Rekonstruktion stationärer Mannigfaltigkeiten der Unterbanddynamik stimmhaft instationärer Sprachsignale

F.R. Drepper
Forschungszentrum Jülich GmbH; email: f.drepper@fz-juelich.de

Stimmhafte Sprache zeichnet sich durch qualitativ vielfältige mode locking Phönomene zwischen harmonisch angeregten akustischen Moden des Vokaltraktes aus, deren differenzierte Analyse jedoch aufgrund der starken Instationarität der Sprache mit dem vorherrschenden linearen Quelle-Filter Modell mit stationärer Quelle nur unvollständig gelingt. Durch Einführung einer an den Phonationsprozess angepassten Unterbandzerlegung mit Gehör angepassten Bandbreiten und Unterband spezifischen Quelle - Filter (Treiber – Response) Modellen mit Fundamentalbandphasen synchronen Quellen wird eine topologisch äquivalente Rekonstruktion einer Vielzahl von akustischen Moden des Sprachtraktes erzeugt, die eine differenzierte Analyse der besagten mode locking Phänomene erlaubt. Das für stimmhafte Tonkomplexe charakteristische mode locking drückt sich in stationären (invarianten) Synchronisations-Mannigfaltigkeiten (Linien oder Flächen) im gemeinsamen Zustandsraum der Unterbänder aus. Die Phonem spezifischen Eigenschaften werden einerseits als topologisch invariante Resonatoreigenschaften der rekonstruierten Responsedynamik und andererseits als Phonemklassen spezifischer Verlust der linearen Phasensynchronisation der Quellen sichtbar (hörbar).

Im Rahmen der Theorie der verallgemeinerten Synchronisation in Treiber - Response Systemen ist in vergleichsweise jüngerer Vergangenheit gezeigt geworden, dass mode-locking bzw. Synchronisation kein undifferenziertes Phänomen darstellt, sondern einen Oberbegriff für eine Vielzahl von insbesondere auch qualitativ unterschiedlichen Koordinationsmöglichkeiten, die durch mehr oder weniger glatte und/oder stetige invariante Mannigfaltigkeiten im gemeinsamen Zustandsraum der Treiber und Responseoszillatoren beschrieben werden können (Haken 1977, Rulkov et al. 1995). Ein im Zusammenhang mit der Spracherkennung wichtiger Spezialfall verallgemeinerter Synchronisation ist die topologische Äquivalenz zwischen Treiber und Response, die durch eine Konjugation (monoton stetige, eindeutig invertierbare Abbildung) zwischen den gekoppelten Subsystemen ausgezeichnet ist (Kocarev and Parlitz 1996).

Ein zweites wesentliches Merkmal stimmhafter Sprache ist die ubiquitär instationäre Variation der Amplitude und Tonhöhe der Phonation. In diesem Zusammenhang ist es von besonderem Interesse, dass das altbekannte Phänomen der Synchronisation nicht auf periodischen oder quasiperiodischen Antrieb beschränkt ist, sondern auch bei stochastischen (Afraimovich et al. 1986) oder deterministisch chaotischen Treibern (Rulkov et al. 1995) auftritt. Die bisherige Anwendung des Quelle Filter Modells auf die Erkennung stimmhafter Sprache beruht auf der Annahme eines stationären Phonationsprozesses (Vary et al. 1998). Diese Annahme beschränkt das Quelle – Filter Modell jeweils auf die Beschreibung relativ kurzer Ausschnitte stimmhafter Sprache (typischerweise 20 ms). Derartig kurze Ausschnitte sind jedoch nur ungenügend geeignet, die für stimmhafte Phoneme charakteristischen invarianten Mannigfaltigkeiten zu erkennen. Der hier vorgestellte Zugang zur Sprachakustik fasst die Atome bzw. Objekte (insbes. die Phoneme) der stimmhaften Sprache nicht mehr als stationäre Prozesse auf sondern als stationäre bzw. invariante Mannigfaltigkeiten im gemeinsamen Zustandsraum der instationären Treiber und Response-Oszillatoren.
Als wesentliches Merkmal des hier vorgestellten Verfahrens werden mithilfe geeignet gewählter Bandpassfilter sowohl eine fundamentale Treibermode als auch höherfrequente (harmonische) Unterbänder bestimmt, die jeweils topologisch äquivalente Rekonstruktionen entsprechend angeregter akustischer Moden des Resonanzraumes darstellen. Die räumliche und frequenzmäßige Konzentration der Bewegungsenergie der Schallquelle bewirkt bzw. begünstigt, dass die Anregung des Schallfeldes durch einen elementaren Oszillator mit zwei vergleichsweise langsam veränderlichen Zustandsgrößen beherrscht wird, die die übrigen, schnelleren Zustandsgrößen der Phonation bzw. deren Auswirkung auf das Schallfeld „versklaven“ (Haken 1977). Hierdurch wird es möglich, die Anregung der akustischen Moden als eine Synchronisationsmannigfaltigkeit eines fundamentalen Treiberoszillators darzustellen, dessen potentiell instationäre Dynamik vollständig durch eine Treiberamplitude und eine Treiberphase beschrieben wird.

