrm{K},n}\}$ mit derselben Beziehung bestimmt werden, indem wir die Zeitindexierung von $\{\hat{s}_{\mathrm{A},n}\}$ umkehren, mit Hilfe von (9) eine fiktive kausale Impulsantwort bestimmen und aus der, durch weitere Zeitumkehrung, die aktuelle antizipative Impulsantwort $\{h_{\mathrm{A},n}\}$ erhalten. Dieser Vorgang gemeinsam mit der Faltung und Einbeziehung von $K = e^{\hat{s}_0}$ ist in Abb. 4 dargestellt.

An dieser Stelle ist Gelegenheit zu einer wichtigen Bemerkung. Obwohl das komplexe Cepstrum auch für eine endliche Folge, für unser gefenstertes Sprachsignal $\{s_n\}$ der Länge N unendlich ist, genügen zur Berechnung der kausalen Impulsantwort $\{h_{\mathrm{K},n}\}$ nur $N/2+1$ Werte des Teilcepstrums $\{\hat{s}_{\mathrm{K},n}\}$. Ähnliche Überlegung dann gilt auch für die Berechnung der

antizipativen Impulsantwort $\{h_{A,n}\}$, hier genügen nur $N/2$ Werte des Teilcepstrums $\{\hat{s}_{A,n}\}$. Die Länge der gesamten Impulsantwort $\{h_n\}$ ist dann gleich *N*.

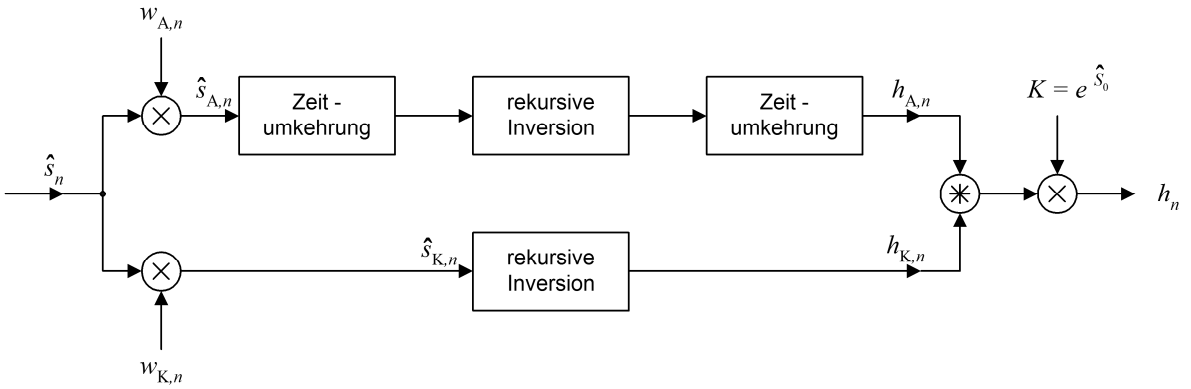

Abbildung 4 – Rekursive Konstruktion der Teilimpulsantworten $\{h_{K,n}\}, \{h_{A,n}\}$ und die durch Faltung und Skalierung erhaltene gesamte Impulsantwort $\{h_n\}$.

Für unser Beispiel finden wir die Teilcepstren, Teilspektren, und Teilimpulsantworten des Lautes „a“ in Abb. 5 und die gesamte Impulsantwort $\{h_n\}$ in Abb. 6. Die Länge der Impulsantwort $\{h_n\}$ ist der Länge des Sprachsegmentes *N* gleich. Die *mischphasige kausale FIR* Impulsantwort wird aus diesem Signal mit üblicher *Fenstermethode* der FIR Synthese konstruiert.

