Modulationsfilterung von Sprache mit Fourier-Spektrogramm und Wavelet-Transformation

Olaf Schreiner und Hans Werner Strube
Drittes Physikalisches Institut, Universität Göttingen

Einleitung

Sprache von Störgeräuschen zu trennen, ist eigentlich ein sogenanntes *ill-posed problem*, d. h. ein Problem, für das es keine unmittelbare analytische Lösung gibt, da man weder das eine noch das andere Signal kennt, sondern nur das Signalgemisch. Dennoch ist das gesunde menschliche Gehör in der Lage, eben diese Trennung bis zu einem gewissen Grad durchzuführen.

Ein Ansatz, diese Fähigkeit des menschlichen Gehörs technisch nachzubilden, ist die Methode der Hervorhebung von modulierten Anteilen im Signal [6]. Wilmers zeigte, dass die Erkennungsraten eines automatischen Spracherkenners für einen begrenzten Wortschatz (die Ziffern Null bis Neun) mit diesem Ansatz deutlich erhöht werden können. Geurts und Wouters [2] zeigten, dass auch die Verständlichkeit von Sprache für Gehörgeschädigte, die eine Innenohr-Prothese (Cochlea-Implantat) tragen, deutlich verbessert werden kann.

Dieser Arbeit liegen nun zwei Fragestellungen zugrunde: Erhöht die Modulationsfilterung auch die Verständlichkeit von Sprache für Normalhörende? Und: Lässt sich der Ansatz von Wilmers durch Berücksichtigung auditorischer Merkmale des menschlichen Gehörs verbessern?

Modulation von Sprache

Wesentliche Anteile der menschlichen Sprache, nämlich alle Vokale und alle stimmhaften Konsonanten, weisen eine besondere zeitliche Struktur auf: Sie sind mit der Grundfrequenz F_0 moduliert.

Durch das harte Schließen der Glottisschwingung bei stimmhaften Lauten entstehen Obertöne bei Vielfachen der Grundfrequenz F_0, die mit etwa 12 dB/Oktave abfallen. Die Überlagerung der Grundfrequenz F_0 und ihrer Vielfachen kann auch als Schwebung mit der Differenzfrequenz F_0 interpretiert werden. Man erhält durch die Anwesenheit von Harmonischen also automatisch eine Amplituden-Modulation des Signals mit F_0. Mit der Amplitude $s(t)$ ist natürlich auch die momentane Signalleistung $s(t)^2$ amplitudenmoduliert. Da der Vokaltrakt als Filter betrachtet eine geringe Güte hat, d. h. seine Impulsantwort gut im Zeitbereich lokalisiert ist, bleibt die zeitliche Struktur der Glottisschwingung, die die Amplitudenmodulation beinhaltet, trotz der Filterung durch den Vokaltrakt bei allen Vokalen erhalten. Dies schließlich macht man sich bei der Modulationsfilterung zunutze.

Gleichzeitig weist das menschliche Gehör einen Mechanismus auf, der gerade diese zeitliche Struktur unabhängig vom Spektrum des Lautes abbildet, nämlich auf die verschiedenen Modulationsfrequenzen. Die Wahrnehmung der Tonhöhe komplexer Töne wie Sprachsignale beruht nicht, wie zunächst angenommen, auf der Frequenzzerlegung in der Cochlea, sondern auf einer nachgeschalteten zeitlichen Analyse im Colliculus Inferior im Hirnstamm. Dort wird für jeden Frequenzkanal der Cochlea eine Art Autokorrelations-Analyse erstellt, indem das Signal mit verzögerten Versionen seiner selbst verglichen wird. Dadurch erhält man neben der Frequenzzerlegung in der Cochlea eine zweite Dimension, das Modulationsspektrum [3].

Die Stimme eines Sprechers wird im Modulationsspektrum besonders stark auf die Grundfrequenz F_0 und deren Harmonische abgebildet. Ein anderes Geräusch, das mit einer anderen Frequenz moduliert ist, wird an eine andere Stelle im Modulationsspektrum abgebildet. Insbesondere wenn ein Geräusch überhaupt keine Struktur in der zeitlichen Einhüllenden seiner Energie hat, sollte es auf die Modulationsfrequenz *Null* abgebildet werden. Es liegt also nahe anzunehmen, dass das menschliche Gehör die Abbildung auf verschiedene Orte im Modulationsspektrum nutzt, um Signale zu trennen – etwa einen Sprecher von einem anderen oder einen Sprecher von einem Hintergrundgeräusch.

