Stimmhafte Sprache als sekundäre Antwort eines selbst-konsistenten Treiberprozesses

F.R. Drepper

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

Das stimmhafte Sprachsignal wird als stochastische Antwort eines verborgenen fundamentalen Treibers dargestellt, der selbstkonsistent und eindeutig anhand des Sprachsignals rekonstruiert wird. Wie in einer begleitenden Studie erläutert wird [1], wird die synchronisierte primäre Antwort (stimmhafte Quelle) durch eine periodische Kopplungsfunktion des fundamentalen Treibers dargestellt. Das zweistufige Treiber-Response Modell stimmhafter Sprachsignale zeichnet sich im Vergleich zum einstufigen Quelle-Filter Modell durch einen vergrößerten Anwendungsbereich in der Sprachanalyse und Synthese aus.

Ein schwieriges Problem der Analyse und Synthese stimmhafter Sprache besteht in der Zeitskalentrennung zwischen der phonetisch relevanten schnellen Dynamik und niederfrequenten Phänomenen wie Mikrotremor und Prosodie [2-3, 12]. Üblicherweise wird hierfür eine breite Frequenzlücke unterhalb von ca. 100 Hz angenommen. In der begleitenden Studie [1] wird ein neuartiger Zugang zur Zeitskalentrennung stimmhafter Kontinuanten beschrieben, der darauf beruht, dass die phase locking Phänomene der aero-akustischen Anregungsdynamik im Kehlkopf bzw. Vokaltrakt des Sprechers als verallgemeinerte Synchronisation in einem Treiber – Response System beschrieben werden können. Im Fall nichtpathologischer stimmhafter Sprache wird außerdem davon ausgegangen, dass der potentiell instationäre gemeinsame Treiberprozess (glottale Masteroszillator) nur *einen* unabhängigen Freiheitsgrad aufweist.

Auf der Empfängerseite der Sprachkommunikation werden den jeweiligen akustischen Moden entsprechende Zeit – Frequenz Atome bzw. Unterbänder rekonstruiert, die geeignet sind, die charakteristische Synchronisation auf der Senderseite auch im Fall instationärer Tonhöhenschwankungen wiederzuspiegeln. Von besonderer Bedeutung ist hierbei die Rekonstruktion des fundamentalen Treiberoszillators, der das topologisch äquivalente Abbild des glottalen Masteroszillators darstellt. Da optimale Zeit – Frequenz Atome als Ergebnis einer logarithmischen Entwicklung bis zur zweiten Ordnung aufgefasst werden können, zeichnen sich optimal an instationäre akustische Objekte angepasste Zeit – Frequenz Atome durch eine quadratische Filterphase bzw. einen linearen Trend der Filterphasengeschwindigkeit (Filtermittenfrequenz) aus. Im Zentrum der vorliegenden Darstellung steht die selbst-konsistente Anpassung des Trends der Filtermittenfrequenzen der Teilbandzerlegung an die zeitliche Entwicklung der Phasengeschwindigkeit des fundamentalen Treiberprozesses.

Im Hinblick auf real time Anwendung wird eine rein autoregressive Approximation komplexer Gammatone Bandpass Filter 5. Ordnung [15] verwendet. Die äquivalente Rechteck Breite (ERB) der Teilbänder mit den Indizes $j=1,\ldots,N$ wird in grober Annäherung an die psycho-akustisch bestimmten Bandbreiten bestimmt [14, 15]. Die charakteristische Autosynchronisation der stimmhaften Sprache legt es nahe, die Filtermittenfrequenzen der relevanten Teilbänder als ganzzahlige (harmonische) Vielfache h_j einer fundamentalen Filtermittenfrequenz anzunehmen. Im Bereich der separablen Teilbänder $1 \le j \le 8$ ist die harmonische Ordnungszahl h_j identisch mit dem Teilbandindex j. Um eine substantielle Übervollständigkeit der Zerlegung im nicht-separablen Bereich $8 < j \le N$ zu vermeiden, wird der Satz von harmonischen Teilbändern entsprechend den jeweiligen ERBs ausgedünnt. Die autoregressive Approximation der Gammatone Filter ermöglicht eine einfache Realisierung der Zeitabhängigkeit der Filterkoeffizienten sowie eine analytische Lösung der Impulsantwort. Hiermit ergeben sich folgende komplexwertige Teilbänder,