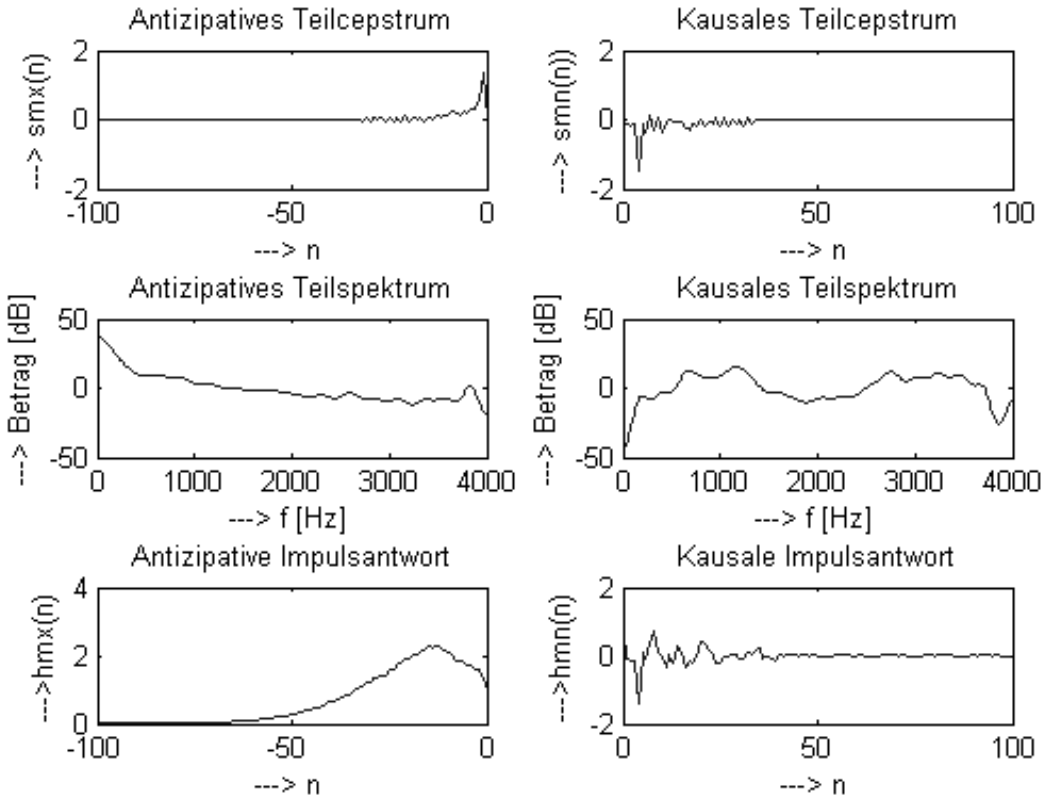

Abbildung 5 – Teilcepstren, Teilspektren, und Teilimpulsantworten des Lautes „a“.

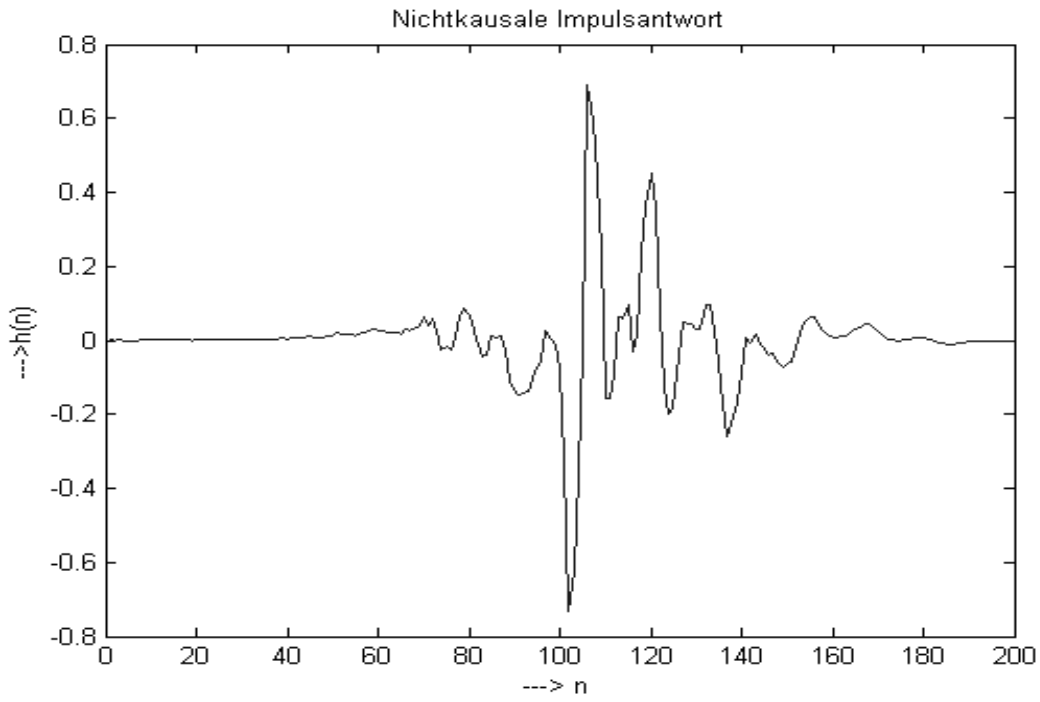

Abbildung 6 –Nichtkausale Impulsantwort des Vokaltraktmodells für den Laut „a“.

Die Synthese des Lautes „a“ wird für unser Beispiel realisiert durch Faltung der FIR Impulsantwort $\{h_n\}$ mit einer Folge $\{p_n\}$ von periodischen Einheitsimpulsen mit der Periode $M_0 = 68$, $p_n = \sum_{i=0}^{[N/68]} \delta_{n-i68}$, wo $\{\delta_n\}$ der diskrete Einheitsimpuls ist.