Dieser Ansatz soll hier technisch nachempfunden werden, um ihn zur Trennung von Sprache und Hintergrundgeräusch zu nutzen. Wenn man also eine Transformation findet, die das Signal ähnlich wie im Gehör abbildet, sollte es möglich sein, den Teil der Abbildung, in dem das Störgeräusch liegt, zu entfernen und den Sprachanteil zu erhalten, so dass nach einer Rücktransformation ins Zeitsignal das Sprachsignal ohne oder zumindest mit einem verringerten Störgeräusch übrig bleibt.

Funktionsweise

Als erstes erfährt das Signal eine Vorverarbeitung, die die Höhen hervorhebt, angelehnt an die Verarbeitung im Mittelohr. Anschließend wird das Signal $s(t)$ mittels einer Filterbank in Frequenzkanäle $s_i(t)$ zerlegt, entsprechend der Verarbeitung in der Cochlea (Abbildung 1).

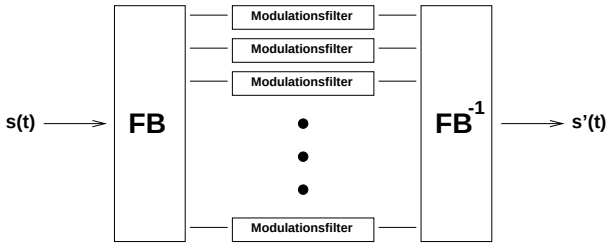

Abbildung 1: Filterbank (FB) zur Modulationsfilterung.

Auf allen Kanälen soll nun einheitlich die Filterung nach der Modulation des Signals stattfinden (Abbildung 2). Dazu wird von jedem Kanal die Hilberteinhüllende gebildet, d. h. der Betrag $|s_i(t)|$ des komplexen Zeitsignals. Aus der Gesamtheit der Einhüllenden aller Kanäle wird dann die Grundfrequenz F_0 bestimmt. Alle Einhüllenden werden dann mit einem schmalen Bandpass-Filter bei F_0 gefiltert.

Da die Einhüllenden (Betragssignale) vorher positiv waren, sollen sie auch hinterher wieder größer als Null sein. Durch den Bandpass wurde jedoch der Gleichanteil entfernt, der anschließend wieder rekonstruiert werden muss. Nachdem der Gleichanteil wieder hinzuaddiert wurde, kann nun zusammen mit der unveränderten Phase $\arg(s_i(t))$, die bei der Betragsbildung zur Seite gelegt wurde, ein komplexes Signal $s_i'(t)$ zurückgewonnen werden. Schließlich wird aus allen Kanälen mit der Filterbankrekonstruktion ein Zeitsignal zurückgewonnen. Am Ende schließt sich noch eine Nachverarbeitung an, die die Vorverarbeitung (s. o.) aufhebt.

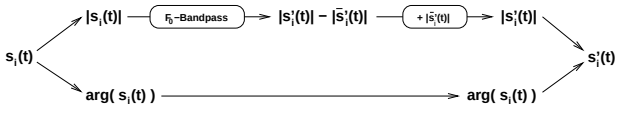

Abbildung 2: Modulationsfilter.

Realisierung

Eine Möglichkeit, die Filterbank zu realisieren, ist die Schnelle Fouriertransformation (FFT) [6]. Zur Rekonstruktion des Gleichanteils addierte Wilmers eine geglättete Version der negativen Halbwelle. Damit konnten bereits Verbesserungen der Erkennungsraten automatischer Spracherkenner erzielt werden. Die lineare Frequenzaufteilung der Fouriertransformation entspricht jedoch nicht der Zerlegung in der Cochlea, die eher logarithmisch verläuft. Ein weiterer Nachteil der Fourier-Filterbank ist, dass im

Bereich der Grundfrequenz das Fenster kürzer ist als die Wellenlänge. Die Grundfrequenz kann daher nicht mehr korrekt aufgelöst werden.

