Stimmhafte Anregung als synchronisierte Antwort eines eindeutigen fundamentalen Treibers

F.R. Drepper

Forschungszentrum Jülich GmbH, 52425 Jülich, f.drepper@fz-juelich.de

Das stimmhafte Sprachsignal wird als stochastische sekundäre Antwort eines bandbegrenzten fundamentalen Treibers dargestellt, der selbstkonsistent aus dem Sprachsignals rekonstruiert wird. Die selbstkonsistente Rekonstruktion stützt sich auf eine stimmenspezifische Teilbandzerlegung und schließt eine Bestätigung der topologischen Äquivalenz zwischen dem fundamentalen Treiber (FD) und einem glottalen Masteroszillator der stimmhaften Phonation ein. Von dem bandbegrenzten glottalen Masteroszillator wird angenommen, dass er die breitbandige stimmhafte Anregung und mehrere Unterbänder des Sprachsignals synchronisiert und beim menschlichen Empfänger die Tonhöhenwahrnehmung hervorruft. Die synchronisierte primäre Antwort wird durch eine periodische Kopplungs- und/oder Amplituden Modulationsfunktion der Phase des fundamentalen Treibers dargestellt. Die topologisch invarianten Eigenschaften der primären Antwort werden als zusätzliche Kenngrößen der automatischen Spracherkennung vorgeschlagen.

Obwohl Sprachsignale sich bekanntlich durch ein hohes Maß an Instationarität auszeichnen, geht die bisherige Ermittlung der akustischen Merkmalsvektoren der maschinellen Spracherkennung von der Annahme aus, dass sich die Sprachprokuktion auf der Zeitskala von etwa 20 ms als ein lineares zeitinvariantes (LTI) System auffassen lässt. Die Annahme eines LTI-Systems wird hierbei entweder als Voraussetzung bei der konsistenten Schätzung von Fourier Spektren benutzt oder bei der Schätzung von autoregressiven Modellen [2, 3]. Im letzteren Fall wird vielfach ein Treiber – Response (Quelle Filter) Modell benutzt, bei dem sich der autoregressive Modellteil auf die Beschreibung der Resonanzeigenschaften des Vokaltrakt beschränkt. Lineare autoregressive Modelle sind in besonderer Weise geeignet, Transienten mit unterschiedlichen Abklingraten in unterschiedlichen Frequenzbereichen zu beschreiben einschließlich relativ langer Transienten (Formanten). Die Abklingraten (Liapunov Exponenten) solcher Transienten stellen bekanntlich topologische Invarianten [4] dar, welche als robust bei Veränderung des Kommunikationskanals angenommen werden können und bekanntlich wichtige Merkmale der Vokale darstellen [2, 3]. Der konventionelle LTI System Zugang stellt sich jedoch im Fall der stimmhaften Sprache als besonders problematisch heraus. Der Vokaltrakt kann nicht als zeitinvariant angesehen werden [2, 3] und von der Quelle kann nicht angenommen werden, dass sie durch ein autonomes lineares System erzeugt wird [5].

Als Ausgangspunkt einer konsistenteren phänomenologischen Beschreibung der Aero-Akustik der stimmhaften Sprache wird ein zusätzlicher Treiber – Response Schritt eingeführt, der die offenbar hoffnungslos irreguläre, breitbandige Sprachanregung als stationäre primäre Antwort eines instationären, bandbegrenzten, fundamentale Treibers im Frequenzbereich der Tonhöhenwahrnehmung beschreibt. Die Bedeutung, Allgemeingültigkeit und Präzision der Tonhöhenwahrnehmung kann als ein wichtiger Hinweis gedeutet werden, dass der verborgene fundamentale Treiber direkt anhand des Sprachsignals ermittelt werden kann. Die Zweistufigkeit der Systemantwort kann dazu benutzt werden, jeweils komplementäre Vereinfachungen der Systemdynamik einzuführen.

