EMOSystem: Ein Demonstrationssystem zur Stimmcharakterisierung

Marko Lugger[1], Bin Yang[2]

[1] *Lehrstuhl für Systemtheorie und Signalverarbeitung, 70550 Stuttgart, Deutschland, Email: marko.lugger@lss.uni-stuttgart.de*

[2] *Lehrstuhl für Systemtheorie und Signalverarbeitung, 70550 Stuttgart, Deutschland, Email: bin.yang@lss.uni-stuttgart.de*

Einleitung

In diesem Beitrag möchten wir unser Demonstrationssystem zur Stimmcharakterisierung **EMOsystem** vorstellen. Das System kann zur Analyse der Stimme und insbesondere zur Detektion des Geschlechts, der Stimmqualität und des emotionalen Zustands eines Sprechers eingesetzt werden. Neben der Auswertung von bestehenden Datenbanken ist das System vor allem darauf ausgelegt, Live-Klassifikationen durchzuführen. Darunter verstehen wir die Klassifikation neu aufgenommener Daten von für das System bekannten aber auch fremden Sprechern.

Die Klassifikation paralinguistischer Eigenschaften ist ein Forschungsthema, welches in den letzten Jahren immer mehr an Bedeutung gewinnt. Ein Hauptanwendungsgebiet ist dabei die Detektion von Emotionen aus dem Sprachsignal [1],[2],[3]. Die Erkennung von Ärger in der Stimme eines Anrufers findet in Callcentern auch schon erste Anwendung in der Praxis [4],[5]. Allerdings wird in diesem Forschungsgebiet meist mit gut validierten Datenbanken gearbeitet. Diese sind oft von professionellen Schauspielern gesprochen. Wir versuchen mit diesem System die Klassifikation auf verallgemeinerte Bedingungen anzuwenden. Darunter verstehen wir Live-Daten von nicht professionellen Sprechern unter verschiedenen Aufnahmebedingungen zu analysieren und zu klassifizieren. In diesem Beitrag wird hauptsächlich auf die Realisierung und die Funktionalität des Systems eingegangen. Es werden die extrahierten Merkmale, die verwendeten Klassifizierer und die mit dem System erzielten Ergebnisse vorgestellt.

Implementierung

Umgebung und Systemanforderungen

Das EMOsystem wurde in der Simulationsumgebung MATLAB entwickelt. Es wurde plattformunabhängig programmiert und kann sowohl unter Windows und Linux verwendet werden. Als Mindestanforderung wird ein leistungsstarker PC und 1 GB RAM empfohlen. Die Sprachaufnahmen wurden mit einem hochwertigen Funkmikrofon von AKG bei 16 kHz Abtastfrequenz und einer Quantisierung von 16 bit durchgeführt.

Oberfläche und Steuerung

Das System besteht aus einer Aufnahmeeinheit, einer Einheit zur Merkmalsextraktion und einer Klassifikationseinheit. Die Steuerung geschieht über graphische Benutzeroberflächen (GUI). Neben einer Hautgui gibt es noch eine weitere zur Steuerung der Live-Klassifikation. Die Oberfläche ist in Abb.1 dargestellt. Im linken Teil der GUI lassen sich alle Einstellungen zu den Klassifikationsanalysen treffen. Diese beinhalten die Wahl der Trainingssprecher, der zu klassifizierenden Sprecher, der Merkmale und der zu verwendenden Klassen. Über Buttens lassen sich diese Einstellungen für alle drei Klassifikationen separat treffen. Die rechte Seite der GUI ist

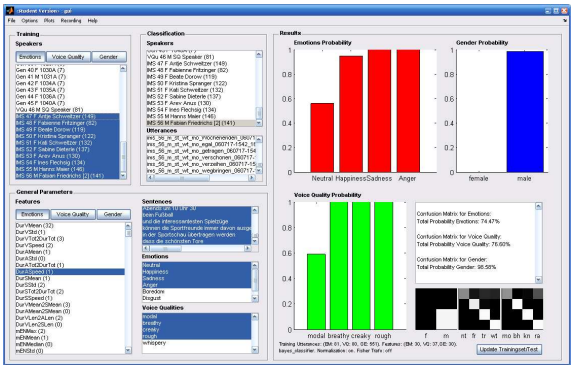

Abbildung 1: Graphische Benutzeroberfläche

für die Darstellung der Ergebnisse in Form von Balkendiagrammen und grafischer Verwechslungsmatrizen reserviert.

