Vergleich von Merkmalsextraktionsverfahren für die automatische Sprecherverifikation bei Nichtübereinstimmung des Stimmaufwands in Trainings- und Testdaten

Corinna Harwardt

Fraunhofer FKIE, 53343 Wachtberg, Deutschland, Email: corinna.harwardt@fkie.fraunhofer.de

Einleitung

Die automatische Sprecherverifikation auf Audiodaten mit weitestgehenden Übereinstimmungen der Signaleigenschaften in Trainings- und Testmaterial liefert für viele Szenarien und Signalqualitäten bereits sehr gute Ergebnisse. Stimmen die Signaleigenschaften jedoch nicht überein, so sinkt die Erkennungsrate häufig rapide ab. Ein Fall der Nichtübereinstimmung ist die Erhöhung des Stimmaufwands in einem der Signale. Erhöht der Sprecher seinen Stimmaufwand, um beispielsweise Hintergrundgeräusche zu übertönen, so ändern sich die akustischen Eigenschaften des Sprachsignals stark. Bisher ist jedoch noch kein klares Muster zur Beschreibung dieser Veränderungen der verschiedenen akustischen Parameter gefunden worden, da die Veränderungen sprecherabhängig zu sein scheinen. Um dieses Problem speziell für die automatische Sprecherverifikation zu untersuchen, werden in dieser Arbeit bestehende Standardmerkmalsextraktionsverfahren auf ihre Leistung in einem solchen Szenario verglichen. Die Ergebnisse dieser Arbeit bilden demnach die Grundlage um bestimmte Merkmale für die weitere Nutzung in einem solchen Szenario auszuschließen und in weiteren Schritten das beste Merkmal auf seine Schwachpunkte für die Sprechererkennung bei verschiedenen Stimmaufwandsgraden zu untersuchen und gegebenenfalls zu verbessern.
Im Folgenden wird zunächst ein Überblick über den Einfluss von erhöhtem Stimmaufwand auf die akustischen Merkmale von Sprache gegeben. Dann wird das Sprechererkennungssystem vorgestellt mit dem die Tests durchgeführt wurden. Ein Schwerpunkt liegt hier auf der Vorstellung der unterschiedlichen Merkmale. Abschließend werden die Ergebnisse dargestellt und diskutiert.

Auswirkungen von erhöhtem Stimmaufwand auf das Sprachsignal

Bei der Produktion eines Sprachsignals verfolgt der Sprecher unterschiedliche Ziele. Möchte er lauter Sprechen - also den Stimmaufwand erhöhen - kann dies emotional oder durch die Kommunikationsgegebenheiten (Hintergrundgeräusche oder große Distanz) bedingt sein. In dieser Arbeit wird nur die Erhöhung des Stimmaufwands auf Grund von Hintergrundgeräuschen betrachtet.
Untersuchungen zum Stimmaufwand können perzeptiv, artikulatorisch und akustisch motiviert sein. In den Perzeptionsstudien wird untersucht ob der Sprecher sein Ziel - die Verbesserung der Verständlichkeit - trotz ungünstiger Kommunikationsgegebenheiten erreicht. Die artikulatorisch motivierten Studien analysieren wie der Sprecher sein Ziel erreicht. Während die akustischen Untersuchungen das Resultat begutachten. In der Sprechererkennung sind vor allem die akustisch motivierten Studien von Interesse. Bei der Untersuchung der Grundfrequenz zeigte sich in zahlreichen Studien (z.B. [1], [2]), dass die Grundfrequenz bei erhöhtem Stimmaufwand steigt. Die Stärke des Anstiegs lässt sich nicht klar definieren, sodass der Grad des Anstiegs möglicherweise sprecherabhängig ist [2]. Auch die Formanten und ihre Amplituden sind unter den verschiedensten Bedingungen analysiert worden. Hier ergibt sich jedoch kein klares Bild. Für F1 besteht ebenso wie für die Grundfrequenz, die Tendenz zum Anstieg (siehe z.B. [1], [3]). Für den zweiten und dritten Formanten lässt sich keine allgemeingültige Aussage machen. Es zeigt sich jedoch, dass die Veränderungen ausreichend stark sind, um die Leistung sprachverarbeitender Systeme negativ zu beeinflussen [4]. Welche der gängigen Standardmerkmale am wenigsten beeinflusst werden wird in dieser Untersuchung vorgestellt.