Auf der Ebene der sekundären Treiber-Response Wechselwirkung hat sich die Annahme linearer autoregressiver (Allpol) Filter mit stabilem Punkt Attraktor offenbar gut bewährt [2, 3]. Demgegenüber sollte angenommen werden, dass der primäre Response stimmhafter Sprache zumindest partiell durch nichtlineare Dynamik mit Attraktoren einer von Null verschiedenen Dimension erzeugt wird. Als ein nützlicher Kontrast zu den langen Transienten einer typischen sekundären Antwort wird von der primären Antwort angenommen, dass sie durch stark dissipative nichtlineare Dynamik erzeugt wird, die überwiegend kurze Transienten erzeugt. Solche Dynamik kann drastisch vereinfacht werden, indem ein perfektes phase locking bzw. eine verallgemeinerte Synchronisation des Antwortoszillators durch einen treibenden Oszillator angenommen wird [6,8]. Die synchronisierte Dynamik ist durch eine Kopplungsfunktion charakterisiert, die eine zentrale Mannigfaltigkeit (Linie oder Fläche) im gemeinsamen Zustandsraum von Treiber und Response darstellt [10, 11]. Die Mannigfaltigkeit zieht die schnellen Transienten an, bzw. ignoriert diese. Die Unterscheidung zwischen der akustischen Quelle und dem fundamentalen Treiber stellt sich als nützlicher Ausgangspunkt heraus, um zusätzliche Merkmale stimmhafter Sprache offen zu legen, die bei Veränderung des Kommunikationskanals invariant bleiben, sowie um die phonetisch relevante Dynamik vom nieder-frequenten Jitter, Mikrotremor [12] und der Prosodie zu trennen. Eine eng verwandte Kopplungsfunktion (wave-shaper function) wurde erfolgreich eingesetzt, um Vokale mit natürlichem und pathologischem Jitter und Mikrotremor zu synthetisieren [13].

Auto-Synchronisation der stimmhaften Anregung

Das Phänomen der Synchronisation gekoppelter periodischer Oszillatoren ist seit der Zeit von C. Huygens bekannt. In neuerer Zeit sind auch Fälle der Synchronisation zwischen aperiodischen Oszillatoren untersucht worden, insbesondere die Fälle einseitiger Kopplung bei stochastischen [10] und chaotischen [11] Treibern. In diesen Fällen beschreibt die attraktive Synchronisationsmannigfaltigkeit den momentanen Zustand der Antwort als stetige Funktion des jeweiligen Treiberzustands.

Stetige Synchronisationsmannigfaltigkeiten sind nicht auf den Fall ausschließlich einseitiger Kopplung beschränkt. Der Begriff eines Treiber-Response Systems wird daher im Folgenden auch im Sinne eines *überwiegend* einseitig gekoppelten Systems verstanden. Aufgrund des großen Dichteunterschiedes (von 1000:1) zwischen Gewebe und Luft passen sich die akustischen Moden (schnellen oszillatorischen Freiheitsgrade der aeroakustischen Dynamik) im Kehlkopfbereich nahezu instantan den veränderlichen Randbedingungen des Strömungskanals an, die insbesondere durch die vergleichsweise langsame Bewegung der Stimmlippen vorgegeben bzw. beeinflusst werden. Deshalb kann davon ausgegangen werden, dass die Ankopplung der akustischen Moden an die Stimmlippen eine dominante Wirkungsrichtung aufweist und dass das der Synchronisationsannahme zugrunde liegende Entrainment vorliegt.

Ein besonders einfacher Fall der Synchronisation zwischen nichtlinearen Oszillatoren tritt im Fall der nahezu symmetrisch gekoppelten Dynamik der beiden Stimmlippen auf. Im Fall nichtpathologischer Phonation kann die Synchronisationsmannigfaltigkeit dieses gekoppelten Systems durch eine ein-eindeutige stetige Funktion (Konjugation) beschrieben werden. D.h. die beiden Oszillatoren werden topologisch äquivalent [4] bzw. verhalten sich wie ein einzelner Oszillator. Außer im pathologischen Fall der sog. Biphonation [12] kann daher ein glottaler Masteroszillator mit nur einem unabhängigen Freiheitsgrad definiert werden, der die anderen oszillatorischen Freiheitsgrade des Kehlkopfes synchronisiert und die akustischen Moden des Vokaltraktes antreibt. Es kann daher davon ausgegangen werden, dass der Zustand des fundamentalen Treibers durch eine fundamentale Amplitude und eine fundamentale Phase eindeutig beschrieben

wird, wobei letztere in Abhängigkeit des Stimmtypus möglicherweise partiell abgewickelt werden muss [6-8, 12].

