Synthetische Vokale als Summe von modulierten Sinustönen

S. Uppenkamp, A. Kothari, J. Bailes, R.D. Patterson

Centre for the Neural Basis of Hearing, Department of Physiology, University of Cambridge, Downing Street, Cambridge, CB2 3EG, U.K. E-mail: stefan.uppenkamp@mrc-cbu.cam.ac.uk

Einleitung

Ziel der Arbeiten ist, mit funktioneller MR Tomographie durch den Kontrast der Aktivierung durch Vokale und geeignete physikalisch ähnliche Kontrollstimuli (Nicht-Vokale) den Ort einer sprachlaut-spezifischen Verarbeitung im menschlichen Gehirn zu finden. Die Produktion von Vokalen kann als zweistufiger Prozess beschrieben werden. Der Luftstrom aus den Lungen regt die Glottis zu periodischen Schwingungen an. Damit wird das Eingangssignal für den Vokaltrakt gebildet, eine Pulsfolge mit einer Rate entsprechend der Tonhöhe der Stimme. Auf dem Weg durch den Vokaltrakt wird dem Eingangssignal durch die veränderlichen Resonanzen das jeweils charakteristische Formantspektrum aufgeprägt. Vokale lassen sich nach genau diesem Prinzip synthetisieren [1]. Ein anderer Ansatz ist es, Sinustöne bei den Formantfrequenzen zu addieren. So synthetisierte Vokale klingen allerdings trotz vollständiger Formantinformation nicht wie Sprache. Prägt man den Sinustönen jedoch eine sich periodisch wiederholende exponentiell abklingende Einhüllende auf, lassen sich sprachähnliche Signale synthetisieren, die eindeutig und zuverlässig als Vokale wahrgenommen werden. Dieser Ansatz erlaubt es, durch einfache Manipulation von wenigen Parametern akustisch ähnliche, nicht-sprachliche Kontrollstimuli zu generieren.

Generierung der synthetischen Vokale

Die Bausteine unseres Ansatzes der Vokalsynthese sind “damped sinusoids” [2], Sinustöne bei den Formantfrequenzen mit einer sich periodisch wiederholenden, exponentiell abklingenden Einhüllenden. Eine Abklingzeit (Halbwertzeit) der Exponentialfunktion von 4 ms ergibt eine realistische Bandbreite für die einzelnen Formanten. Die Wiederholrate der Einhüllenden wird aus dem Bereich der Tonhöhen von menschlichen Stimmen gewählt. Abbildung 1 zeigt ein Beispiel für einen auf diese Weise synthetisierten Vokal /a/. Derartige Vokale klingen zwar etwas künstlich, sie werden aber eindeutig als Sprachlaute wahrgenommen, wenn die Länge in etwa einer typischen Silbenlänge (300-400ms) entspricht. Die anderen Beispiele in Abb. 1 zeigen, wie sich aus derartigen Vokalen durch Randomisierung der beiden Parameter Trägerfrequenz und Einhüllenden-Onset über der Zeit auf einfache Weise akustisch – im Sinne des Langzeitspektrums – ähnliche Kontrollstimuli generieren lassen, die entweder sprachähnlich oder aber gar nicht mehr wie Sprache klingen.

Experiment 1: "Damped/ramped"-Diskrimination

Für drei verschiedene Abklingzeiten wurden psychometrische Funktionen für die Diskrimination von "damped" und "ramped" Signalen mit einem, zwei und vier Formanten in Abhängigkeit von der Einhüllendenperiode ermittelt. Signale mit zwei Formanten waren in der Regel leichter zu diskriminieren als einzelne modulierte Sinustöne. Bei weiterer Erhöhung der Zahl der Formanten zeigte sich jedoch eine Verschlechterung der Diskriminationsleistung (s. Abbildung 2). Eine mögliche Interpretation dieses Ergebnis ist, daß es für das Gehör einen sprachlautspezifischen Verarbeitungsmodus gibt. Wenn die Signale eindeutig als Vokale gehört werden, ist die Leistung in einer nichtsprachlichen Aufgabe offenbar nicht optimal.

