Robuste Spracherkennung mit spektro-temporalen Filterbankmerkmalen

Marc R. Schädler, Bernd T. Meyer, Birger Kollmeier
Medizinische Physik - Carl von Ossietzky Universität Oldenburg, 26111 Oldenburg, Deutschland
Email: marc.r.schaedler@uni-oldenburg.de

Einleitung

Automatische Spracherkennungssysteme erreichen unter realen Bedingungen längst nicht die Erkennungsleistung des Menschen. Der Grund hierfür ist, dass der Mensch Sprache auch dann noch zu erkennen vermag, wenn das Sprachsignal stark verändert wurde. Er weist somit eine gewisse Robustheit gegenüber Veränderungen des Sprachsignals durch Hintergrundgeräusche oder geänderte Übertragungscharakteristiken - zusammengefasst, extrinsischer Variabilität - auf. Die Robustheit automatischer Systeme gegenüber dieser Variabilität wurde mittels physiologisch motivierter Merkmale, die spektro-temporale Modulationen kodieren können, schon verbessert [3, 1]. Beim Sprechen modulieren Menschen ein Anregungssignal mit ihrem Vokaltrakt sowohl temporal als auch spektral. Die daraus resultierenden spektro-temporalen Muster sind in Abb. 1 zu beobachten, wo auch veranschaulicht ist wie diese mit einem Filter extrahiert werden können. Die zitierten Studien verwen-

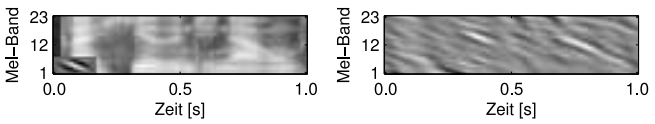

Abbildung 1: Das links dargestellte Log-Mel-Spektrogramm eines gesprochenen Satzes wird mit der ihm unten links überlagerten Filterfunktion gefaltet. Durch die Filterung werden zwei Frequenztransienten deutlich extrahiert (rechts). Helle Bereiche bedeuten viel Energie.

den zu diesem Zweck 2D-Gaborfilter. Zusätzlich finden dort statistische und stochastische Methoden der Modellierung, wie zum Beispiel eine Hauptkomponentenanalyse oder Neuronale Netze, Anwendung. Letztere erfordern spezielles Trainingsmaterial und erschweren so den Vergleich mit Referenzmerkmalen. Hier wird eine Filterbank von 2D-Gaborfiltern zur Merkmalsextraktion genutzt, die dieser zusätzlichen Modellierung nicht bedarf. Die damit gewonnenen Gabor-Filterbank (GBFB)-Merkmale werden mit traditionellen Merkmalen in einem Spracherkennungsexperiment mit dem Aurora 2 Framework bezüglich ihrer Robustheit verglichen.

Spektro-temporale Filterbankmerkmale

Eine spektro-temporale Darstellung des Sprachsignals in Form eines logarithmisch skalierten Mel-Spektrogrammes mit 23 Bändern zwischen 64 Hz und 4000 Hz und 100 Hz Abtastrate bildet wie auch bei den traditionellen Mel Frequency Cepstral Coefficients (MFCCs)-Merkmalen den Ausgangspunkt für die weitere Verarbeitung.

Eindimensionale Gaborfilter sind das Produkt einer Einhüllenden der Breite b und eines komplexwertigen sinusoiden Trägers der Kreisfrequenz ω, vgl. Gln. 1 und 2. Ein 2D-Gaborfilter $g_{\omega_k,\omega_n,\nu_k,\nu_n}(k,n)$ ist das Produkt zweier 1D-Gaborfilter und kann, wie in Gl. 3, auch als Produkt eines Trägers und einer Einhüllenden aufgefasst werden, wobei k und n die Laufindizes der spektralen und temporalen Dimension sind.

$$h_b(x) = \begin{cases} 0.5 - 0.5\cos\left(\frac{2\pi x}{b}\right) & -\frac{b}{2} < x < \frac{b}{2} \\ 0 & \text{sonst} \end{cases} \quad (1)$$

$$s_\omega(x) = \exp(i\omega x) \quad (2)$$

$$g_{\omega_k,\omega_n,\nu_k,\nu_n}(k,n) = \underbrace{s_{\omega_k}(k)\, s_{\omega_n}(n)}_{\text{Träger}} \cdot \underbrace{h_{\frac{\nu_k}{2\omega_k}}(k)\, h_{\frac{\nu_n}{2\omega_n}}(n)}_{\text{Einhüllende}} \quad (3)$$

Die Ausdehnung der Filter ist durch die zentralen Modulationsfrequenzen (MF) (ω_k, ω_n) und die Anzahl der Halbwellen unter der Einhüllenden (ν_k, ν_n) in der jeweiligen Dimension gegeben. Es wird nur der Realteil der Filterfunktionen verwendet, wobei ein möglicher Gleichanteil durch Subtraktion eines geeigneten Anteiles der Einhüllenden kompensiert wird.