$$z_{j,t} = \sum_{t'=0}^{t} \exp\left(i h_j \sum_{k=t'+1}^{t} \omega_k\right) \lambda_j^{t-t'} \frac{(t-t'+\Gamma-1)!}{(\Gamma-1)!\,(t-t')!} s_{t'} \quad . \quad (1)$$

mit den Teilband abhängigen Dämpfungsfaktoren λ_j und der zeitabhängigen Filter-Phasen Geschwindigkeit $\omega_t = 2\pi F_{1,t}$, welche eng mit der fundamentalen Filtermittenfrequenz $F_{1,t}$ des ersten Teilbands verknüpft ist. Gleichung (1) kann als eine hochgradig übersamplete Zeit – Frequenz Zerlegung des Eingangssignals $s_{t'}$ interpretiert werden, wobei das überkritische sampling auf die Zeitachse beschränkt bleibt. Die Amplitudenverteilung der Impulsantwort hat angenähert die Gestalt einer Γ-Funktion, welche dafür bekannt ist, bei höherer Ordnungszahl Γ eine Gaussglocke zu approximieren. Um bei negativen Chirpraten eine Singularität der Periodenlänge der Impulsantwort zu vermeiden, wird die Zeitabhängigkeit der Filter-Phasen Geschwindigkeit ω_t in Abhängigkeit vom Vorzeichen der Chirprate c gewählt,

$$\omega_k = \begin{Bmatrix} \omega_0 (1+c\,k) \\ \omega_0 / (1-c\,k) \end{Bmatrix} \quad for \quad \begin{Bmatrix} c \ge 0 \\ c < 0 \end{Bmatrix} . \quad (2)$$

Die selbstkonsistente Anpassung der Filtermittenfrequenzen wird auf zweierlei Typen von Teilbandphasen gestützt. Im Bereich der separablen Teilbänder (j = 1, 2, …, 8) wird die Trägerphase $\varphi_{j,t} = \arctan(im(z_{j,t})/re(z_{j,t}))$ bzw. die normierte Trägerphase $\varphi_{j,t}/h_j$ benutzt. Im Bereich der nicht-separablen Teilbänder wird die Hilbertphase der Modulationsamplitude (Envelope) benutzt ($h_j = 1$). Im Gegensatz zu den Trägerphasen benötigen die Envelopephasen eine Korrektur der Gruppenverzögerung. Der Teilband abhängige Teil dieser Korrektur wird wie in [15] gewählt.

Rekonstruktion der Fundamentalen Phase

Der beschriebene Satz von Bandpassfiltern wird auf eine chirpende Folge von synthetischen glottalen Pulsen in Form von Sägezähnen angewandt, die in grober Annäherung ein typisches glottales Anregungsspektrum aufweist [3]. D.h. wir wählen

$$S_{t'} = \min(t'-T_m^*,\, s\,(T_{m+1}^*-t')) \quad , \quad (3)$$

wobei T_m^* und T_{m+1}^* die direkten Nachbarn von t' in der Folge

$$T_m = \sum_{m'=1}^{m} \tau_{m'} \qquad (m = 1,\ldots,\infty) \quad (4)$$

darstellen, und wobei die Periodenlängen $\tau_{m'}$ implizit mittels der zeitabhängigen Phasengeschwindigkeit $\omega'(t')$ definiert sind, die analog zu Gleichung (2) gewählt wird, jedoch mit potentiell unterschiedlicher Chirprate c' und Anfangsphasengeschwindigkeit ω_0'. Der Parameter s bestimmt die negative Steigung des Sägezahns. Durch Integration von $\omega'(t')$ über eine volle Periode der zugehörigen Phase wird die Periodenlänge τ_m in expliziter Form erhalten,

$$\tau_m = \begin{Bmatrix} \frac{1+c'm}{c'} \left(\sqrt{1 + \frac{4\pi c' \omega_0'}{(\omega_0'(1+c'm)^2}} - 1 \right) \\ \frac{1-c'm}{c'} \left(1 - \exp(-2\pi c'/\omega_0') \right) \end{Bmatrix} \quad for \quad \begin{Bmatrix} c \ge 0 \\ c < 0 \end{Bmatrix} .$$