Experiment 2: Vokal-Identifikation

Zur Evaluierung der synthetischen Vokale wurde ein Identifikationsexperiment durchgeführt. Den Versuchspersonen wurden aus insgesamt 16 Schallbedingungen zufällig ausgewählte Vokale vorgespielt. Die Aufgabe war, jeweils zu entscheiden ob es /a/, /e/, /i/, /o/ oder /u/ war. Die verwendeten Bedingungen waren die Schalle 1-12 aus Tabelle 1, sowie jeweils eine Version mit exponentiell ansteigender Einhüllender ("ramped") der ersten vier Schalle aus der Tabelle. Die resultierenden Verwechslungsmatrizen zeigten, daß sowohl "ramped" als auch "damped" Signale mit zwei und vier Formanten eine zuverlässige Identifikation erlauben. Die Identifikationsleistung für Vokale aus reinen Sinustönen bei den Formantfrequenzen ist im Gegensatz dazu signifkant schlechter.

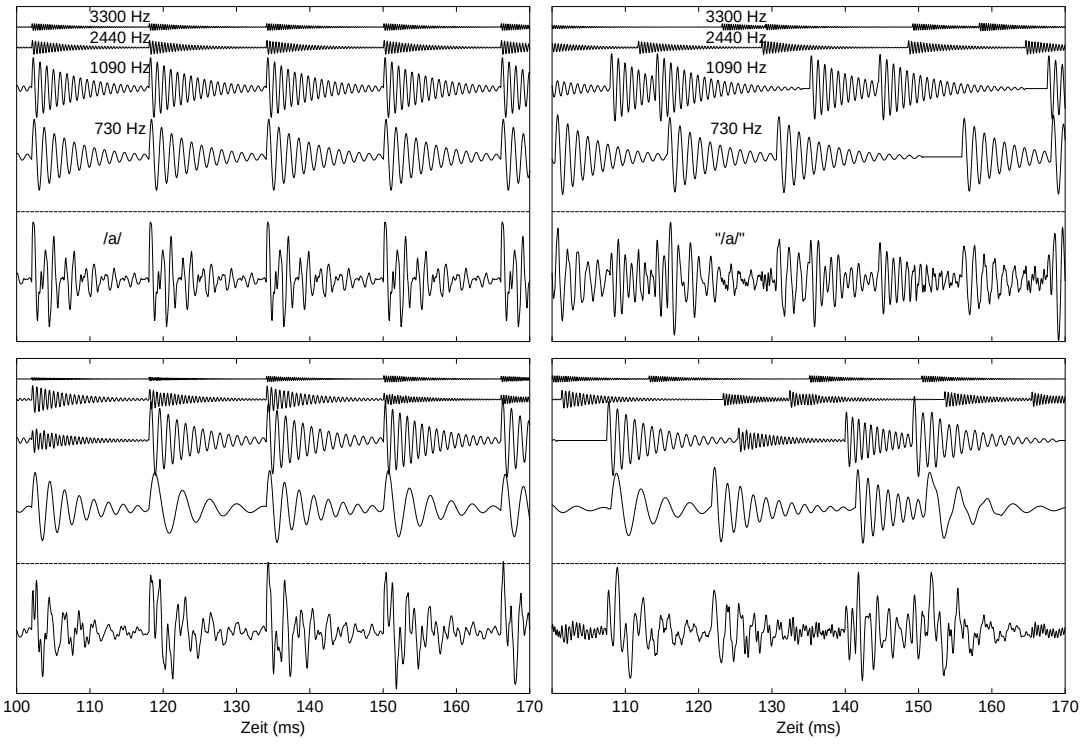

Abbildung 1: *Prinzip der Synthese von Vokalen und eng verwandten Kontrollstimuli durch Addition von modulierten Sinustönen mit periodisch wiederholter exponentiell abklingender Einhüllender ("damped sinusoids"). Oben links: Vokal* /a/*, feste Formantfrequenzen und periodische Einhüllende; oben rechts: "pathologischer" Vokal, feste Formantfrequenzen, randomisierte Einhüllenden-Onsets; unten links: periodische Einhüllende, randomisierte Trägerfrequenzen von Zyklus zu Zyklus; unten rechts: Randomisierung von Einhüllenden-Onsets und Trägerfrequenzen ("musical rain").*

