Erkennung negativer Emotionen in Sprachsignalen mittels Bags-of-Audio-Words

Florian Pokorny[1], Franz Graf[1], Franz Pernkopf[2]

[1] *Joanneum Research Forschungsgesellschaft mbH*
DIGITAL - Institut für Informations- und Kommunikationstechnologien
Steyrergasse 17, 8010 Graz, Österreich

[2] *Technische Universität Graz*
Institut für Signalverarbeitung und Sprachkommunikation
Inffeldgasse 16c, 8010 Graz, Österreich

Einleitung

Mitbedingt durch ein breites, potentielles Anwendungsfeld stellt die Emotionsspracherkennung heute ein großes Forschungsgebiet dar, das auf die computergestützte Erkennung menschlicher Emotionen in Sprachsignalen fokussiert und dabei mit verschiedenen Herausforderungen, wie der Findung eines geeigneten, zugrundeliegenden Trainingsmaterials konfrontiert ist. Eine Vielzahl an Studien befasste sich insbesondere mit Systemen zur Erkennung negativer Emotionen. Allerdings wurden dabei nur selten diverse Anforderungen berücksichtigt, die bei einem möglichen Einsatz unter realen Bedingungen zu beachten wären, wie z.B. Unempfindlichkeit gegenüber Störschall, Echtzeitfähigkeit, oder die Verwendung ausreichend realistischer Sprachkorpora.
Ausgehend davon wurde ein robustes, echtzeitfähiges Klassifikationssystem für die Erkennung negativer Emotionen in Sprachsignalen implementiert und offline evaluiert.

Methode

Am Eingang des Systems liegt ein Audiosignal an, das in überlappende Frames geteilt wird. Anschließend erfolgt die Extraktion eines Feature-Vektors aus jedem Frame mittels openSMILE [1]. Verwendet wurde das offizielle Feature-Set der INTERSPEECH 2009 Emotion Challenge [2], das 384 Features höherer Ordnung umfasst (12 statistische Funktionale für die Verläufe von 16 Low-Level Descriptors (LLDs) und deren erste zeitliche Ableitungen (Δ)) und in Tabelle 1 spezifiziert ist.

Tabelle 1: LLDs und Funktionale des verwendeten Feature-Sets. Zero-Crossings Rate (ZCR), Root Mean Square (RMS), Grundfrequenz (F0), Harmonics-to-Noise Ratio (HNR), Mel-Frequency Cepstral Coefficient (MFCC), Mean Square Error (MSE).

384 (16 x 2 x 12) Higher Order Features	
LLDs (16 x 2)	Functionals (12)
ZCR, ΔZCR RMS Energy, ΔRMS Energy F0, ΔF0 HNR, ΔHNR MFCCs 1-12, ΔMFCCs 1-12	mean, standard deviation, kurtosis, skewness, linear regression (offset, slope, MSE), extremes (values, relative positions, range)

Den Kern des implementierten Systems bildet das in der Emotionsspracherkennung bislang kaum verwendete und ursprünglich aus dem Bereich der Text-Analyse stammende Prinzip der Generierung von Bags-of-Audio-Words (BoAWs; vgl. [3]), das zunächst eine Transformation der extrahierten Feature-Vektoren in diskrete Symbole erfordert. Hierzu wird Vektorquantisierung (VQ) angewandt. Allgemein werden zur VQ ein Codebuch, ein Clustering-Algorithmus sowie ein Distanzmaß benötigt [4]. Aufgrund besserer Klassifikationsergebnisse gegenüber der Verwendung herkömmlicher VQ in Vorversuchen kommt im implementierten System ein hierarchisches Teilvektor-Quantisierungsverfahren, bekannt als Split Vector Quantization (SVQ; vgl. [5]), mit dem k-Means-Clustering-Algorithmus und dem Euklidischen Distanzmaß zum Einsatz. Wie in Abbildung 1 dargestellt, beruht die SVQ auf der Verwendung von unabhängigen Codebüchern für mehrere Feature-Unterräume S_1, …, S_n.

