Evaluierung von natürlichen Emotionen in Sprachsignalen

Michael Grimm, Kristian Kroschel
Universität Karlsruhe (TH), Institut für Nachrichtentechnik, 76128 Karlsruhe, Deutschland
Email: grimm@int.uni-karlsruhe.de

Einleitung

Die Erkennung von Emotionen im Sprachsignal ist in den letzten Jahren ein wichtiger Bestandteil der Forschung im Bereich Mensch-Maschine-Schnittstelle geworden. Bisher implementierte Erkennungssysteme verwenden in der Regel Sprachproben, die von Schauspielern gespielte Emotionen enthalten [1, 2]. Da diese jedoch häufig stark übertrieben sind, verfolgen wir den Ansatz, nur natürliche Emotionen zu berücksichtigen. Diese Emotionen sind in Sprachsignalen enthalten, die von nichtprofessionellen Sprechern aufgenommen wurden und außerdem in Situationen entstanden sind, in denen die Sprecher sich nicht bewusst sind, dass ihre Sprache zu diesem Forschungszweck aufgenommen wird [3]. Bisher gibt es nur wenig Forschungsarbeit in diesem Bereich, z.B. von Douglas-Cowie *et al.* [4].
In diesem Beitrag wird die Problematik der Verwendung natürlicher Emotionen behandelt: Da die Emotion des Sprechers nicht bekannt ist, muss eine Schätzung basierend auf den Aussagen mehrerer menschlicher Evaluierer erfolgen. Neben einem geeigneten Evaluierungsverfahren wird auch eine Methode zur unterschiedlichen Gewichtung der einzelnen Evaluierer vorgestellt.

Evaluierungsmethode

In der Forschung zur Emotionserkennung werden zwei unterschiedliche Ansätze zur Beschreibung von Emotionen verfolgt: der *kategorische* und der *dimensionale Ansatz*. Während in ersterem die Beurteilung eines Emotionsausdrucks durch Auswahl eines deskriptiven Begriffs aus einer in der Regel zwei bis zehn Einträge enthaltenden Liste erfolgt, wird beim zweitgenannten, dem dimensionalen Ansatz, der Emotionsgehalt anhand mehrerer, unterschiedlicher Kriterien beurteilt. Diese Kriterien stellen die Basisdimensionen des Emotionsraumes dar.
Wir verwenden im Folgenden einen dreidimensionalen Emotionsraum mit den Basisdimensionen *Valenz*, *Aktivierung* und *Dominanz* [5]. Valenz beschreibt, wie positiv oder negativ der Emotionseindruck ist. Mittels Aktivierung wird erfasst, wie erregt oder unerregt die Emotion zum Ausdruck kommt, während Dominanz mit Werten zwischen schwach und stark das Auftreten gegenüber der adressierten Person berücksichtigt.
Eine für die Praxis sinnvolle Umsetzung dieses Prinzips stellt die Evaluierung mit den *Self Assessment Manikins* (SAMs) nach Lang dar [7]. Sie bieten für jede Basisdimension fünf diskrete Werte an, von denen der Evaluierer jeweils genau einen selektieren muss (siehe Abb. 1). Das Evaluationsergebnis ist also ein Zahlentripel, das einen Punkt im diskretisierten Emotionsraum repräsentiert.

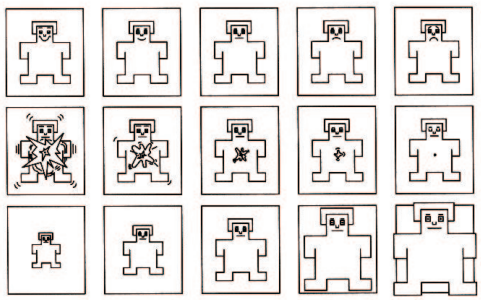

Abbildung 1: Die Self Assessment Manikins zur textfreien dimensionalen Beurteilung eines Emotionseindrucks, aus [6].

Emotionsschätzung

Das Evaluationsergebnis $\hat{\mathbf{x}}_{n,k}$ eines Evaluierers $\Upsilon_k, k = 1, \ldots, K$ für eine Sprachprobe $s_n, n = 1, \ldots, N$ kann als fehlerbehaftete Schätzung der wahren Emotion $\mathbf{x}_n$ angesehen werden. Der Fehler entsteht bei der Sprachprobengenerierung, aber auch durch den Evaluierer aufgrund des Wiedergabesystems und seiner eigenen emotionalen Verfassung.