Das Sprechererkennungssystem

Um die verschiedenen Merkmalsextraktionsverfahren zu testen wurde ein GMM-UBM basiertes Sprecherverifikationssystem [5] verwendet. Man unterscheidet zwischen der Trainings- und der Testphase. Für beide Phasen muss vorab eine Vorverarbeitung durchgeführt werden, die eine energiebasierte Sprach-Pause-Detektion und die Merkmalsextraktion umfasst. Als Merkmalsextraktionsverfahren wurde variiert zwischen Mel-Cepstrum Koeffienten (MFCC), den Reflektionskoeffizienten der linearen Prädiktion (LPREFC) und der perzeptuellen linearen Prädiktion (PLP) als Mischung der beiden Verfahren, extrahiert mit HTK [6].
Bei der Berechnung der MFCCs wird zunächst eine Fourier Transformation durchgeführt. Anschließend wird eine Logarithmierung vorgenommen um später eine Trennung zwischen Anregung und Vokaltraktformung vornehmen zu können. Um die menschliche Wahrnehmung besser zu modellieren, werden die Frequenzen nach der Mel-Skala gewichtet. Die Frequenzen werden dann mittels Diskreter Cosinus Transformation in den cepstralen Bereich transformiert. Auf Grund der Trennung zwischen

Anregung und Vokaltraktformung enthalten die MFCCs Informationen über die Formantstruktur, nicht aber über die Grundfrequenz, sodass nur die Veränderung der Formanten durch erhöhten Stimmaufwand auf die MFCCs Einfluss nimmt, nicht aber die der Grundfrequenz.
Ein alternativer Ansatz ist die LPC Analyse. Sie folgt dem Grundsatz, dass sich die Parameter aus den vorherigen bestimmen lassen. Die Parameter Schätzung erfolgt rekursiv.
Die PLP Merkmale sind eine Kombination der beiden vorab genannten Verfahren. Eine detaillierte Beschreibung der Merkmale findet sich in [7].
Sind die Merkmale extrahiert können hiermit in der Trainingsphase ein Hintergrundmodell (UBM - Universal Background Model) und das Sprechermodell trainiert werden. Da für die einzelnen Sprecher häufig nicht viel Trainingsmaterial vorhanden ist, wird das UBM als Grundlage des Sprechermodells benutzt indem eine MAP (maximum a posteriori) Adaption des UBMs mit Sprecherdaten durchgeführt wird. Aus diesen Modellen wird dann ein Detektor zusammengestellt (siehe Abb. 1). Die Ausgabe des Sprecherverifikationssystems ergibt sich aus dem Verhältnis der logarithmischen Likelihood Werte (LLR - Log Likelihood Ratio) von UBM und Sprechermodell.