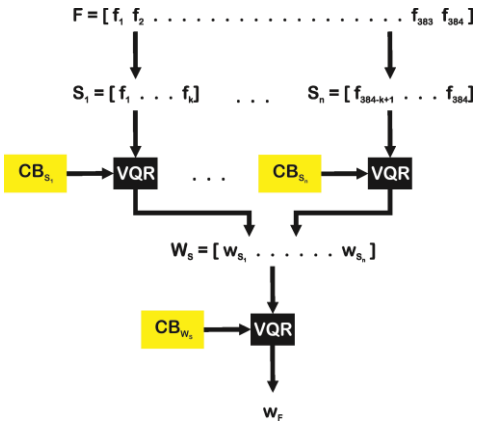

Abbildung 1: SVQ-Verfahren exemplarisch angewandt auf einen Feature-Vektor F. Codebuch (CB), Vektorquantisierer (VQR).

Für die Berechnung von BoAWs wird ein Vokabular festgelegt, das in diesem Fall alle diskreten Symbole des finalen Codebuchs der vorangegangenen SVQ-Stufe umfasst. Die mittels SVQ erzeugte Sequenz an diskreten Symbolen wird nun in überlappende Frames geteilt. Anschließend erfolgt das Mapping eines jeden Frames auf einen Vektor, der für alle Wörter des festgelegten Vokabulars in definierter Reihenfolge (hier aufsteigend) die Anzahl an Vorkommnissen im Frame kodiert (Histogramm) und jeweils einen BoAW darstellt. Abbildung 2 veranschaulicht die Generierung von BoAWs aus einer Sequenz von diskreten Symbolen, die zuvor in einer VQ-Stufe mit einem 2-Bit-Codebuch erzeugt wurde.

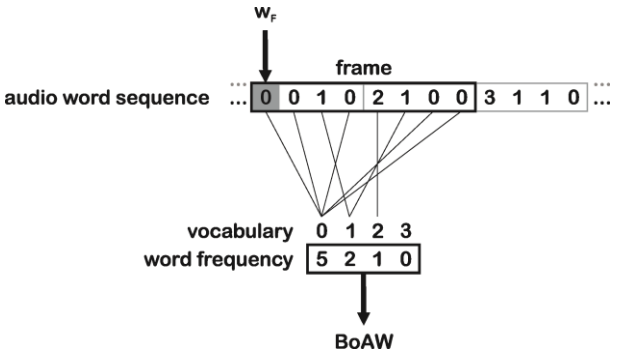

Abbildung 2: Prinzip der Generierung von BoAWs aus einer Sequenz von 2-Bit-Symbolen.

In einem letzten Schritt wird schließlich jeder BoAW einer von zwei diskreten Emotionsklassen, nämlich einer negativen, oder einer nicht-negativen zugeordnet. Dies erfolgt aufgrund der geringen Rechenleistung (und somit guten Echtzeittauglichkeit) sowie einer guten Eignung für Applikationen mit realen Datensätzen mittels Naive Bayes-Klassifikator [6].

Evaluierung

Nach intensiver Suche eines zu Training und Evaluierung des implementierten Klassifikationssystems geeigneten Sprachkorpus fiel die Entscheidung auf die Vera am Mittag German Audio-Visual Emotional Speech Database [7]. Diese umfasst 947 spontansprachliche Äußerungen von 47 unterschiedlichen SprecherInnen mit einer durchschnittlichen Länge von 3s, die aus aufgezeichneten Diskussionen zwischen Gästen der deutschen Fernseh-Talk Show Vera am Mittag extrahiert wurden. Weiters bietet der Korpus Annotationen für jede Äußerung hinsichtlich der drei Emotionsattribute Valenz, Erregung und Dominanz. Tabelle 2 zeigt die Verteilung der für Training und Evaluierung des Systems verwendeten Äußerungen (Instanzen) des Korpus nach deren Einteilung in emotional negative und nicht-negative Äußerungen anhand ihrer Annotationen für das Attribut Valenz. Um eine Sprecherunabhängigkeit gewährleisten zu können, wurden keine Äußerungen von selben SprecherInnen in sowohl Trainings- als auch Testpartition derselben Emotionsklasse zugeteilt.