Maximum-Likelihood-Schätzer. Das Ziel, basierend auf K Evaluiereraussagen für jede Sprachprobe den optimalen Schätzwert zu finden, der den Erwartungswert des quadratischen Fehlers minimiert, führt auf den *Maximum-Likelihood-Schätzer* (ML-Schätzer)

$$\hat{\mathbf{x}}_n^{ML} = \frac{1}{K} \sum_{k=1}^{K} \hat{\mathbf{x}}_{n,k}. \tag{1}$$

Ein Maß für die Übereinstimmung der Evaluierer stellt die Standardabweichung mit den für jede Dimension einzeln berechneten Vektorelementen $d_n = \sqrt{\mathcal{E}\{(\hat{x}_n - x_n)^2\}}$ dar [8].

Gewichteter Schätzer. Da der Erwartungswert $\mathcal{E}\{\hat{\mathbf{x}}_n^{ML}\}$ mit dem wahren Wert $\mathbf{x}_n$ übereinstimmt, stellt die Ähnlichkeit zwischen den Evaluiereraussagen $\hat{\mathbf{x}}_{n,k}$ und $\mathcal{E}\{\hat{\mathbf{x}}_n^{ML}\}$ ein Maß für die Güte des Evaluierers Υ_k dar. Als Ähnlichkeitsmaß dient der für jede Vektorkomponente separat berechnete Pearsonsche Korrelationskoeffizient r_k [8].
Mit Hilfe dieses Gütemaßes kann nun ein *Gewichteter Schätzer* $\hat{\mathbf{x}}_n^{GS}$ definiert werden, dessen Vektorelemente nach

$$\hat{x}_n^{GS} = \frac{1}{\sum_{k=1}^{K} r_k} \sum_{k=1}^{K} r_k \hat{x}_{n,k} \tag{2}$$

berechnet werden.

Ergebnisse

Datenbasis. Das vorgestellte Evaluierungsverfahren wurde an einer Sprachdatenbank mit 165 emotionalen Äußerungen einer nichtprofessionellen Sprecherin getestet. Die einzelnen Äußerungen enthalten vollständige Sätze mit bis zu 25 Wörtern. Sie wurden durch manuelle Segmentierung des *Lego-Korpus* von Kehrein [3] erstellt, das eine Aufzeichnung emotionaler Sprache im Dialog zweier sich vertrauter Menschen enthält.
Diese Sprachdatenbank wurde von $K = 13$ Evaluierern nach ihrem Emotionsgehalt mit Hilfe der SAMs beurteilt. Die Auswahl der SAMs wurde für jede Dimension der Anordnung in Abb. 1 entsprechend von links nach rechts auf eine der natürlichen Zahlen $\hat{x}_{n,k} \in \{1, \ldots, 5\}$ abgebildet.

Beurteilung der Ergebnisse. Zur Beurteilung der Datenbank wurde nun für beide Schätzer die Standardabweichung d_n bei jedem Sprachfile und für jede Dimension berechnet. Der Wert $d_n^{opt} = 0$ wird nur erreicht, wenn alle Evaluierer das gleiche SAM ausgewählt haben. Durch die Diskretisierung können jedoch auch $d_n > 0$ noch ideale Übereinstimmung beschreiben. Die maximale optimale Standardabweichung $d_{n,\max}^{opt}$ beläuft sich mit $K = 13$ auf $d_n^{opt} \leq 0{,}52$. Für den Fall geringster Übereinstimmung lässt sich die Obergrenze $d_{n,\max} = 2{,}08$ berechnen [8].
Ein Großteil der Sprachfiles wurde mit einer Standardabweichung im Bereich um $d_{n,\max}^{opt}$ evaluiert; bei manchen konnte sogar der Optimalfall $d_n = 0$ erreicht werden. Der schlechtestmögliche Fall trat dagegen bei keiner der evaluierten Äußerungen ein.
Es zeigt sich sogar, dass der über die ganze Datenbank gemittelte Wert der Standardabweichung, $\overline{d} = 0{,}7$, schon sehr nahe an die mittlere Optimalschwelle $\overline{d}_n^{\,opt} = 0{,}26$ herankommt [8].
Dieses Ergebnis ließ sich mit Einsatz des Gewichteten Schätzers noch weiter verbessern. Für die 13 Evaluierer zeigten sich mit dem Bereich $0{,}22 \leq r_k \leq 0{,}84$ sehr unterschiedliche Korrelationswerte zu dem Mittelwert. Abb. 2 zeigt die Ergebnisse für jede der drei Dimensionen.

