Robustheit automatischer Spracherkennung mit Amplitudenmodulationsspektrogrammen

Niko Moritz[1], Bernd T. Meyer[2], Jörn Anemüller[2] und Birger Kollmeier[1,2]

[1] *Fraunhofer IDMT / Hör-, Sprach- und Audiotechnologie, 26129 Oldenburg, E-Mail: morino@idmt.fraunhofer.de*
[2] *Medizinische Physik, Carl-von-Ossietzki Universität Oldenburg*

Einleitung

Heutige automatische Spracherkennersysteme erreichen bei weitem nicht die Leistung des menschlichen Gehörs, wenn sie in akustisch schwierigen Bedingungen, wie etwa störgeräuschbehafteten oder halligen Umgebungen, eingesetzt werden. Hier wird zur Erhöhung der Robustheit der Ansatz einer am Gehör orientierten Musterextraktion durch Amplitudenmodulationsspektrogramme (AMS) beschrieben. Mit den AMS wird die menschliche Verarbeitung von Modulationsfrequenzen nachempfunden, indem für jeden Frequenzkanal der Kurzzeitspektralanalyse auch das Modulationsspektrum bestimmt wird. Die Evaluation der Robustheit mit den AMS als Merkmale zur Spracherkennung wird mit dem Aurora-2 Framework durchgeführt.

Amplitudenmodulationsspektrogramme (AMS)

AMS sind motiviert durch psycho-physische und neurophysiologische Erkenntnisse über die Verarbeitung von Amplitudenmodulationen im auditorischen System von Säugetieren. Untersuchungen von Langner und Schreiner (1988) haben eine periodotope, auf bestimmte Modulationsfrequenzen abgestimmte Neuronenanordnung im Colliculus inferior nachgewiesen, die nahezu orthogonal zu der tonotopen, auf die Mittenfrequenzen abgestimmten Anordnung, liegt. Zudem haben psychoakustische Untersuchungen von Dau et al. (1997) die These einer Modulationsfrequenzanalyse für jedes Mittenfrequenzband belegt.

In die Signalverarbeitung wurden diese Erkenntnisse von Kollmeier und Koch (1994) in Form des Amplitudenmodulationsspektrogramms umgesetzt, welches durch die Analyse der Mitten- und Modulationsfrequenzen gewonnen wird.

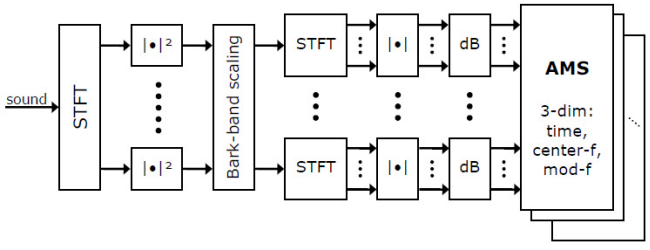

Abbildung 1: Blockschaltbild der Signalverarbeitungsschritte für die Berechnung der AMS.

Abbildung 1 zeigt ein Blockschaltbild der Signalverarbeitungsschritte für die Berechnung der AMS. Das zu analysierende Sprachsignal wird zunächst mittels einer STFT („short-time fourier transformation“) für kurze Signalabschnitte einer Frequenzanalyse unterzogen. Mit dem Betragsquadrat wird im nächsten Schritt die Einhüllende des aus der STFT resultierenden Spektrogramms berechnet. Anschließend werden die einzelnen Frequenzkanäle zu Bark-Bändern zusammengefasst. Im nächsten Schritt der AMS-Berechnung wird jeder einzelne Frequenzkanal des betragsquadrierten und gefilterten Spektrogramms erneut einer Kurzeitspektralanalyse unterzogen. Auf diese Weise wird das Modulations-Spektrum für jede Mittenfrequenz der Bark-Filterbank bestimmt. Im letzten Schritt folgt die Betragsbildung und Logarithmierung der komplexen AMS-Koeffizienten.

Problematik der Analysefensterlänge

Resultierend aus den beiden STFTs haben die AMS folglich vier Parameter, um die zeitliche sowie spektrale Auflösung der Mitten- und Modulationsfrequenzen zu steuern. Problematisch ist hierbei vor allem die zeitliche Länge des Fensters für die Analyse der Modulationsfrequenzen. In Anlehnung an Kanedera et al. (1999) trägt nur der sehr schmale Modulationsfrequenzbereich zwischen 1 Hz und 16 Hz zur robusten automatische Spracherkennung (ASR) bei. Um in diesen schmalen Frequenzbereich ausreichend viele Abtastwerte zu erhalten, werden sehr lange Analyseblocklängen benötigt. Andererseits wird für die zuverlässige Kurzzeitspektralanalyse vorausgesetzt, dass sich die interessierenden Eigenschaften des Signals innerhalb eines Analyseabschnittes nicht ändern und somit als näherungsweise stationär betrachtet werden können. D.h. die Amplitudenmodulationen dürfen innerhalb eines Analyseabschnittes zeitlich nicht zu stark variieren, da sonst die Änderungen im AMS-Muster ausgemittelt werden.