Tabelle 2: Verteilung an Instanzen auf Trainings- und Test-Partitionen für die Emotionsklassen 0 (nicht negativ) und 1 (negativ) zur Evaluierung des Klassifikationssystems.

#	0	1	∑
Training	318	420	738
Test	133	76	209
∑	451	496	947

Die besten Klassifikationsergebnisse konnten erzielt werden, wenn die extrahierten Feature-Vektoren zur SVQ in drei Feature-Untervektoren mit jeweils 128 Features geteilt und 2-Bit-Untervektor-Codebücher, bzw. ein finales 5-Bit-Codebuch verwendet wurden. Diese Systemkonfiguration ergab klassen-ungewichtete und gewichtete Erkennungsraten von 64,2 % bzw. 65,6 %. Ähnliche, jedoch komplexere und daher rechenintensivere Klassifikationssysteme (z.B. aufgrund einer Klassifizierung mittels Support Vector Machines gegenüber des hier eingesetzten Naive Bayes-Klassifikators) erreichten in vergleichbaren Szenarien ungewichtete und gewichtete Erkennungsraten von maximal 67 % und 69 % (vgl. [5]).

Diskussion und Ausblick

Es konnte ein recheneffizientes und echtzeitfähiges System zur Erkennung negativer Emotionen in Sprachsignalen entwickelt und auf einem emotional gefärbten Spontansprachkorpus trainiert werden. Die erzielten Erkennungsraten lagen durchaus im Bereich vergleichbarer, jedoch komplexerer Systeme. Somit scheint auch das im Bereich der Emotionsspracherkennung innovative Prinzip der Transformation von extrahierten Feature-Vektoren in BoAWs Potenzial für akustische Klassifikationsaufgaben aufzuweisen. Aus Sicht der Implementierung ist das entwickelte System jedoch auf keine spezielle Anwendungsdomäne bzw. Klassifikationsaufgabe beschränkt. Als mögliches Ausgangssetup für verschiedene potenzielle Applikationen wäre eine praktische Installation und Evaluierung des entwickelten Systems in Szenarien unter realen Bedingungen von großem Interesse. Die Gewinnung von geeigneten Sprachdaten ausreichender Menge stellt allerdings in vielen Fällen ein großes Hindernis in der Entwicklung von Klassifikationssystemen zum Einsatz jenseits von Studiobedingungen dar und bringt, gerade im Bereich der Erkennung negativer Emotionen, meist praktische und ethische Probleme mit sich. Daher könnte vorerst die bereits verwendete Vera am Mittag German Audio-Visual Emotional Speech Database z.B. durch künstliches Überlagern der bestehenden Samples mit Alltagsgeräuschen adaptiert und so szenariospezifisch der Realität weiter angenähert werden.

Literatur

[1] Eyben, F., Wöllmer, M., Schuller, B.: openSMILE: The Munich versatile and fast open-source audio feature extractor. International Conference on Multimedia (2010), 1459-1462

[2] Schuller, B., Steidl, S., Batliner, A.: The INTERSPEECH 2009 emotion challenge. Interspeech (2009), 312-315

[3] Joachims, T. Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms. Kluwer Academic Publishers, 2002

[4] Gray, R.: Vector quantization. ASSP Magazine 1 (1984), 4-29

[5] Han, W. et al.: Towards distributed recognition of emotion from speech. International Symposium on Communications Control and Signal Processing (2012), 1-4

[6] Theodoridis, S., Koutroumbas, K. Pattern Recognition. Academic Press, 2009

[7] Grimm, M. et al.: The Vera am Mittag German audio-visual emotional speech database. International Conference on Multimedia and Expo (2008), 865-868