Merkmalsextraktion

Dem System stehen insgesamt über 300 Sprachmerkmale zur Verfügung. Diese entstammen den Merkmalsgruppen: Sprachgrundfrequenz(52), Energie(81), Dauer(16), Artikulation(35), Stimmqualität(82), Nulldurchgangsrate(16) und Cepstralkoeffizienten(90). Diese Merkmale werden alle aus dem Sprachsignal extrahiert und durch die nachfolgend beschriebenen Methoden zur Dimensionsreduktion selektiert bzw. transformiert.

Dimensionsreduktion

Im System sind verschiedene Methoden zur Dimensionsreduktion implementiert. Zur Merkmalsselektion wurde der leistungsstarke „sequential floating forward selection“(SFFS) [6] Algorithmus implementiert. Aus der Gruppe der Transformationen sind die bekannten Methoden „principle component analysis“(PCA) und „linear discriminant analysis“(LDA) [7] verfügbar. Diese können zur Vermeidung von „curse of dimensionality“ verwendet werden, um ein zur Klassifikation adäquates Merkmalsset zu finden und gute Klassifikationsergebnisse zu erhalten.

Klassifikation

Die eigentliche Klassifikation findet in zwei Phasen statt. Im Trainingsteil können die für das Training verwendeten Sprecher ausgewählt werden. Im Klassifikationsteil werden die zu klassifizierenden Sprecher gewählt. Wird die Option „speaker independent“ gesetzt, so wird automatisch eine sprecherunabhängige „leave-one-speaker-out cross-validation“ mit allen ausgewählten Sprechern durchgeführt.

Es sind eine Reihe von Klassifizierern im System implementiert. Diese können mit den entsprechenden Parametern zur Klassifikation der extrahierten und transformierten Merkmalsvektoren verwendet werden. In die Klasse

der parametrischen Klassifizierer fallen der „Bayesian“ und der „Gaussian mixture model“ Klassifizierer. In die Klasse der nichtparametrischen Klassifizierer gehören der „linear discriminant function“ Klassifizierer und das Neuronale Netz. Klassifizierer, die auf einfachen Abstandsmaßen basieren, sind der „nearest mean“ und der „k-nearest-neighbours“ Ansatz.

Weitere Funktionen des Systems

Zur Charakterisierung der Stimme gehört neben der Klassifikation auch die Ausgabe verschiedener Plots, Konturen und Verteilungen von Merkmalen und Basisgrößen. Abb.2 zeigt einige mit dem System erzeugte Plots und Konturen wie z.B. das Zeitsignal, die Sprachaktivität, die Energie, die Nulldurchgangsrate und die Sprachgrundfrequenz. Zur Erfassung der Relevanz der