Da die Erzeugung von Modulationen durch Sprache physikalisch-physiologischen Grenzen unterliegt ist nur ein gewisser Bereich spektraler und temporaler MF für die Spracherkennung von Bedeutung. Der Ansatz der Filterbank ist, den für die Spracherkennung relevanten Bereich gleichmäßig abzudecken, um:

a) durch die Filterbankstruktur bedingte Reduktion der Anzahl der Parameter den Einfluss stochastischer Prozesse zu verringern.

b) durch *gleichmäßige* Abdeckung Redundanzen, und damit den Einsatz einer den Vergleich erschwerenden Redundanzreduktion, zu vermeiden.

Für a) werden alle Filter mit gleicher Güte angenommen, haben also dieselben Werte für ν_n und ν_k. Für b) wird der (modulations)spektrale Überlapp benachbarter Filter konstant angenommen, wodurch die zentralen MF logarithmisch anordnet sind, vgl. Abb. 2. Dabei wird die Ausdehnung der Filter auf 69 Bänder[1] und 400 ms beschränkt. Als gute Werte ergaben sich in Pilotexperimenten: $\nu_n = 3{,}5\,;\ \nu_k = 3{,}5\,;$

$$\omega_k = \pm\{\,25{,}0\,;\ 12{,}23\,;\ 5{,}99\,;\ 2{,}93\,;\ 0{,}0\,\} \quad \cdot 10^{-2}\ \tfrac{\text{Zyklen}}{\text{Band}}$$

$$\omega_n = \pm\{\,25{,}0\,;\ 15{,}70\,;\ 9{,}86\,;\ 6{,}19\,;\ 0{,}0\,\} \quad \text{Hz}$$

Durch Kombination der spektralen und temporalen MF ergeben sich 41 unterschiedliche Filter. Diese extrahieren

[1] die Filter dürfen über das Spektrogramm hinausragen

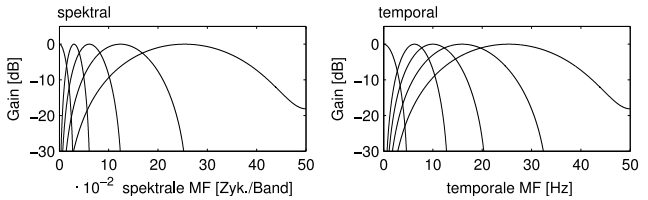

Abbildung 2: Übertragungsfunktionen der Filterbank in spektraler und temporaler Dimension.

teils rein spektrale ($\omega_n = 0$), rein temporale ($\omega_k = 0$), oder spektro-temporale Muster, vgl. Filterfunktionen in Abb. 3. Zur Berechnung der GBFB-Merkmale wird das Log-Mel-Spektrogramm des Sprachsignals, wie in Abb. 3 skizziert, mit jedem der 41 Filter gefiltert, und anschließend systematisch eine repräsentative Auswahl der Bänder getroffen, um Redundanzen zu vermeiden. Zu dieser Auswahl gehört immer das zentrale Mel-Band (12) und jene, die durch Verschiebung des Filters um $\frac{1}{4}$ seiner Ausdehnung in spektraler Dimension hervorgehen. Die repräsentativen Bänder aller Filter zusammen bilden die 311-dimensionalen GBFB-Merkmale.