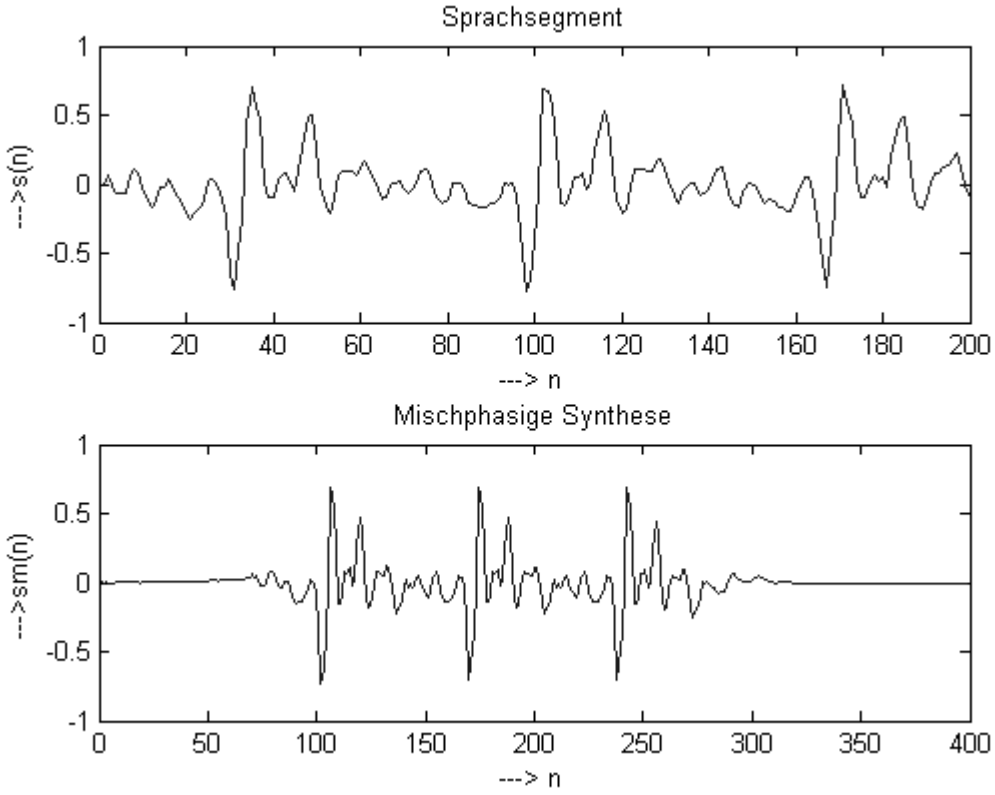

Abbildung 7 – Stationärer Teil des Lautes „a“ und dessen mischphasige Synthese.

In Abb. 7 sind die mischphasige Synthese $\{s_{\mathrm{M},n}\}$ zusammen mit dem Lautsegment $\{s_n\}$ dargestellt. Die Synthese kann auch schrittweise implementiert werden, in dem man das Eingangssignal des Vokaltraktes zuerst mit der kausalen Impulsantwort $\{h_{\mathrm{K},n}\}$ und nachher mit der antizipativen Impulsantwort $\{h_{\mathrm{A},n}\}$ faltet und mit dem Verstärkungsfaktor K skaliert. Man kann sehen, dass die mischphasige Synthese ziemlich treu den Vokalzeitverlauf approximiert. Der hörbare Unterschied zwischen beiden Synthesen kann auch festgestellt werden.

4 Schlussbetrachtung

Die Konstruktion des mischphasigen FIR cepstralen Vokaltraktmodells ist relativ einfach und bietet natürlichere Sprachsynthese. Nachteilig sind aber die höheren Speicheransprüche und auch die größere numerische Komplexität. Das verwendete Beispiel ist numerisch sehr stabil und führt zu keinen Schwierigkeiten. In manchen Fällen bei gewissen Lauten kommen große Werte des komplexen Cepstrums vor, die wahrscheinlich zusammen mit der cepstralen Überfaltung bei der Berechnung von Teilimpulsantworten ein Überlaufen verursachen. Diese Probleme entstehen aber nicht bei der direkten Berechnung der gesamten mischphasigen Impulsantwort $\{h_n\}$ mit Hilfe der FFT. An möglichen Vereinfachungen und der Implementierung der cepstralen Synthese in das tschechische triphonbasierte Text-to-Speech System wird gearbeitet.

Danksagung

Die vorliegende Arbeit wurde im Rahmen von COST2102 vom Ministerium für Bildung, Jugend und Sport der Tschechischen Republik, unter dem Kennzeichen OC08010 und der Projekte „Gezielte Forschung“ von der Grantagentur der Akademie der Wissenschaften der Tschechischen Republik unter dem Kennzeichen 1QS108040569 unterstützt.

Literatur

[1] Oppenheim, A. V., Schafer, R. W.: Discrete-Time Signal Processing. Prentice Hall, 1989.
[2] Quatieri, T., F.: Discrete-Time Speech Signal Processing. Principles and Practice. Prentice Hall, 2002.
[3] Vích, R.: Komplexes Cepstrum in der Sprachsynthese. In: A. Lacroix (ed.): Festschrift aus Anlass des 80. Geburtstags von Herrn Prof. Dietrich Wolf. Studientexte zur Sprachkommunikation, TUD Dresden, 2009, (in Druck).
[4] Vích, R.: Z-Transform Theory and Application. Dordrecht, D. Reidel Publishing Company, 1987.