In der Situation der Signalanalyse stützt sich die selbstkonsistente Bestimmung der Filtermittenfrequenzen der Band-

passfilter insbesondere auf die Schätzung des Trends der Phasengeschwindigkeit der durch die Bandpassfilter erzeugten Teilbänder. Die Selbstkonsistenz besteht hierbei darin, dass die Schätzwerte für die Verbesserung des Trends der Filtermittenfrequenzen der besagten Bandpassfilter benutzt werden. Zur Reduzierung der Abhängigkeit der Schätzwerte von Größe und Lage des Analysefensters wird ein Zeitskalen Trennungsansatz benutzt, der die durch die Synchronisation hervorgerufenen regelhaften Schwankungen der Phasengeschwindigkeit mittels einer 2π periodischen Funktion $P_j(\varphi)$ berücksichtigt,

$$\dot{\varphi}_{j,t} / h_j \;=\; \alpha_j\, t \;+\; P_j(\varphi_{j,t} / h_j). \qquad (5)$$

$P_j(\varphi)$ kann durch eine Fourier Reihe niedriger Ordnung approximiert werden. Sowohl die Fourierkoeffizienten als auch die gesuchte Steigung α_j der Phasengeschwindigkeit werden mittels multipler linearer Regression bestimmt. Im Vorfeld der Rekonstruktion der fundamentalen Treiberphase interessiert die Frage, welches Teilband für die Anpassung seiner eigenen Filtermittenfrequenz am besten geeignet ist. Es zeigt sich, dass diese Frage graphisch beantwortet werden kann, wenn wir uns auf die Anpassung des Filter Chirp Parameters c beschränken. Für die graphische Nachbildung der Erzielung von Selbstkonsistenz benutzen wir einen Graphen, der für mehrere Teilbänder jeweils den Trend α_j der normierten Teilband Phasengeschwindigkeit $\dot{\varphi}_{j,t} / h_j$ als Funktion der fundamentalen Filter-Chirprate c darstellt. Bild 1 zeigt die Schätzwerte der relativen Steigung $\alpha_j /(\omega_0' c')$ mehrerer Teilband Phasengeschwindigkeiten (j = 10, 2, 4, 6, von unten nach oben) als Funktion der relativen fundamentalen Filterchirp rate c/c', beide relativ zur vorgegebenen Chirprate des Eingangssignals $\omega_0' c'$.

Mittels zweier graphischer Schritte kann eine iterative Anpassung der Filtermittenfrequenzen durchgeführt werden, die bei geeigneten Startwerten der Filtermittenfrequenz zu einer, je nach Teilband, mehr oder weniger guten Anpassung im stabilen Fixpunkt der Iteration führt. Aus der graphischen Analyse ergeben sich folgende Ergebnisse: Trägerphasen sind besser geeignet, eine präzise Anpassung der Filtermittenfrequenz zu liefern als Envelopephasen. Trägerphasen der höher harmonischen Teilbänder liefern genauere Fixpunkte als die der nieder harmonischen. Konvergenzbereich und Konvergenzgeschwindigkeit der iterativen Verbesserung sind jedoch bei den nieder harmonischen

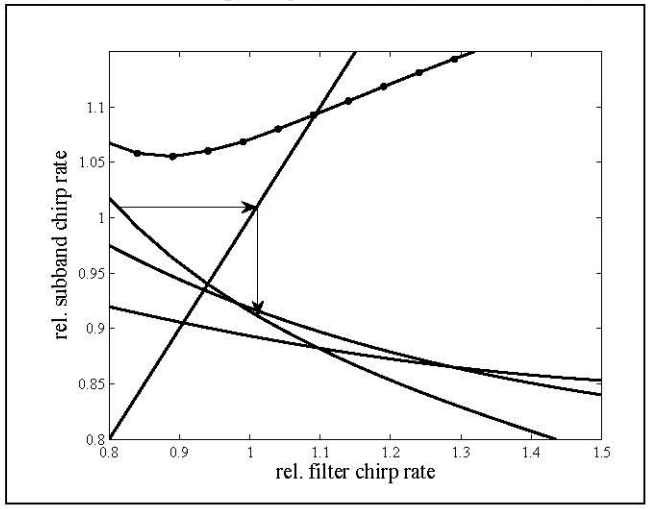

Bild 1: Relative Teilband Chirpraten als Funktion der Chirprate der Filtermittenfrequenz für die drei Trägerphasen der Teilbänder 2, 4, und 6 (von unten nach oben) und für die Envelopephase des 10. Teilbandes (oben, Kreise). Die Pfeile und die Winkelhalbierende des ersten Quadranten erläutern den Algorithmus zur Bestimmung selbst-konsistenter Filtermittenfrequenzen.

