Vergleich von Verfahren zur Sprachaktivitätsdetektion für robotisches Hören

Lehrstuhl für Multimediakommunikation und Signalverarbeitung
der Friedrich-Alexander-Universität Erlangen-Nürnberg

Forschungspraktikum

vorgelegt von

Mack,Wolfgang
Matrikelnummer: 21622394

im April 2016

Betreuer: Dipl.-Ing Stefan Meier

Inhaltsverzeichnis

1 Sprache als Schnittstelle zwischen Mensch und Maschine

Menschen kommunizieren untereinander auf verschiedene Arten und Weisen. So können Briefe geschrieben werden, oder miteinander gesprochen werden. Bei der Kommunikation, der Steuerung von hergestellten Produkten, wie beispielsweise einem Computer, ist die Schriftform über die Tastatur schon lange als Schnittstelle Standard. Je fortschrittlicher technische Systeme werden, desto vielfältiger werden die Möglichkeiten, mit ihnen zu kommunizieren, sei es durch Gesten oder Sprache. Als Ergebnis wird erhöhter Komfort für den Benutzer, wie leichtere Bedienung, erhalten. Diese Arbeit beschäftigt sich mit der Detektion von gesprochener Sprache, so dass entschlüsselnde Algorithmen pausiert werden können, bis ein gesprochenes Wort detektiert wird. Es werden verschiedene Merkmale aus kleinen Sprachfragmenten, 25 ms, extrahiert und diese anschließend bezüglich ihres Erfolges bei verschiedenen Signal-Rausch-Leistungsverhältnissen verglichen. Als Rauschen werden Weißes-Gaußsches Rauschen und Babble-Rauschen verwendet. Die extrahierten Eigenschaften werden sowohl im Zeit- als auch im Frequenzbereich betrachtet. Für die Detektion wird ein neuronales Netz verwendet. Zu Beginn wird auf die Merkmale eingegangen und anschließend werden sie evaluiert.

2 Merkmalsextraktion

In diesem Kapitel wird beschrieben, wie Merkmale aus den Sprachframes extrahiert werden. Die Frames haben eine Länge von jeweils 25 ms und die Anfangspunktverschiebung von aufeinanderfolgenden Teilen beträgt 10 ms. Sie überlappen sich somit.[1] [2]

2.1 Zeitbereichsmerkmale

Hier werden Zeitbereichsmerkmale von Sprachsignalen extrahiert.

2.1.1 Signal-Rausch-Leistung

Sprechen Menschen, passen sie die Lautstärke der Umgebung an. Dieser Effekt wird als Lombard-Effekt bezeichnet. In einer rauscharmen Umgebung kann also die Energie des Signals zur Erkennung von Sprache verwendet werden. Dies ist ein sehr simples Merkmal und ist in Situationen mit Hintergrundrauschen nicht mehr verwendbar, sofern sich die Rauschleistung in Höhe der Sprachsignalleistung befindet. Zum Test des Algorithmus sind Sprach- und Rauschdateien vorhanden. Die Leistung wird verwendet, um Sprachpausen aus den sehr wenig verrauschten Sprachdateien zu extrahieren. Ist die Energie eines Frames unter einer heuristisch festzulegenden Grenzenergie, wird dieser als Pause erkannt.

2.1.2 Autokorrelation

Stimmhafte Phone werden durch Schwingung der Stimmbänder und anschließender Filterung des Impulses durch den Vokaltrakt erzeugt. Dieser Grundimpuls hat eine sprecherabhängige Frequenz von 50 - 400 Hz [4]. Damit werden sehr tiefe Männerstimmen und sehr hohe Kinderstimmen abgedeckt. In der Implementierung wurde die Autokorrelation berechnet, auf ihre Energie normiert und anschließend um obigen Frequenzbereich gefiltert. Sollte es sich um stimmhafte Sprache handeln, müssen durch den Grundimpuls mehrere Peaks auftreten. Der Wert des Maximums dieser gefilterten Autokorrelation ist somit ein Indiz für stimmhafte Sprache, da dieser aus dem Grundimpuls resultiert. Störungen mit einer energiereichen Grundschwingung in diesem Bereich senken den Wert dieses Merkmals drastisch. [4]

2.2 Spektralbereichsmerkmale

Der Mensch hört im Bereich von ca. 16 - 20000 Hz. Die spektrale Zusammensetzung ist für Sprachsignale signifikant und wird in diesem Abschnitt analysiert.