Für stimmhafte Kontinuanten (verlängerbare Laute mit aktiver Phonation) wird die phonetisch relevante Dynamik überwiegend durch die zeitliche Entwicklung der fundamentalen Phase bestimmt. Die Phasengeschwindigkeit des fundamentalen Treibers kann als akustisches Korrelat der Tonhöhenwahrnehmung identifiziert werden. Neben der Tonhöhenwahrnehmung hat die Psychoakustik die Lautheitswahrnehmung als eine noch allgemeingültigere akustische Wahrnehmung herausgestellt. Es bietet sich daher an, die Amplitude des fundamentalen Treibers in enger Beziehung zum akustischen Korrelat der Lautheitswahrnehmung [7, 8] zu wählen.

Da Funktionen von Phasen periodisch sein müssen, kann die stetige Abhängigkeit der Kopplungsfunktion von der fundamentalen Phase gut durch eine endliche Fourierreihe approximiert werden, die die geeignete Periodizität aufweist [6-8]. Im Fall der höherfrequenten akustischen Moden des Kehlkopfes wird die Synchronisation notwendigerweise durch eine multimodale Kopplungsfunktion beschrieben, die ein (n:1) phase locking zur fundamentalen Phase aufweist. Das phase locking der höherfrequenten akustischen Moden des Kehlkopfes wird jedoch durch die Resonanzen (Formanten) und eine optionale Verengung des Vokaltraktes zum Teil zerstört. Eine wesentliche Eigenschaft des Übertragungsprotokolls der stimmhaften Sprache besteht offenbar darin, dass im Frequenzbereich, in dem der menschliche Hörnerv Phaseninformation weiterleitet, d.h. im Frequenzbereich bis ca. 3 kHz mehrere Unterbänder des Sprachsignals bestehen bleiben, für die der Vokaltrakt hinreichend transparent ist, um anhand dieser Unterbänder den glottalen Masteroszillator zu rekonstruieren. Die nachgewiesene Synchronisation der Phasen mehrerer unabhängiger Unterbänder kann sowohl zur Bestätigung der topologischen Äquivalenz zwischen dem fundamentalen Treiber und einem glottalen Masteroszillator benutzt werden als auch für die Rekonstruktion einer eindeutigen Phasengeschwindigkeit des fundamentalen Treibers [6, 8, 9]. Als eine charakteristische Eigenschaft der selbstkonsistenten Rekonstruktion der Phase des fundamentalen Treibers wird die Teilband Zerlegung des Sprachsignals mittels zeitabhängiger Bandpassfiltern durchgeführt, deren Filtermittenfrequenzen an die Phasengeschwindigkeit des fundamentalen Treibers angepasst werden [9].

Als eine weitere bemerkenswerte Eigenschaft des stimmhaften Übertragungsprotokolls stellt es sich heraus, dass die Rekonstruktion eines kohärenten fundamentalen Treibers typischerweise auf einem größeren Zeitsegment gelingt als die Rekonstruktion der jeweiligen Treiber - Responses Beziehung zwischen dem bandbegrenzten fundamentalen Treiber und dem breitbandigen Sprachsignal [6, 8]. Während letztere Rekonstruktion zweckmäßigerweise in einem nur leicht vergrößerten Zeitfenster von ca 30-40 ms erfolgt, gelingt die Rekonstruktion des fundamentalen Treibers für Zeitsegmente, die dem ununterbrochen stimmhaften Segment einer Silbe entsprechen, d.h. über einer Zeitspanne von typischerweise mehr als 100 ms. Gemäß einer vergleichsweise strengen phonologischen Regel enthält jedes Sprachsegment mit ununterbrochener Phonation in aller Regel jeweils einen Vokalkern. Letzterer kann dazu benutzt werden, die (aufgewickelte) fundamentale Phase in Bezug auf das Schließereignis der Glottis zu eichen und somit die Mehrdeutigkeit des Anfangswertes der fundamentalen Phase aufzuheben.