In Anlehnung an das weitverbreitete lineare Quelle-Filter Modell der Spracherzeugung liegt es nahe, die Unterband spezifische Anregung als Produkt der Treiberamplitude A und einer oszillierenden Treiberphasen abhängigen Anregungsfunktion $G(\psi)$ darzustellen. Die Versklavung der schnellen Freiheitsgrade der Anregung bewirkt eine Periodizität der Treiberphasen abhängigen Anregungsfunktion, wobei es im Zusammenhang der Behandlung instationärer Prozesse von entscheidender Bedeutung ist, dass sich die Periodizität nicht auf die Zeit sondern auf die Treiberphase bezieht. Die sprecherabhängige Periodenlänge fällt vielfach mit der fundamentalen Periode der Treibermode zusammen. Aufgrung ihrer Bandlimitierung lassen sich die Anregungsfunktionen gut durch eine endliche Fourier Reihe approximieren, deren Terme jeweils als rein harmonische Elementaranregungen interpretiert werden können, die durch die fundamentale Treiberphase synchronisiert werden. In Anlehnung an das lineare Quelle-Filter Modell liegt es nahe, die Unterbanddynamik des $j$ten Bandes als einen endlich dimensionalen, linearen Response einer Treiberphasen synchronen Quelle zu approximieren. Aufgrund der gehörangepassten Bandbegrenzung dieser Unterbänder (sowie aufgrund einer Unterband angepassten Wahl der Zeitschrittweite Δ) reicht hierbei eine zweidimensionale lineare Responsedynamik aus.

$$X_{j,(n+1)\Delta} = -a_j X_{j,n\Delta} - b_j X_{j,(n-1)\Delta} + A_{n\Delta} G_j(\psi_{n\Delta}) \qquad (n = 0,1,...)$$

Das Ziel der Phonationsprozess angepassten Bandpasszerlegung besteht darin, Unterbänder zu erzeugen, die jeweils als linearer Response auf möglichst nur eine der rein harmonischen Elementaranregungen dargestellt bzw. approximiert werden können. Die Filtermittenfrequenzen der Bandpassfilter stellen hierbei die wesentlichen Anpassungsgrößen bei der Diagonalisierung der Unterbandanregungen dar.

Wenn der Einfluss einer zu starken Instationarität des Treiberprozesses ausgeschlossen werden kann, deutet das weitgehende Scheitern der Diagonalisierung auf eine konsonantische Artikulations-