Die Fixpunkte zeigen den asymptotischen Fehler der Filterchirprate an, der nicht nur vom Teilbandindex sondern auch von der Länge des Analysefensters abhängt (welche im Fall von Bild 1 etwa 5 Oszillationen bzw. einem sechstel der Verdopplungszeit der Frequenz entspricht). Die Kleinstquadrate Schätzung von Gleichung (5) führt zu einer systematischen Unterschätzung des Betrags der Steigung α_j.

Ein gleichzeitig robuster, schneller *und* präziser Algorithmus zur selbst-konsistenten Bestimmung der fundamentalen Phasengeschwindigkeit wird erreicht, wenn man zunächst Envelopephasen und/oder eine Trägerphase niedriger Ordnung dazu benutzt, um separable Teilbänder höherer Ordnung zu finden, die (n:m) phasensynchron zu Trägern anderer separabler Teilbänder sind, und vorzugsweise die Trägerphase der höchsten Ordnung dazu benutzt, um die fundamentale Treiberphase und die zugehörigen Filtermittenfrequenzen zu bestimmen. Es stellt sich heraus, dass dieses Entwurfsprinzip des Algorithmus geeignet ist, wohlbekannte Eigenschaften der Tonhöhenwahrnehmung zu erklären, einschließlich der Bevorzugung der harmonischen Teilbänder mit den Ordnungszahlen 4-6 [14]. Im Fall der stimmhaften Frikative wird die höchste Harmonische mit einer strengen Phasensynchronisation nicht nur durch die Bandbreite der Teilbandfilter bestimmt sondern auch durch charakteristische Eigenschaften des Sprachsignals.

SCHLUSSFOLGERUNG: Ein Übertragungsprotokoll stimmhafter Sprache wurde beschrieben, das auf der charakteristischen Auto-Phasen-Synchronisation der Teilbänder stimmhafter Kontinuanten beruht. Die Auto-Phasen-Synchronisation instationär stimmhafter Sprache kann nur mittels einer an das Sprachsignal angepassten Teilbandzerlegung im vollen Umfang decodiert werden. Die selbst-konsistent angepasste Teilbandzerlegung *beruht* auf der präzisen Rekonstruktion einer eindeutigen fundamentalen Phasengeschwindigkeit. Gleichzeitig *unterstützt* die selbst-konsistente Teilbandzerlegung die präzise Rekonstruktion eines fundamentalen Treibers sowie dessen Bestätigung als topologisch äquivalentes Abbild eines glottalen Masteroszillators. Die Verteilung der Phasensynchronisation und deren Periodizitäten auf die unterschiedlichen Teilbänder sowie charakteristische Zeit- bzw. Phasenverschiebungen, relativ zur fundamentalen Phase, stellen (zusätzliche) topologische Invarianten der phonetisch relevanten Dynamik dar, die durch Veränderungen des Kommunikationskanals minimal gestört werden. Die Frage, in wieweit dieses Übertragungsprotokoll bei der Erkennung von Phonemen und Sprechern tatsächlich benutzt wird, muss zum jetzigen Zeitpunkt noch weitgehend offen bleiben. Das erstaunlich robuste Übertragungsprotokoll kann jedoch als Resultat einer kombiniert phylogenetischen und ontogenetischen Ko-evolution des Stimmerzeugungssystems und des Hörpfades in einer hochgradig variablen akustischen Umwelt entstanden sein, die durch Lautäußerungen der jeweiligen Zeitgenossen der eigenen Spezies stark beeinflusst wurde. Die efferente Innervation der äußeren Haarzellen der Hörschnecke zeigt, dass es zumindest keinen unmittelbar offensichtlichen Widerspruch seitens der Hörphysiologie gibt.

Der Author dankt V. Hohmann, B. Kollmeier, Oldenburg, M. Kob, C. Neuschaefer-Rube, Aachen, N. Stollenwerk, Lisboa, J. Schoentgen, Brussels, P. Grassberger, M. Schiek and P. Tass, Jülich für hilfreiche Diskussionen.

References siehe [1]

[1] Drepper F.R., „Stimmhafte Anregung als Antwort eines eindeutigen fundamentalen Treibers “, *Fortschritte der Akustik-DAGA'06* (2006)