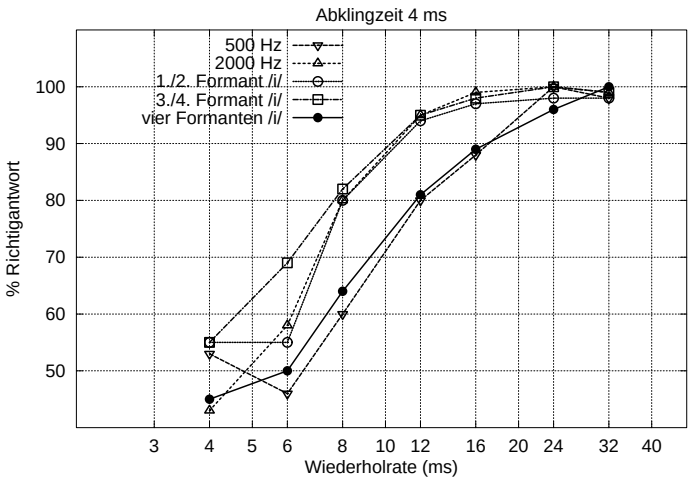

Abbildung 2: *Psychometrische Funktionen für die Diskrimination von "damped" und "ramped" Signalen mit einem, zwei und vier Formanten als Funktion der Periode der Einhüllenden (entsprechend der empfundenen Tonhöhe der Signale); gemittelte Ergebnisse von 5 Versuchspersonen. Die Abklingzeit (Halbwertzeit) für die exponentielle Einhüllende betrug 4 ms*

dmp_vow	* damped vowels, four tracks of damped sinusoids at formant frequencies
dmp_two	as dmp_vow, but only first and second formant
dmp_fst	as dmp_vow, but first formant only
dmp_snd	as dmp_vow, but second formant only
flt_vow	as dmp_vow, but no lowpass slope in spectrum
jit_vow	as dmp_vow, 10% jitter in envelope timing
pth_vow	as dmp_vow, irregular envelopes (100% jitter in timing), i.e. no pitch
noi_vow	* as dmp_vow, but narrow bands of noise as carriers of triangles
sin_vow	four sinusoids at formant frequencies, no damped envelope
sin_two	as sin_vow, but only first and second formant
sin_fst	as sin_vow, but first formant only
sin_snd	as sin_vow, but second formant only
noi_pit	four tracks of damped noise bands, one octave wide
noi_ran	as above, but irregular envelope (no pitch)
fxr_pit	four tracks of damped sinusoids, random change of carrier frequencies within limited bandwidth, regular timing
fxr_ran	as above, but random timing (no pitch)
mus_pit	* complete randomisation of carrier frequencies for each track, regular timing
mus_ran	* randomisation of carrier frequencies and timing

Tabelle 1: *Schallbedingungen in den Experimenten 2 (Vokal-identifikation) und 3 (Paarvergleiche der Sprachqualität). Die Bedingungen aus Abb. 1 sind durch * gekennzeichnet.*

Experiment 3: Qualität der synthetisierten Vokale

Die Sprachähnlichkeit der synthetisierten Vokale und Nicht-Vokale wurde in einem Paarvergleichsexperiment quantifiziert. Die beiden Stimulusintervalle in jedem Paar enthielten jeweils drei zufällig ausgewählte Vokale aus einer Bedingung. Insgesamt wurden 18 Schallbedingungen miteinander verglichen (siehe Tabelle 1). Aufgabe der Versuchsperson war es, jeweils das sprachähnlichere Signal aus einem Paar auszuwählen. Jeder Vergleich wurde zweimal durchgeführt, einmal als A-B und einmal als B-A. Damit hat jede Versuchsperson 306 Urteile abgegeben. Es nahmen 15 Versuchspersonen teil. Mit der Bradley-Terry-Luce Methode [3] wurde aus den insgesamt 4590 Urteilen eine relative Skala der Sprachähnlichkeit der 18 untersuchten Bedingungen gewonnen (siehe Abb. 3).

Wie erwartet sind reguläre Formantfrequenzen eine notwendige Voraussetzung für Sprachähnlichkeit. Der zweite wesentliche Faktor ist das Aufprägen der exponentiell abklingenden Einhüllenden, entweder periodisch oder aperiodisch, was zu einer natürlichen Bandbreite der einzelnen Formanten führt. Der dritte Faktor ist die Regularisierung der Einhüllenden-Onsets, die den Signalen eine Tonhöhe aufprägt. Die Signale am unteren Ende der Skala sind sehr verschieden von menschlichen Vokalen. Die vollständige Randomisierung der Einhüllenden-Onsets und der Trägerfrequenzen innerhalb der einzelnen Perioden der Einhüllenden führt zu einem nicht-periodischen Signal, das überhaupt nicht mehr wie Sprache klingt, sondern eher als “musical rain” beschrieben werden kann.