Die spektrale Auflösung der Modulationsfrequenzen ergibt sich aus der Gleichung

$$\frac{f_s}{N} = \left(\frac{1000}{Shift_{cf}}\right) \Big/ \left(\frac{BL_{\mathrm{mod}}}{Shift_{cf}}\right) = \frac{1000}{BL_{\mathrm{mod}}} \qquad (1)$$

wobei $Shift_{\mathrm{cf}}$ den Fenstervorschub der ersten STFT und BL_{mod} die Blocklänge der zweiten STFT in Millisekunden darstellen. Die Variable f_s ist die Abtastfrequenz und N ist die Anzahl der Abtastpunkte innerhalb des Analysefensters. Aus der Gleichung folgt, dass um bspw. die Modulationsfrequenzauflösung von 2 Hz zu erzielen ein 500 ms langes Analysefenster nötig ist. Diese Anforderung würde allerdings mit der zeitlichen Länge der Phoneme oder sogar mit der Länge eines ganzen Wortes kollidieren, wodurch diese langen Analysefenster vermutlich nicht für den Zweck der Spracherkennung geeignet sind.

Experimentelle Auswertung

Es wurden Experimente durchgeführt, um die zuvor beschriebene Problematik mit der zeitlichen Länge der Analysefenster zu evaluieren. Die Testumgebung für die hier dargestellten Experimente basiert auf dem Aurora-2

Framework. Die Testsätze A und B aus dem Framework beinhalten jeweils vier unterschiedliche Rauschszenarien, die mit den SNRs von 20 dB bis -5 dB zu den rauschfreien Testdaten gemischt wurden. Der Testsatz C umfasst insg. zwei der Rauschszenarien aus den Testsätzen A und B. Zusätzlich wurden die Testdaten aus C mit einer Telefonübertragungscharakteristik gefaltet.

Die Merkmale werden aus dem AMS-Muster gewonnen, indem zunächst der Modulationsfrequenzbereich mit den Grenzfrequenzen 1 Hz und 16 Hz bandpassgefiltert wird. Anschließend werden die bandpassgefilterten AMS-Muster mittels einer Hauptkomponentenanalyse (PCA) auf 39 Merkmale reduziert, um die Vergleichbarkeit mit den MFCCs zu gewährleisten. Für das Experiment werden folgende Einstellungen der AMS-Parameter verwendet:
BL_{cf} = 25 ms, $Shift_{cf}$ = 10 ms, BL_{mod} = ? ms, $Shift_{mod}$ = 10 ms.
Die Parameter BL_{cf} und BL_{mod} beschreiben die Blocklänge der ersten bzw. der zweiten STFT. $Shift_{cf}$ und $Shift_{mod}$ geben entsprechend den Fenstervorschub der beiden STFT Analysen an. Nun wurde die Analysefensterlänge BL_{mod} für die Berechnung der Modulationsfrequenzen über einen weiten Zeitbereich variiert und die durchschnittlichen Worterkennungsraten mit Aurora-2 ermittelt. Das Ergebnis ist in Abbildung 2 dargestellt.

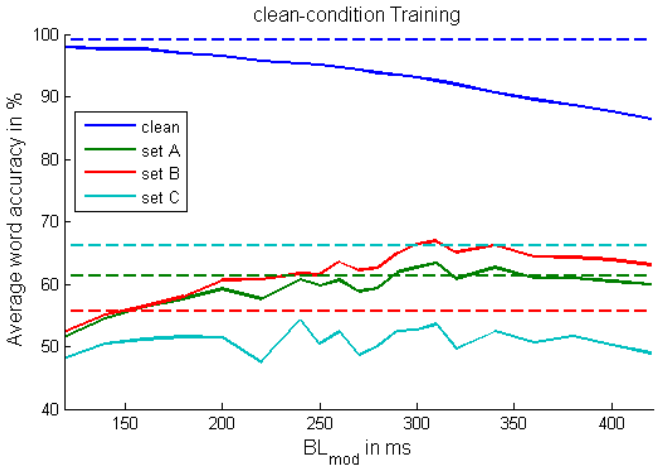

Abbildung 2: Dargestellt ist die durchschnittliche Worterkennungsrate (gemittelt über SNRs von 0 bis 20 dB) der Aurora 2 Testmethoden A, B und C für variierende Analysefensterlängen BL_{mod}. Zusätzlich ist die durchschnittliche Worterkennungsrate der rauschfreien Testaufnahmen („clean“) dargestellt. Die gestrichelten Linien repräsentieren die entsprechenden durchschnittlichen Worterkennungsraten der Aurora-2-Baseline mit den MFCC-Merkmalen. Die Ergebnisse wurden mit dem „clean-condition“ Training erzielt.