2.2.1 Spektrale Entropie

Die spektrale Entropie ist ein Maß für die Unterschiedlichkeit der vorkommenden Spektralanteile. Sie ist maximal, wenn alle Spektralanteile die gleiche Höhe besitzen. Die Frequenzachse wird in eine Dichte verwandelt, dabei werden alle Elemente aufsummiert und jedes einzelne durch die Summe geteilt.

$$Dichte = \frac{fftWert}{\sum_{-20000Hz}^{20000Hz} fftWerte} \tag{2.1}$$

Die einzelnen, diskreten Werte der Dichte werden als p_x bezeichnet. Aus der so erhaltenen Dichte wird die spektrale Entropie berechnet:

$$Entropie = \sum_{alleX} p_x \cdot log(p_x) \tag{2.2}$$

Der Wert der Entropie ist ein Maß für die Unterschiedlichkeit der Energie der vorkommenden Frequenzen. Je nach Art des Rauschens kann sie Sprachsignalaktivitätsdetektion also unterstützen oder sogar gewährleisten.

2.2.2 Mel-Spektrum, Delta-Koeffizienten und Delta-Delta-Koeffizienten

Das menschliche Gehör kann Frequenzen nicht genau unterscheiden. Es hört gewisse Bereiche und ordnet diese zu. Je tiefer die Frequenz, desto kleiner ist dieser Bereich. Die Mel-Skala ist eine an dieses Phänomen angepasste Frequenzskala. In der Implementierung werden die Frequenzen von 300 bis 8000 Hz berücksichtigt. Es werden 26 Mel-Filter verwendet und von den erhaltenen Koeffizienten die Nummern 2-13 verwendet. Diese stellen teile der Frequenzachse dar und sind signifikante Merkmale für Sprache. Man erhält diese Koeffizienten, indem über berechnete Abschnitte der Frequenzachse aufsummiert wird. Die so erhaltenen Werte werden logarithmiert und mithilfe der diskreten Cosinustransformation in Melmerkmale umgewandelt. Da bei Sprache viel Information im dynamischen Verhalten liegt, wird die Änderung dieser Koeffizienten über verschiedene Frames, sowie die Änderung der Änderung zusätzlich mit einbezogen. Der so erhaltene Merkmalsvektor enthält 36 Einträge und ist ein sehr gutes Indiz für Sprache.

3 Aufbau der Implementierung

Dieses Kapitel bietet einen kurzen, schematischen Überblick, wie der Code aufgebaut ist. Zum Erhalt eines neuronalen Netzes kann das Script GetNet.m ausgeführt werden. Die Graphik stellt den Aufbau dieses Skriptes dar. Jedes Viereck ist ein Modul. Oben steht der Funktionsname, in der Mitte die Eingabeparameter und unten die Ausgaben. Zuerst werden die Audiodateien eingelesen und fragmentiert, das geschieht in der Funktion Read_Fragment_Audio. Die so erhaltenen Fragmente können, je nach Bedarf, mit der Funktion mixFragments bei verschiedenen SNR überlagert werden. Um die Fragmentpositionen der Sprache zu finden, kann die Funktion Filter_Fragments_By_Energy verwendet werden. Diese Funktion erkennt Sprachpausen, da deren Energie kleiner als eine heuristisch bestimmte Grenzenergie ist. Anschließend wird mit der Überlagerung und dem Rauschen eine Extraktion der Merkmale vorgenommen. Die verwendete Funktion heißt Get_Features_From_Fragments. Die Funktion GetNet gibt aus einer festgelegten Anzahl von Iterationen das beste erhaltene neuronale Netz zurück. Ihr werden die Featurevektoren und die zu behaltenden Positionen übergeben. Daraus berechnet die Funktion die Parameter für ein neuronales Netz und startet dieses.