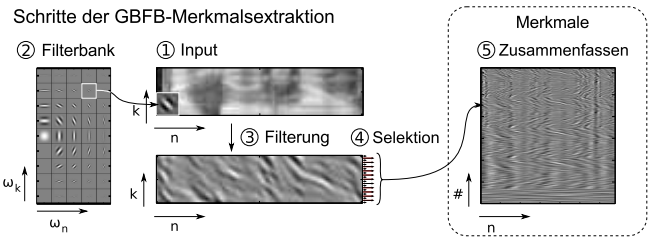

Abbildung 3: Berechnung der Gaborfilterbank-Merkmale. Das Log-Mel-Spektrogramm ① wird mit allen 41 Gaborfiltern ② gefiltert ③. Anschließend werden in Abhängigkeit der Ausdehnung der Filter in spektraler Dimension Sätze repräsentativer Bänder gewählt ④, welche zusammen die GBFB-Merkmale bilden ⑤.

Spracherkenner

Zur Bewertung der Robustheit der Merkmale wird das Aurora 2 Framework für ein Spracherkennungsexperiment verwendet [2]. Die englischsprachigen Aufnahmen bestehen aus Ziffernketten denen Alltagshintergrundgeräusche bei Signal-Rausch-Verhältnissen (SNR) zwischen -5 dB und 20 dB überlagert wurden. Zwei Trainingsmodi existieren: Beim *Clean* Training werden lediglich Aufnahmen *ohne*, beim *Multi* Training auch Aufnahmen *mit* Störgeräuschen verwendet. Die Sprache wird mit Gaußschen Mischverteilungsmodellen und Hidden Markov Modellen modelliert. Neben den MFCC Merkmalen mit erster und zweiter diskreter Ableitung, werden als Referenz auch Ergebnisse für RASTA-PLP (RPLP) Merkmale angegeben. Ausgewertet werden die Wortfehlerraten bei SNRs zwischen 20 dB und 0 dB.

Ergebnisse und Zusammenfassung

Zur Bewertung der Robustheit interessant ist vor allem die Trainingskondition *Clean*, da der Erkenner die Störgeräusche hier nicht modellieren kann. Wie aus Tab. 1 hervorgeht sind die Wortfehlerraten mit GBFB Merkmalen nicht nur im Mittel niedriger als mit MFCC oder RPLP Merkmalen, sondern macht das System auch bei relativer Betrachtung 30% weniger Fehler als mit MFCC Merkmalen. Des Weiteren bleiben die Ergebnisse für GBFB Merkmale, im Gegensatz zu RPLP Merkmalen, auch in der Trainingskondition *Multi* besser als mit MFCCs.

Tabelle 1: Aurora 2 Ergebnisse. Mittlere Wortfehlerrate in % und relative Verbesserung in % gegenüber den MFCC Merkmalen sind angeben. Dabei bedeuten 50% relative Verbesserung eine Halbierung der Wortfehlerrate im Mittel.

	Wortfehlerrate [%]		rel. Verbesserung [%]	
	clean	multi	clean	multi
MFCC	41,9	13,0	0,0	0,0
RPLP	35,8	15,1	16,2	-31,1
GBFB	33,4	11,9	30,0	17,7

Die Unterschiede sind nach Abb. 4 besonders bei SNRs über 10 dB ausgeprägt, was auf eine nicht nur *robuste*, sondern auch *gute* Darstellung der Sprache durch GBFB Merkmale schließen lässt. *Gut* im dem Sinne, dass die erhöhte Robustheit nicht durch schlechtere Erkennungsleistung bei der Erkennung unverrauscher Aufnahmen, und in der Trainingskondition *Multi* erkauft wird.

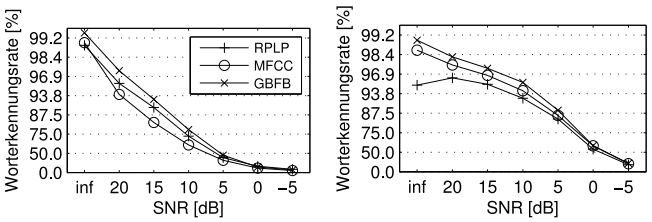

Abbildung 4: Mittlere Worterkennungsraten auf logarithmischer Wortfehlerraten-Skala in Abhängigkeit des SNR. Der Abstand der horizontalen Linien entspricht einer Halbierung der Wortfehlerrate.

Literatur

[1] B.T. Meyer and B. Kollmeier. Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition. *Speech Commun.*, 2010.

[2] D. Pearce and H.G. Hirsch. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proc. of ICSLP 2000*, volume 4, pages 29–32, 2000.

[3] S.Y. Zhao, S. Ravuri, and N. Morgan. Toward a many-stream framework of cortically-inspired spectro-temporal modulation features for automatic speech recognition. *Speech Commun.*, 2010.