Die beschriebene Synchronisationsannahme erweist sich nicht nur im Fall der Vokale als nützlich sondern insbesondere auch im Fall anderer stimmhafter Kontinuanten, für die die einfache Interpretation des Quelle-Filter Modells in Form einer ebenen Welle in einem unverzweigten Vokaltrakt seine Gültigkeit verliert und die Nichtlinearität der Strömungsdynamik verstärkt in Erscheinung tritt. Die stimmhaften Konsonanten (Frikative) weisen zum Teil eine zweite Schallquelle in der Nähe einer zweiten Verengungsstelle des Vokaltraktes auf, die einen intermittierend turbulenten Luftstrom erzeugt. Die Umwandlung der kinetischen Energie dieses Luftstromes in akustische Energie erfolgt hierbei (z.B. an der Oberkante der unteren Zähne) mit einer für den jeweiligen Wirkmechanismus charakteristischen Wirkungsverzögerungszeit, die sich aus der im Vergleich zur Schallgeschwindigkeit stark verminderten Konvektionsgeschwindigkeit des jeweils relevanten Luftstroms ergibt [5]. Die Phonemspezifischen Zeit- bzw. Phasenverschiebungen, relativ zur fundamentalen Phase können als weitere Beispiele topologischer Invarianten angesehen werden, die vergleichsweise wenig auf Störungen des Kommunikationskanals reagieren.

Es wird daher vorgeschlagen, die automatische Spracherkennung um ein Phasenmodulationsprotokoll mit mehrfachen, teilweise synchronisierten Träger und/oder Envelope Phasen zu erweitern, wobei die Phasenbeziehungen zwischen den Unterbändern mit einem gemeinsamen Treiber kompatibel sind. Stimmhaftigkeit wird erkannt als ein Satz von Phasenbeziehungen, der hinreichend groß ist, um einen instationären fundamentalen Treiber zu rekonstruieren und seine topologische Äquivalenz mit einem glottalen Masteroszillator auf der Senderseite zu bestätigen. Die Rekonstruktion der Phase des fundamentalen Treibers erfordert eine extrem genaue Rekonstruktion der Phasengeschwindigkeit des glottalen Masteroszillators. Die menschliche Tonhöhenwahrnehmung spielt daher im hypothetischen Phasendecoder der Stimmen Wahrnehmung eine zentrale Rolle. In einer begleitenden Studie [9] wird eine selbstkonsistente Rekonstruktion der Phasengeschwindigkeit des fundamentalen Treibers beschrieben, die sich auf eine an das jeweilige Sprachsignal optimal angepasste Zerlegung in Zeit – Frequenz Atome stützt und Eigenschaften der sog. virtuellen Tonhöhe aufweist.

References:

[1] Lippmann R., „Speech recognition by machines and humans“, *Speech Communication* **22**, 1-15 (1997)
[2] Gold B. and N. Morgan, *Speech and audio signal processing*, John Wiley & Sons (2000)
[3] Vary P., U. Heute, W. Hess, *Digitale Sprachsignalverarbeitung*, B.G. Teubner Verlag, Stuttgart (1998)
[4] Kantz H., T. Schreiber, *Nonlinear time series analysis*, Cambridge Univ. Press (1997)
[5] Jackson P.J.B. and C.H. Shadle, *IEEE trans. speech audio process.,* vol. **9**, pp. 713-726 (2001)
[6] Drepper F.R., „Selfconsistent time scale separation of instationary speech signals“, *Fortschritte der Akustik-DAGA'05* (2005)
[7] Drepper F.R., "Voiced excitation as entrained response of a reconstructed glottal oscillator", *Interspeech 2005,* Lisboa (2005)
[8] Drepper F.R., "A two-level drive-response model of non-stationary speech signals", in M. Faundez-Zanuy et al. (Eds), NOLISP 2005, *LNAI* 3817, 125-138, Springer (2005)
[9] Drepper F.R., „Stimmhafte Sprache als sekundäre Antwort eines selbst-konsistenten Treiberprozesses “, *Fortschritte der Akustik-DAGA'06* (2006)
[10] Afraimovich V.S., N.N. Verichev, M.I. Rabinovich, "Stochastic synchronization of oscillation in dissipative systems" *Radiophys. Quantum Electron*. **29**, 795 ff (1986)
[11] Rulkov N.F. , M.M. Sushchik, L.S. Tsimring, H.D.I. Abarbanel, "Generalized synchronization of chaos in directionally coupled systems", *Phys. Rev. E* **51**, 980-994 (1995)
[12] Schoentgen J., "Stochastic models of jitter", *J. Acoust. Soc. Am.* **109** (4): 1631-1650 (2001)
[13] Hanquinet J., F. Grenez and J. Schoentgen, "Synthesis of disordered speech", *Interspeech 2005,* Lisboa (2005)
[14] Moore B.C.J., *An introduction to the psychology of hearing*, Academic Press (1989)
[15] Hohmann V., "Frequency analysis and synthesis using a Gammatone filterbank" *Acta Acustica* **10**, 433-442 (2002)