Die größte Verbesserung ergab sich bei der Valenz-Komponente: Die Veränderung der mittleren Standardabweichung von 0,66 auf 0,58 entspricht einer relativen Verbesserung von 12,1%. Für die anderen Dimensionen fielen die Verbesserungen mit 7,3% für die Aktivierung bzw. 3,6% für die Dominanz etwas geringer aus.

Diskussion. Insgesamt ist die Kombination von Self Assessment Manikins und Schätzeinrichtung ein gut geeignetes Instrument, um natürliche Emotionen in Sprache zu evaluieren.
Das Verhalten, dass nicht alle Evaluationsergebnisse in den Bereich der Optimalschwelle fallen, zeigt eine spezifische Schwierigkeit in der Evaluation natürlicher Emotionen. Bei manchen Äußerungen in der Datenbank sind die Emotionen in Semantik und Prosodie widersprüchlich zu einander, was eine eindeutige Evaluation unmöglich macht.
Das Auftreten solch schwierig zu beurteilender Emotionen zeigt auf der anderen Seite jedoch auch die Komplexität menschlicher Emotionen. Es unterstützt die erläuterte Absicht, das gesamte Emotionsspektrum nicht durch wenige Begriffe abzudecken, sondern durch einen mehrdimensionalen Emotionsraum zu erfassen.

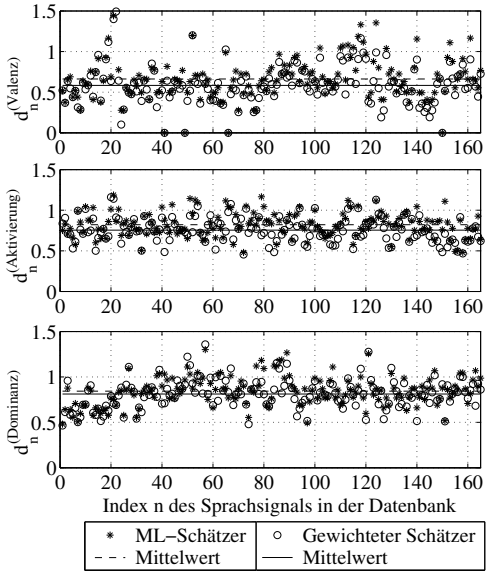

Abbildung 2: Standardabweichung des ML- und des Gewichteten Schätzers für jedes Sprachfile in der Datenbank.

Danksagung

Vielen Dank an Dr. Roland Kehrein vom Forschungsinstitut für Deutsche Sprache in Marburg für die Bereitstellung des Sprachdatenmaterials.

Literatur

[1] F. Dellaert, T. Polzin, A. Waibel, *Recognizing Emotion in Speech.* Proc. ICSLP (1996), 1970-1973

[2] S. Yildirim et al., *An Acoustic Study of Emotions Expressed in Speech.* Proc. ICSLP (2004), 2193-2196

[3] R. Kehrein, *Prosodie und Emotionen.* Max Niemeyer Verlag, Tübingen, 2002

[4] E. Douglas-Cowie, R. Cowie, M. Schröder, *The Description of Naturally Occurring Emotional Speech.* Proc. ICPhS (2003), 2877-2880

[5] R. Cowie, *Describing the Emotional States Expressed in Speech.* Speech Communications **40** (2003), 5-32

[6] L. Fischer, D. Brauns, F. Belschak, *Zur Messung von Emotionen in der angewandten Forschung.* Pabst Science Publishers, Lengerich, 2002

[7] P.J. Lang, *Behavioral Treatment and Bio-behavioral Assessment: Computer Applications,* in „Technology in Mental Health Care Delivery Systems", 119-137. Ablex Publishing, Norwood (NJ), USA, 1980

[8] M. Grimm, K. Kroschel, *Zur Eignung der Self Assessment Manikins für die Emotionsbeurteilung von Sprachsignalen.* Interner Bericht, Universität Karlsruhe, INT, 2005