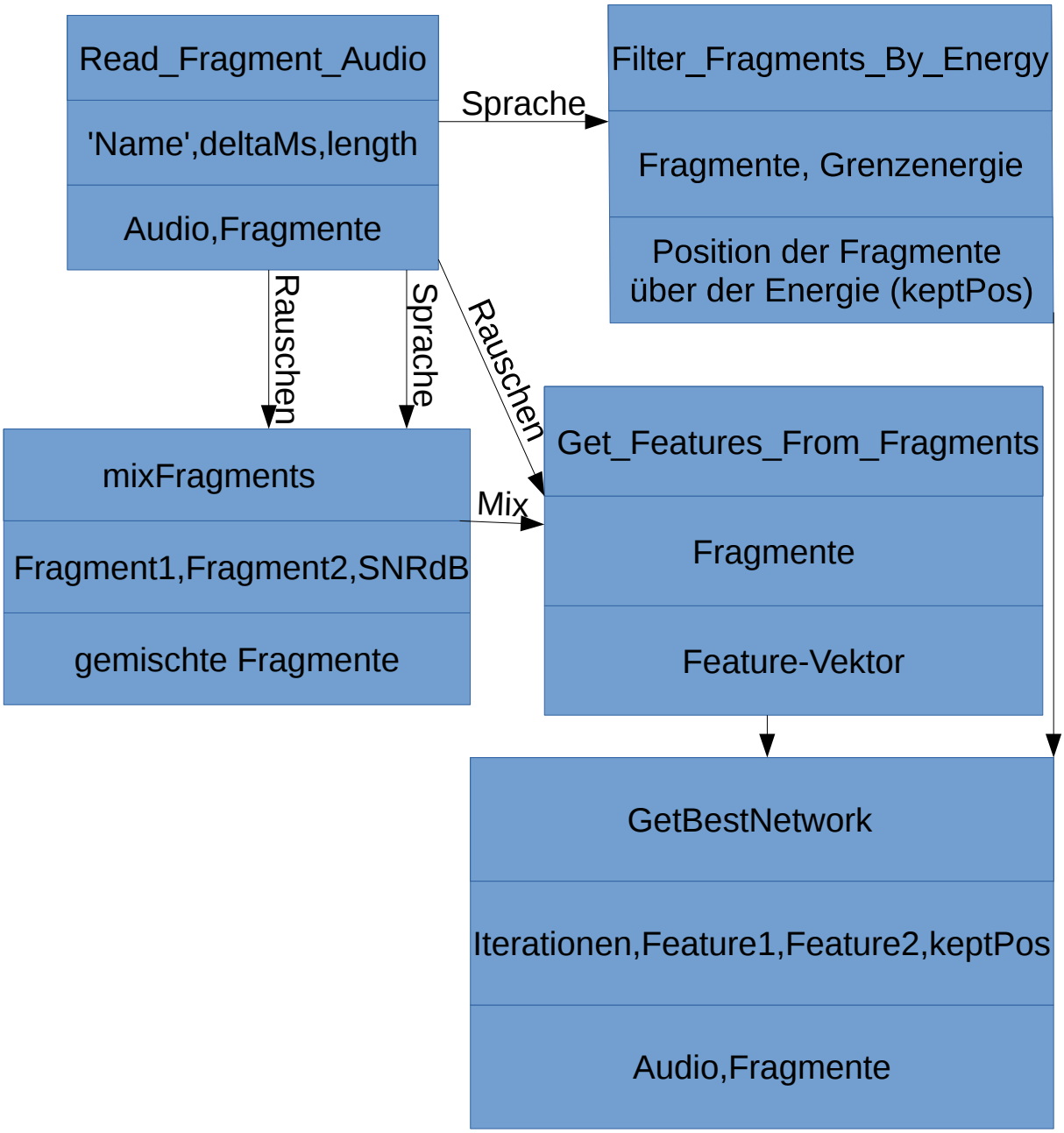

Abbildung 3.1: Schematischer Überblick über den Aufbau des Programmiercodes

4 Evaluation der Merkmale

In diesem Kapitel werden die Merkmale anhand der Fläche unter der ROC-Kurve analysiert.