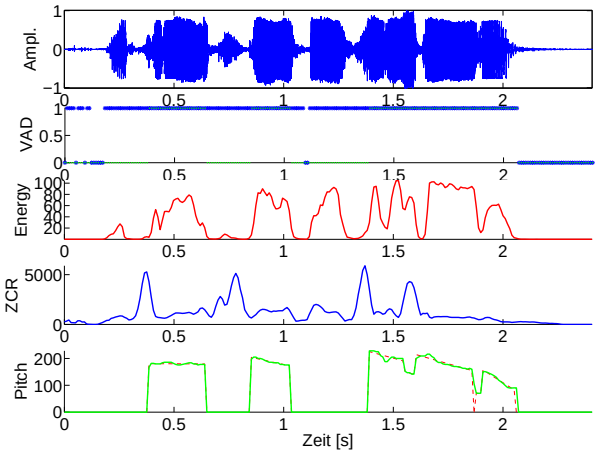

Abbildung 2: Plots und Konturen

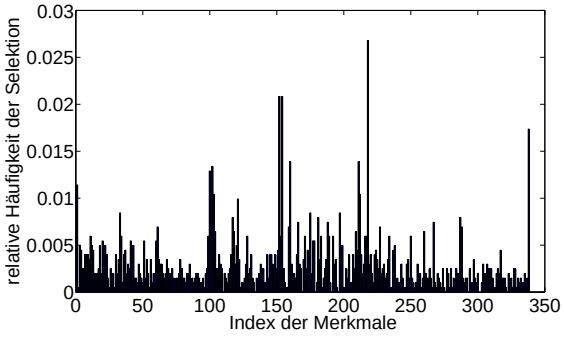

Abbildung 3: Merkmalsstatistik

einzelnen Merkmale wird die in Abb.3 dargestellte Merkmalsstatistik in Form eines Histogramms gebildet. Ein hoher Wert im Histogramm zeigt an, dass dieses Merkmal häufig selektiert wurde und es demnach eine große Relevanz für die Klassifikation besitzt. Weitere Funktionalitäten sind über verschiedene Menus in Form von Fenstern in der GUI erreichbar.

Live-Klassifikation

Die Hauptanwendung dieses System ist der Einsatz zur Live-Klassifikation. Hierzu steht dem Benutzer eine zweite graphische Benutzeroberfläche zur Verfügung. Die mit der Aufnahmeeinheit aufgezeichneten Sätze werden umgehend geschnitten und mit den eingegebenen Metadaten gelabelt. So lassen sich auch relativ komfortabel neue Datenbanken erzeugen. Aus dem Signal werden dann die selektierten Merkmale berechnet. Der Merkmalsvektor wird mit der ebenfalls gewählten Klassifikationsmethode einer Emotion, einer Stimmqualität und einem Geschlecht zugewiesen. Die Klassifikation erfolgt annähernd in Echtzeit. Alle bisher angesprochenen Funktionalitäten des Systems lassen sich auch auf die Live-Klassifikation anwenden.

Ergebnisse

In diesem Abschnitt werden einige mit dem System erarbeitete Ergebnisse zur Live-Klassifikation vorgestellt. Diese beinhalten sowohl sprecherabhängige als auch sprecherunabhängige Klassifikationsergebnisse für die Detektion des Geschlechts, der Stimmqualität und des emotionalen Zustands. Alle Klassifikationen finden auf Satzbasis statt. Die verwendeten Sprachdaten sind ausschließlich mit dem EMOsystem an unserem Lehrstuhl aufgezeichnet. Es sind insgesamt 4700 Sätze von 55 Sprechern enthalten. Bei der Klassifikation der Stimmqualität wird zwischen rauer, knarriger, behauchter und modaler Stimmgebung unterschieden. Bei der Detektion des emotionalen Zustands wird zwischen den 4 Emotionen Freude, Trauer, Wut und Neutral unterschieden. Tab.1 zeigt die durchschnittliche Erkennungsrate der verschiedenen Analysen unter Verwendung des GMM-Klassifizierers und der jeweils 15 besten Merkmale.

Tabelle 1: Klassifikationsergebnisse

	Studie		
	Geschlecht	Stimmqualität	Emotion
Spr. abh.	-	83,7%	70,9%
Spr. unabh.	94,6%	74,5%	62,3%

Literatur

[1] Michael Grimm and Kristian Kroschel, “Evaluierung von natürlichen Emotionen in Sprachsignalen,” *31. Deutsche Jahrestagung für Akustik*, 2005.

[2] Chul Min Lee and S. Narayanan, “Toward detecting emotions in spoken dialogs,” *Transaction on speech and audio processing*, vol. 13, no. 2, pp. 293–303, 2005.

[3] B. Schuller, D. Arsic, F. Wallhoff, and G. Rigoll, “Emotion recognition in the noise applying large acoustic feature sets,” *Speech Prosody, Dresden*, 2006.

[4] F. Burkhardt, M. van Ballegooy, R. Englert, and R. Huber, “An emotion-aware voice portal,” *Electronic Speech Signal Processing Conference*, 2005.

[5] Donn Morrison, Ruili Wang, and Liyanage C. De Silva, “Ensemble methods for spoken emotion recognition in call-centres,” *speech communication*, vol. 49, pp. 98–112, 2007.

[6] P. Pudil, Ferri. F., Novovicova J., and J. Kittler, “Floating search method for feature selection with nonmonotonic criterion functions,” *Pattern Recognition*, vol. 2, pp. 279–283, 1994.

[7] R.A. Fisher, “The use of multiple measurements in taxonomic problems,” *Annals of Eugenics*, vol. 7, pp. 465–475, 1936.