konstellation im Vokaltrakt hin. Hierbei kann ein partielles Scheitern der Diagonalisierung als ein stimmhafter Konsonant (z.B. mit einer zweiten von der ersten Schallquelle nur partiell abhängigen bzw. versklavten Schallquelle) gedeutet werden und ein vollständiges (alle Unterbänder betreffendes) Scheitern als ein stimmloser Konsonant. Sowohl im Fall des weitgehenden Gelingens der Diagonalisierung als auch im Fall des unvollständigen Gelingens liefert der Vergleich der invarianten Mannigfaltigkeiten der Anregung mit den Synchronisations bzw. Koordinationseigenschaften der hierdurch jeweils hervorgerufenen Responsprozesse wertvolle Hinweise zur Art des Phonems. Bei weitgehender Diagonalstruktur der Anregungen kann der Fall nahezu rein harmonischer Responsemannigfaltigkeiten als Vokal und der Fall des weitgehenden Verlustes der sog. Identischen Synchronisation der Elementaranregungen als Nasal interpretiert werden. Im Fall der unvollständigen Diagonalisierung der Anregungen richtet sich die Phonemerkennung sowohl auf qualitative Eigenschaften der bandspezifischen Anregungs-Mannigfaltigkeiten als auch auf die Untersuchung des Synchronisationsverlustes auf dem Weg von der Anregung zum Response.

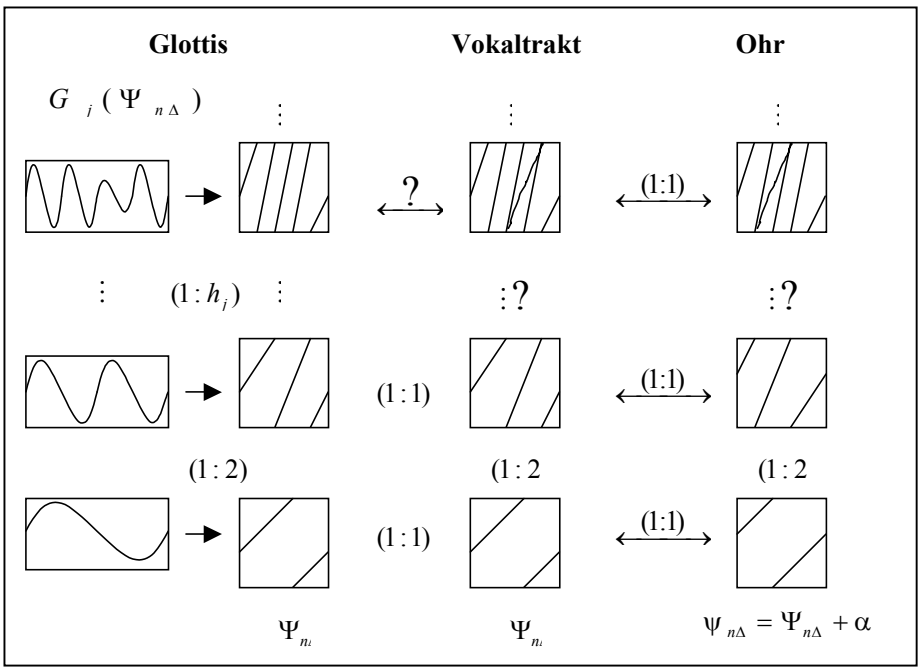
Die Bandbreiten der besagten Bandpassfilter sollten nach Möglichkeit kleiner sein als der doppelte Frequenzabstand zur nächsthöheren Harmonischen, andererseits jedoch auch hinreichend breitbandig sein, sodass die relative Bandbreite oberhalb der für das jeweilige Betrachtungsintervall relevanten, relativen Bandbreite des instationär oszillierenden fundamentalen Treiberprozesses bleibt. Die aus Maskierungsexperimenten der Psychoakustik bekannten, gehörangepassten Bandbreiten z.B. nach dem ERB (equivalent rectangular bandwidth) Modell (Moore 1989) stellen einen offenbar evolutionär erprobt günstigen Kompromiss dar.