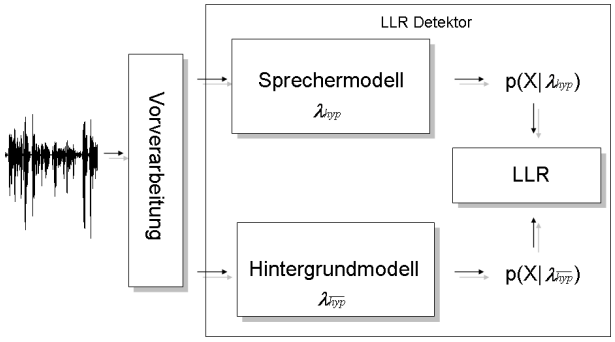

Abbildung 1: Detektor zur Sprecherverifikation

Ergebnisse

Zum Vergleich der Merkmalsextraktionsverfahren wurde das Pool 2010 Korpus verwendet [2]. Es wurden 105 Sprechermodelle auf spontansprachlichen Daten normalen Stimmaufwands mit durchschnittlich ca. 50 Sekunden Sprachanteil trainiert. Zum Testen wurden je Sprecher zwei Aufnahmen mit erhöhtem Stimmaufwand verwendet, die aus der gleichen Aufnahmesitzung stammen. Um den erhöhten Stimmaufwand hervorzurufen wurde den Sprechern weißes Rauschen per Kopfhörer zugeführt. Der durchschnittliche Sprachanteil liegt hier ebenfalls bei ca. 50 Sekunden.
Die Erkennungsergebnisse können in der DET (Detection Error Tradeoff) Kurve in Abbildung 2 abgelesen werden. Wählt man die Gleichfehlerrate (EER - Equal Error Rate) als Vergleichswert, so schneiden die MFCCs am besten ab (4,8%). Die LPC Analyse liefert mit Abstand die schlechtesten Ergebnisse (46,2%). Diese Merkmale scheinen für das gegebene Szenario nicht geeignet zu sein. Die PLP Merkmale hingegen schneiden ähnlich

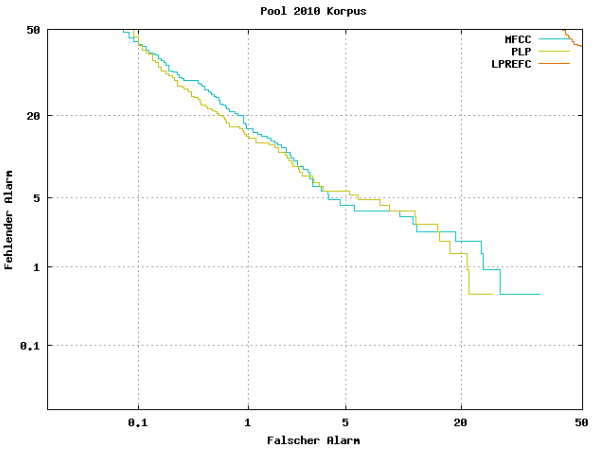

Abbildung 2: Ergebnisse verschiedener Merkmalsextraktionsverfahren auf dem Pool 2010 Korpus

gut wie die MFCCs ab (5,3%). Je nach Operationspunkt auf der DET Kurve liefern die PLPs sogar bessere Ergebnisse.

Schlussfolgerung

Die PLPs und die MFCCs scheinen gleichermaßen geeignet zu sein für Sprechererkennung mit erhöhtem Stimmaufwand. Die EER beider Verfahren ist relativ gering. Verglichen mit den Erkennungsraten für Sprachdaten gleichen Stimmaufwands (0,95% für MFCC) ist die Fehlerrate jedoch ungefähr um das Fünffache erhöht. Eine Verbesserung der Erkennungsrate durch zusätzliche oder an das Szenario angepasste Merkmale ist in nachfolgenden Studien zu untersuchen.

Literatur

[1] Schulman, R.: Articulatory Targeting and Perceptual Constancy of Loud Speech. Phonetic experimental research at the Institute of Linguistics (1985),University of Stockholm

[2] Jessen, M., Köster, O. und Gfroerer, S.: Influence of vocal effort on average and variability of fundamental frequency. International Journal of Speech, Language and the Law (2005)

[3] Liénard, J., Benedetto, M.: Effect of vocal effort on spectral properties of vowels. Journal of the Acoustical Society of America 106 (1999), 411–422

[4] Bořil H.: Robust Speech Recognition: Analysis and Equalization of Lombard Effect in Czech Corpora. Czech Technical University in Prague (2008)

[5] Reynolds, D., Quatieri T. und Dunn R.: Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing 10 (2002), 19-41

[6] Young et al.: The HTK Book (for HTK Version 3.4). 2009, `http://htk.eng.cam.ac.uk/`

[7] Wendemuth A.: Grundlagen der stochastischen Sprachverarbeitung. 2004