Aus diesen Gründen wird hier die Wavelet-Transformation verwendet, die darüber hinaus den Vorteil hat, dass mit einem Transformationsschritt gleich ein komplettes Spektrogramm, also eine zweidimensionale Zeit-Frequenz-Darstellung erzeugt wird. Verwendet wird hierzu die Schnelle Kontinuierliche Wavelet-Transformation FCWT nach Dress [1], die eine quasikontinuierliche Wavelet-Transformation mit frei wählbarer Zeit- und Frequenzauflösung mit Hilfe von zwei Matrixmultiplikationen realisiert.

Die Höhenanhebung entsprechend der Funktion des Mittelohres wird durch die Gewichtung der Kanäle entsprechend ihrer Mittenfrequenzen bereits durch die Wavelet-Transformation erledigt.

Es wurden hier Spektrogramme der Länge 42 ms mit einem entsprechenden Vorschub von 10,5 ms benutzt. Um Randeffekte beim *Overlap-Add* nach der Rücktransformation zu vermeiden, wurde der Signalausschnitt mit einem Hanning-Fenster versehen. Der Vorschub von 10,5 ms ist etwa um Faktor 3 kleiner als der Spektrogrammvorschub in der Fourier-Methode, was eine bessere Auflösung von Diphthongen (Vokalübergängen) sowie Übergängen von stimmhaften und stimmlosen Bereichen erlaubt.

Als Frequenzaufteilung wurden 64 Kanäle benutzt, von 64 Hz bis 16 kHz. Die Mittenfrequenzen des m-ten Kanals sind

$$f_m = 1024\,\mathrm{Hz} \cdot 16^{m/32-1}.$$

Als auditorisches Wavelet wurde der Gammaton verwendet, der die Impulsantwort der Basilarmembran annähert [4]:

$$g_t(t) = a \cdot t^{n-1} \cdot \exp(-2\pi b \mathrm{ERB}(f_c)\, t) \cdot \cos(2\pi f_c t + \phi)$$

mit $f_c = 1024\,\mathrm{Hz}$, $n = 4$ und $b = 4$. Die Frequenzkanäle wurden jedoch nicht einheitlich abgetastet, wie in der Dress-Methode vorgesehen, sondern jeweils in 8 Kanälen einheitlich. Dazu wurde die Dress-Methode jeweils blockweise auf 8 Kanäle angewendet, sodass die Abtastfrequenz jeweils etwa der doppelten höchsten Mittenfrequenz der 8 Kanäle entspricht. Dies bringt gegenüber einer insgesamt einheitlichen Abtastung der Kanäle einen deutlichen Vorteil in der Verarbeitungsgeschwindigkeit.

Die weitere Verarbeitung geschieht dann entsprechend Abbildung 2: Die einzelnen Kanäle werden in Betrag und Phase getrennt. Die Absolutwerte der Kanäle werden noch einmal in Zeitrichtung in das Modulationsspektrum Fourier-transformiert. Da alle Kanäle, was die absolute Zeit anbetrifft, weiterhin gleich lang sind, haben sie nach dem Sampling-Theorem auch die gleiche Frequenzauflösung. Daher sind die Kanäle im Modulationsspektrum wieder gleichartig abgetastet, unabhängig von der Zeitauflösung. Aus der Gesamtheit aller Kanäle wird die Grundfrequenz bestimmt und der Bereich der Grundfrequenz mit Hilfe eines Hanning-Fensters ausgeschnitten.

Erweiterungen

Im Gegensatz zu Wilmers' Ansatz wird hier zunächst die Anwesenheit bestimmter Laute geprüft und dann diese gezielt erhalten.

Ein robusteres Maß für die Ermittelung der Grundfrequenz lässt sich bilden, indem man zu jeder Modulationsfrequenzkomponente gewichtet ihre Vielfachen hinzuaddiert und davon das Maximum bildet, bei Berücksichtigung einer unteren Modulationsfrequenzschranke.

Stimmhafte Laute sind daran zu erkennen, dass ihr Spektrumsmaximum bei der Modulationsfrequenz deutlich höher ist als die mittlere Leistung im Spektrogramm. Neben der eigentlichen Modulationsfilterung wurde eine zusätzliche sigmoide Schwelle eingeführt, die spektrale Anteile mit geringer Leistung vollständig unterdrückt, da es sich dabei meist um Rauschen handelt, das durch die Modulationsfilterung zu unangenehmem Knistern würde.