Tabelle 4.1: AUC

AUC	WGN -10dB	Babble -5dB	Babble 0dB
AllFeature	0.9955	0.8503	0.9796
Mel/Deltas	0.9812	0.8536	0.9609
ohne Entropie	0.9971	0.8401	0.9746
ohne Autokorr.	0.9933	0.8406	0.9643
Entropie	0.8887	0.5582	0.6599
Autokorr.	0.9734	0.5572	0.7335

Aus der Tabelle ist ersichtlich, dass die Melkoeffizienten und deren Deltas sowohl bei weißem Gaußschen Rauschen, als auch beim Babble Rauschen sehr gut abschneiden. Die Fläche unter der ROC-Kurve ist nur unwesentlich kleiner, wenn die Autokorrelation oder die Entropie entfernt werden. Bei Babble Rauschen liefern Entropie und Autokorrelation sehr schlechte Ergebnisse. Dies liegt daran, dass Babble Rauschen bereits Sprache im Hintergrund aufweist bzw. harmonische Strukturen besitzt. Die Entropie erkennt Verteilungsunterschiede bei den vorkommenden Frequenzen. Ist die Verteilung ähnlich unterschiedlich, liefert sie unabhängig von den Frequenzen, ein ähnliches Ergebnis. Sie liefert somit bei weißem Gaußschen Rauschen sehr gute Ergebnisse, bei Babble Rauschen, welches starke spektrale Unterschiede aufweist, jedoch sehr schlechte. Hat die Autokorrelation einen Peak, resultierend aus einer Schwingung zwischen 50 und 400 Hz, so ist dies ein Indiz für Sprache. Da Babble Rauschen harmonisch ist, kann dies auch in dessen Autokorrelation vorkommen. Das Feature liefert also schlechtere Ergebnisse. Bei weißem Gaußschen Rauschen jedoch sehr gute. Zusammenfassend kann gesagt werden, dass eine Implementierung von Melkoeffizienten und deren Deltas sehr robust ist, sowohl in Bezug auf verschiedene SNR, als auch auf Rauscharten. Als Ergänzung kann die Entropie oder die Autokorrelation verwendet werden. Ist das Rauschen weiss, so kann die Entropie oder die Autokorrelation verwendet werden, da sie weniger aufwändig zu berechnen sind und sehr gute Ergebnisse erzielen.

5 Erweiterung der Sprachaktivitätserkennung

Sprache kann in stimmhafte und nicht stimmhafte Phone unterteilt werden. Erstere werden erzeugt, indem die Stimmbänder eine Grundschwingung zwischen 50 und 400 Hz erzeugen, welche im Vokaltrakt gefiltert wird. In dieser Arbeit wird nicht auf deren Unterscheidung eingegangen. Bei stimmloser Sprache fliesst die Luft durch die Stimmbänder. Zischlaute, wie ein s, sind ein Beispiel für letzteres. Die Energie stimmloser Sprachanteile ist kleiner als die stimmhafter. Da die Autokorrelation bei weissem Gausschen Rauschen auch bei minus 10 dB sehr gute Ergebnisse erzielt, kann es sein, dass diese Phone in der Implementierung unter die Energiegrenze der Sprachpausenfilterfunktion fallen, oder sehr selten vorkommen. Eine mögliche weiterführung der Arbeit wäre also die automatische Anpassung der Featurevektoren und Netze an verschiedene Rauscharten und die Unterscheidung von stimmhafter und stimmloser Sprache. Wird der Algorithmus verwendet, um mit nachgeschaltener Software Worte und Sätze zu erkennen, kann eventuell vorhandene schlechtere Erkennung von stimmloser Sprache anderweitig verringert werden. So ist es möglich, aus den erkannten Buchstaben und einer Tabelle möglicher auftretender Wörter, die fehlenden zu berechnen und einzufügen. Dies ist nur ein möglicher Erweiterungsansatz.

Literaturverzeichnis

[1] Features for voice activity detection a comparative analysis ,
author = Simon Graf,TobiasHerbig, Markus Buck, Gerhard Schmidt
eurasip journal on advances in signal processing 2015.

[2] Practical Cryptography,
url = http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/,
urldate = 2016-04-03.

[3] Sprachverarbeitung grundlagen und methoden der sprachsynthese und spracherkennung,
author = Dr. Beat Pfister, Tobias Kaufmann,
isbn = 978-3-540-75909-6
springer.

[4] Voicing features for robust speech detection,
author = Trausti Kristjansson, Sabine Deligne, Peder Olsen,
interspeech 2005.