Dargestellt sind die durchschnittlichen Worterkennungsraten („word recognition rates“, WRR) der Testsätze A, B und C. Die durchschnittlichen WRRs beziehen sich hierbei auf die SNRs von 20 dB bis 0 dB. Zusätzlich ist in der Abbildung auch die durchschnittliche WRR der rauschfreien Testaufnahmen eingezeichnet, um den allgemeinen Trend der Erkennungsraten für rauschfreie Sprache zu zeigen. Offenbar nimmt die allgemeine Erkennerleistung für rauschfreie Sprache mit zunehmender Analysefensterlänge BL_{mod} stark ab. Dieser Effekt hängt mit der vorher beschriebenen Problematik der Analysefenster zusammen. Demgegenüber nimmt die Robustheit für verrauschte Sprache mit zunehmender Blocklänge zu. Ein Maximum der durchschnittlichen WRR für die Testsätze A, B und C ist bei ca. 310 ms zu erkennen. Das bedeutet, dass die Robustheit durch eine höhere Abtastung der Modulationsfrequenzen verbessert werden kann. Bessere durchschnittliche WRRs werden jedoch wieder durch die allgemein verschlechterte Spracherkennerleistung verhindert.

Mehrere simultan verwendete Analysefenster

Um den Konflikt zwischen der verbesserten Robustheit und der allgemein verschlechterten Erkennerleistung für lange Analysefenster mit den AMS-Merkmalen zu kompensieren, können gleichzeitig mehrere Blocklängen für die Analyse der Modulationsfrequenzen herangezogen werden. Tabelle 1 zeigt ein Beispiel für die gesteigerten Erkennungsraten durch ein zusätzliches Analysefenster BL_{mod}.

Tabelle 1: Durchschnittliche WRR (zwischen 0 und 20 dB) für die Testsätze A – C. Zusätzlich ist die durchschnittliche WRR für die rauschfreien Testdaten zu sehen. Eingetragen sind die Ergebnisse für ein einzelnes Analysefenster BL_{mod}, sowie für zwei Analysefenster. Zusätzlich sind zur Referenz die MFCC-Erkennungsraten eingetragen.

BL_{mod} in ms	Clean	Set A	Set B	Set C
310	92,6%	63,4%	66,89%	53,6%
310 + 190	96,22%	65,96%	69,15%	59,46%
MFCCs	99,02%	61,34%	55,74%	66,14%

Fazit

Es konnte gezeigt werden, dass die AMS durch die Betrachtung der Amplitudenmodulationen sehr robust gegenüber Störschall sind. Zur Analyse der Modulationsfrequenzen sind jedoch lange zeitliche Blöcke nötig, wodurch die zeitliche Auflösung der Amplitudenmodulationen verringert wird. Dies führt zu einer Reduzierung der Erkennungsraten für unverrauschte Sprache. Im Gegenzug wird durch eine verbesserte Modulationsfrequenzauflösung die Robustheit der AMS-Merkmale erhöht. Durch die Anwendung verschiedener Blocklängen zur Analyse der Amplitudenmodulationen kann der Trend der allgemein verschlechterten Erkennungsraten jedoch kompensiert und zusätzlich die Robustheit weiter erhöht werden.

Literatur

[1] Tchorz, J.; Kollmeier, B.: Automatic classification of acoustical situation using amplitude modulation spectrograms. J. Acoust. Soc. Amer. 105 (2), 1157 (1999)

[2] Tchorz, J.: Auditory-based Signal Processing for Noise Suppression and Robust Speech Recognition. Universität Oldenburg, Diss (2000).

[3] Kanedera, N.; Arai, T.; Hermansky, H.; Pavel, M.: On the relative importance of various components of the modulation spectrum for automatic speech recognition. Speech Communication 28 (1999) 43-55.

[4] Lippmann, R.: Speech recognition by machines and humans. J. Speech Communication (Vol. 22, pp. 1-15). Elsevier (1997).