Liegt kein stimmhafter Laut vor, so werden stimmlose Konsonanten an einem relativ gleichmäßigen Plateau im Leistungsspektrum zwischen 1,5 und 10 kHz erkannt. Dieses Maß hat den Vorteil, dass es auch bei instationärem Störgeräusch stabil ist.

Nasale Konsonanten (z.B. „m“, „n“) fallen in keine der beiden Kategorien. Sie zeichnen sich durch stimmhafte Anregung aus, das Spektrum fällt aber oberhalb der Grundfrequenz so schnell ab, dass nahezu nur der Grundfrequenzpeak übrig bleibt. Liegt ein solcher vor, nachdem die beiden anderen Kategorien ausgeschlossen wurden, wird dieser Peak isoliert erhalten.

Ist keines der drei Kriterien erfüllt, so wird in diesem Ansatz davon ausgegangen, dass kein Sprachlaut in dem fraglichen Bereich enthalten ist, und es wird Stille zurückgeliefert.

Hat man es mit einem instationären Störgeräusch zu tun, wie etwa bei dem Geräusch einer durcheinander redenden Menschenmenge, so kann man sich das direkt zunutze machen: Die Leistungsfluktuationen sind dabei häufig deutlich schneller als die mittlere Dauer eines Vokals. Solche Geräusche können daher durch Bildung des Minimums dreier aufeinander folgender Spektrogramme von vornherein verringert werden.

Verständlichkeitsmessung

Die Leistungsfähigkeit der Methode wurde mit Hörtests an Normalhörenden gemessen. Dazu wurde das Material des Göttinger Satztests [5] verwendet. Zusätzlich zu dem stationären Sotschek-Rauschen wurde das eher instationäre Geräusch einer realen Menschenmenge getestet. Für die beiden Rauschtypen wurden jeweils 20 verrauschte Sätze unbearbeitet, mit der FFT-Methode gefiltert und mit der Wavelet-Methode gefiltert über Kopfhörer dargeboten. Der Signal-Rausch-Abstand betrug etwa 0 dB Rauschen gegen die mittlere Vokalleistung. Die Ergebnisse der Messung sind in Tabelle 1 dargestellt:

	Menschenmenge	Sotscheck-Rauschen
unverarbeitet	98,7%	93,4%
Fourier-Methode	76,4%	74,9%
Wavelet-Methode	95,1%	87,4%

Tabelle 1: Erkennungsraten

Die Wavelet Methode erlaubt allgemein eine bessere Erhaltung der relevanten Sprachanteile. Gleichzeitig erlaubt sie mit den beschriebenen Erweiterungen eine Störgeräuschverminderung von 10 dB (gegenüber 6 dB bei der FFT-Methode). Eine Erhöhung der Erkennungsraten konnte nicht erzielt werden, jedoch eine deutliche Erhöhung des Signal-Rausch-Abstands ohne wesentliche Verluste bei der Sprachinformation.

Wir bedanken uns herzlich beim Hörzentrum Oldenburg, das uns das Hörtestmaterial des Göttinger Satztests kostenlos zur Verfügung stellte.

Literatur

[1] Dress, D. B.: *Applications of a Fast Continuous Wavelet Transform.* SPIE Proc. Wavelet Applications IV, 3078:570–580, 1997.

[2] Geurts, L. und J. Wouters: *Enhancing the Speech Envelope of Continuous Interleaved Sampling Processors for Cochlear Implants.* J. Acoust. Soc. Am., 105:2476–2494, 1999.

[3] Langner, G., C. E. Schreiner und U. W. Biebel: *Functional Implications of Frequency and Periodicity Coding in the Auditory Midbrain.* In: Palmer, A. R. et al. (Herausgeber): *Psychophysical and physiological Advances in Hearing*, Seiten 277–285. Whurr Publ. Ltd., London, 1998.

[4] Patterson, R. D.: *Auditory Filter Shapes Derived with Noise Stimuli.* J. Acoust. Soc. Am., 59:640–654, 1976.

[5] Wesselkamp, M., K. Kliem und B. Kollmeier: *Erstellung eines optimierten Satztestes in deutscher Sprache.* In: Kollmeier, B. (Herausgeber): *Moderne Verfahren der Sprachaudiometrie*, Buchreihe Audiologische Akustik. median-verlag, Heidelberg, 1992.

[6] Wilmers, H.: *Hervorhebung von Signalen durch Operationen im Modulationsspektrum.* Fortschritte der Akustik – DAGA 98, 1998.