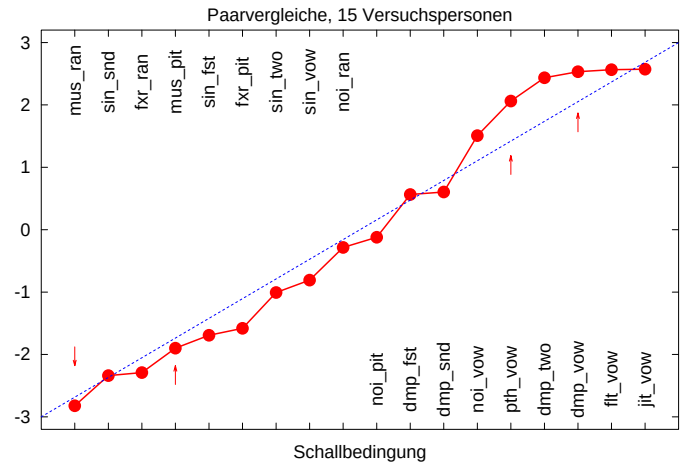

Abbildung 3: *Relative Skala der Sprachähnlichkeit für synthetische Vokale und Nicht-Vokale. Die Skala wurde aus Paarvergleichsurteilen von 15 Versuchspersonen gewonnen. Die Bedingungen aus Abb 1 sind durch Pfeile gekennzeichnet.*

Funktionelle MR Tomographie mit synthetischen Vokalen

In einem Pilotexperiment in einem 3-Tesla-fMRT Scanner wurde mit einer Versuchsperson geprüft, wieviele Bilder notwendig sind, um einerseits eine zuverlässige Aktivierung von auditorischen Strukturen im Gehirn zu sehen, und um andererseits auch den Kontrast zwischen zwei verschiedenen Schallbedingungen mit ausreichendem Signal/Rausch-verhältnis darzustellen. Dabei wird angenommen, daß Regionen in denen Sprachlaute signifikant mehr Aktivierung hervorufen als nichtsprachliche Kontroll-Stimuli Kandidaten für eine sprachlaut-spezifische Verarbeitung im Gehirn sind.

Für die akustische Stimulation wurde ein Magnetfeld-taugliches elektrostatisches Kopfhörersystem des *MRC Institute of Hearing Research* in Nottingham verwendet [4]. Zusätzlich wurde eine spezielle Technik – "sparse temporal sampling" – eingesetzt, um Scanner-Lärm und akustischen Stimulus zeitlich voneinander zu trennen [5]. Die Stimuli aus den beiden in diesem Pilotexperiment verwendeten Bedingungen ("damped vowels" und "musical rain") ergaben im Kontrast mit Ruhe als Kontrollbedingung ein sehr ähnliches Aktivierungsmuster in beiden Schläfenlappen. Neben dem primären Hörkortex fand sich eine starke Aktivierung von akustischen Assoziationsfeldern und des Wernicke-Areals. Der Kontrast zwischen den beiden Schallbedingungen zeigte dagegen eine deutlich zur linken Hemisphäre hin lateralisierte Aktivierung durch die Sprachlaute, unterhalb und posterior zum primären auditorischen Kortex.

Diskussion

Die Synthese von Vokalen mit "damped sinusoids" bei Formant-frequenzen ermöglicht die Generierung einer Klasse von Stimuli mit ähnlichem Langzeitsspektrum, die in ihrer Klangqualität einen großen Bereich von "sehr sprachähnlich" bis zu "überhaupt nicht wie Sprache" abdecken. Wir nehmen an, daß solche Signale primäre Zentren der Schallverarbeitung im Gehirn in sehr ähnlicher Weise aktivieren werden. Zentren, die mit der Verarbeitung von sprachlaut-spezifischen Merk-malen im Schall befaßt sind, sollten dagegen einen starken Kontrast in der Aktivierung durch Sprache und Nicht-Sprache zeigen. Die im Pilot-experiment gefundene Region als Kandidat für die angenommene sprachlaut-spezifische Verarbeitung in der linken Hemisphäre soll in Zukunft in einer umfangreicheren Studie mit weiteren Kontroll-bedingungen und mehr Versuchspersonen weiter untersucht werden.

Die Arbeiten wurden vom UK Medical Research Council unterstützt. Die fMRT Experimente wurden am Wolfson Brain Imaging Centre in Cambridge durchgeführt. Für die Analyse dieser Daten wurde SPM99 (Wellcome Department of Cognitive Neurology, London) verwendet.

Literatur

1. Klatt HD (1980) *J Acoust Soc Am* **67**, 971-995.
2. Patterson RD (1994) *J Acoust Soc Am* **96**, 1419-1428.
3. David HA. *The method of paired comparisons*. 2nd ed., Oxford University Press, N.Y. 1988
4. Palmer AR, Bullock DC, Chambers JD (1998) *Neuroimage* **7**, S359.
5. Hall DA, et al. (1999) *Hum. Brain Mapping* **7**, 213-223.