Ein wesentlicher Vorteil der beschriebenen Unterbandzerlegung besteht darin, dass der Erfolg einer Rekonstruktion der Dynamik der akustischen Moden des Resonanzraumes auch unabhängig von den bisher genannten Voraussetzungen überprüft werden kann. Das für stimmhafte Signale wesentliche mode locking bezieht sich bekanntlich auf akustische Moden, deren Frequenzen Vielfache der Grundfrequenz betragen. Die Einführung von zeitabhängigen und zeitverwandten (partiell stetig fortgesetzen bzw. abgewickelten) Phasen als Zustandsvariable der Anregungs und Responsedynamik schafft die Möglichkeit auch (1:n) bzw. (n:m) Mode- bzw. Phase locking als (mehr oder weniger lineare, diffeomorphe) Konjugation zu erkennen. Aufgrund der Transitivität und Umkehrbarkeit von Konjugationen in einer Kette von konjugiert gekoppelten Oszillatoren wird die Evidenz einer nahezu linearen Konjugation zwischen den Unterbandoszillatoren eines stimmhaften Signals zu einer Bestätigung für die topologische Äquivalenz aller beteiligten Oszillatoren, einschließlich der Äquivalenz zwischen den jeweiligen harmonisch angeregten Schwingungsmoden im Resonanzraum des Untersuchungssystems und den entsprechenden bandpassgefilterten Unterbändern (Figur 1).

Als eine hochgradig nichtzufällige Eigenschaft zeichnet sich stimmhafte Sprache bei niederharmonischen Unterbändern generell durch eine gute Überprüfbarkeit der topologischen Äquivalenz aus (Figur 1). Insbesondere die fundamentale Treiber-Phase kann als fast lineare Konjugation zu niederharmonischen Unterbandphasen bestimmt and somit als ein topologisch äquivalentes Abbild der Phase der fundamentalen akustischen Mode bestätigt werden. Hierdurch wird eine zuverlässige und präzise Bestimmung der momentanen Tonhöhe stimmhafter Sprache ermöglicht.

Da sowohl die Anregungs-als auch die Response-Mannigfaltigkeiten durch endliche Fourier Reihen approximiert werden, können die phänomenologischen Parameter mit Hilfe multipler linear Regression bestimmt werden. Der wesentliche Unterschied des hier vorgestellten Modells zum vorherrschenden Breitband Quelle – Filter Modell besteht in der Darstellung der Quelle als Überlagerung einer Reihe verallgemeinert synchronisierter Anregungen, deren gemeinsamer, fundamentaler Treiber zusammen mit den zugehörigen Responsbändern sowohl Phonationsprozess als auch Vokaltraktresponse äquivalent bzw. angepasst bestimmt wird.

Afraimovich V.S., N.N. Verichev, M.I. Rabinovich, *Radiophys. Quantum Electron*. 29, 795 (1986)
Haken H., *Synergetics*, Springer Verlag, Berlin (1977)
Kocarev L., U. Parlitz, *Phys. Rev. Lett*. 76, 1816 (1996)
Moore B.C.J., *An introduction to the Psychology of hearing*, Academic Press (1989)
Rulkov N.F. , M.M. Sushchik, L.S. Tsimring, H.D.I. Abarbanel, *Phys. Rev. E* 51, 980-994 (1995)
Vary P., U. Heute, W. Hess, *Digitale Sprachsignalverarbeitung*, B.G. Teubner Verlag, Stuttgart (1998)



Figur 1 Stimmhafte Tonkomplexe der Sprache zeichnen sich durch stationäre Mannigfaltigkeiten (Linien oder Flächen) aus, die sich sowohl hinsichtlich des Abstandes zur Tonerzeugung in der Glottis als auch hinsichtlich der jeweiligen Oszillations- bzw. Windungszahl der Unterband spezifischen Anregung unterscheiden. Zur besseren Anschauung werden sowohl die Anregungen als auch die hierdurch erzeugten Responseprozesse durch zeitverwandte (aufgewickelte) Phasenvariablen dargestellt. Bei Vokaltrakt-äquivalenter Bandpassfilterung überträgt sich bei niederfrequenten harmonischen Anregungen die für glottale Anregung charakteristische, angenähert lineare Phasensynchronisation zur fundamentalen Treiberphase sowohl auf den Vokaltrakt als auch auf die Unterbänder des Schalldrucksignals am Ohr. Bei höherfrequenten harmonischen Anregungen geht die lineare Phasensynchronisation je nach Phonemklasse mehr oder weniger weitgehend im Vokaltrakt verloren.