Zur Eignung der Self Assessment Manikins für die Emotionsbeurteilung von Sprachsignalen

—Interner Bericht—

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, 3, 4]. 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 [5]. Bisher gibt es nur wenig Forschungsarbeit in diesem Bereich, z.B. von Douglas-Cowie *et al.* [6, 7].

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.

Im folgenden Abschnitt wird zunächst auf das Evaluierungsverfahren eingegangen und die Abbildung auf einen dreidimensionalen Merkmalsraum erläutert. Anschließend wird die Schätzung der tatsächlich vorhandenen Emotion in der Sprachprobe anhand der Merkmalsvektoren der einzelnen Evaluieraussagen betrachtet. Im letzten Abschnitt wird die Abbildung auf den Merkmalsraum mit drei emotionalen Basisdimensionen am Beispiel einer Datenbank, die 165 emotionale Sprachsignale enthält, aufgezeigt.

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.
Ein Evaluierer muss beim dimensionalen Ansatz also eine differenziertere Beurteilung abgeben; das System erhält mehr Information über den Emotionseindruck des Evaluierers. Eine spätere Reduktion auf wenige Begriffe ist durch die Definition von Klassengrenzen im Emotionsraum einfach möglich.

Wir verwenden im Folgenden einen dreidimensionalen Emotionsraum mit den Basisdimensionen

- *Valenz*,
- *Aktivierung* und
- *Dominanz*.

Dieser Ansatz wurde auch bereits an anderer Stelle zur Beschreibung von Emotionen vorgeschlagen [8, 9]. 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.
Diese Wahl der Basisdimensionen geschieht in Übereinstimmung mit einigen empirischen Studien der Emotionspsychologie und ermöglicht zusammen mit einer Angabe für *überrascht/nicht überrascht*, sämtliche Ausprägungen menschlicher Emotionen in einem kontinuierlichen Raum zu erfassen [5]. Im Gegensatz zum kategorischen Ansatz ist es so zusätzlich möglich, die Intensität einer Emotion zu erfassen. Man denke z.B. an das Kontinuum zwischen Zufriedenheit und ausgelassenem Jubeln, das im kategorischen Ansatz einheitlich mit *Freude* benannt würde.

Eine für die Praxis sinnvolle Umsetzung dieses Prinzips stellt die Evaluierung mit den *Self Assessment Manikins* (SAMs) nach Lang dar [11]. Sie bieten für jede Basisdimension fünf diskrete Werte an, von denen der Evaluierer jeweils genau einen selektieren muss. Es hat sich gezeigt, dass die Übereinstimmung verschiedener Evaluierer beim kategorischen Ansatz nur zwischen 50 und 75% beträgt, je nach Anzahl der Klassen, was neben kognitiven Aspekten hauptsächlich auf unterschiedliche Assoziationen mit den deskriptiven Emotionsbegriffen zurückzuführen ist. Um dieses Problem zu umgehen, wählen wir mit den SAMs ikonenhafte, bildliche

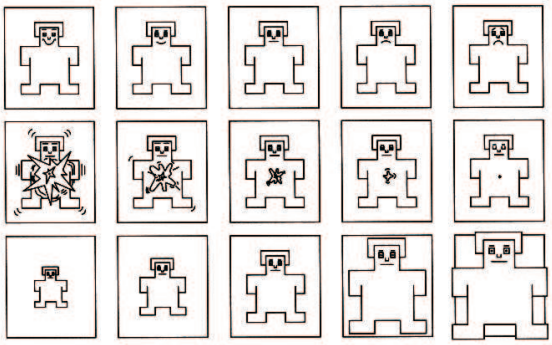

Abbildung 1: Die Self Assessment Manikins, aus [10]. Dieses Evaluationstool ermöglicht eine textfreie dimensionale Beurteilung eines Emotionseindrucks.

Darstellungen (siehe Abb. 1), die ohne linguistische Beschreibungen auskommen und dennoch dem Evaluierer eine klare Zuordnung sowohl nach Dimension als auch nach Intensität ermöglichen. Das Ergebnis einer Evaluation ist also ein Zahlentripel, das einen Punkt im endlichen, diskretisierten Emotionsraum repräsentiert.

Emotionsschätzung

Die Basis für die folgenden Berechnungen bildet die Tatsache, dass der Emotionseindruck des Evaluierers nicht unbedingt der tatsächlich vorhandenen emotionalen Verfassung des Sprechers entsprechen muss, da das Evaluationsergebnis von vielen technischen und menschlichen Faktoren beeinflusst wird. Aus diesem Grund wird das Evaluationsergebnis des Evaluierers $\Upsilon_k, k = 1, \ldots, K$ für jede Sprachprobe $s_n, n = 1, \ldots, N$ als fehlerbehaftete Schätzung $\hat{\mathbf{x}}_{n,k}$ der wahren Emotion $\mathbf{x}_n$ angesehen mit

$$\hat{\mathbf{x}}_{n,k} = \mathbf{x}_n + \mathbf{e}_{n,k}. \tag{1}$$

Der Fehlerterm $\mathbf{e}_{n,k}$ kann als Realisierung eines mittelwertfreien, gaußschen Rauschprozesses $\mathbf{E}_{n,k}$ mit der Kovarianzmatrix

$$\mathcal{E}\left\{\mathbf{E}_{n,k}\mathbf{E}_{n,k}^T\right\} = \mathbf{R}_{n,k}^{EE} = \sigma_{n,k}^2\mathbf{I} \tag{2}$$

modelliert werden.
Die Größe $\sigma_{n,k}^2$ wird zum einen durch Fehler von Seiten der Sprachprobengenerierung bestimmt (Aufnahmesystem, Umgebungsgeräusch, Datenformat, ...), zum anderen aber auch seitens des Evaluierers, der aufgrund des Wiedergabesystems, seiner eigenen emotionalen Verfassung und seiner Konzentration nicht jede Sprachprobe gleich gut bewertet.

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) [12]

$$\hat{\mathbf{x}}_n^{ML} = \frac{1}{K}\sum_{k=1}^{K}\hat{\mathbf{x}}_{n,k}. \tag{3}$$

Dies entspricht der Schätzung des Mittelwertes einer K Werte umfassenden Stichprobe. Der ML-Schätzer gewichtet jede Evaluiereraussage gleich stark; es wird kein *A priori*-Wissen verwendet.
Ein Maß für die Übereinstimmung der Evaluierer stellt die Standardabweichung $\mathbf{d}_n, n = 1, \ldots, N$ mit den für jede Dimension einzeln berechneten Vektorelementen

$$d_n = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\left(\hat{x}_{n,k} - \hat{x}_n^{ML}\right)^2} \tag{4}$$

dar. Der Mittelwert über alle $\mathbf{d}_n$ schließlich lässt eine Abschätzung zu, wie gut sich die Sprachdatenbank insgesamt evaluieren lässt.

Gewichteter Schätzer

Da der Erwartungswert $\mathcal{E}\left\{\hat{\mathbf{x}}_n^{ML}\right\}$ mit dem wahren Wert $\mathbf{x}_n$ übereinstimmt, also die Gesamtheit der Evaluierer im Mittel die wahre Emotion bestimmt, stellt die Ähnlichkeit zwischen den Evaluiereraussagen $\hat{\mathbf{x}}_{n,k}$ und $\mathcal{E}\left\{\hat{\mathbf{x}}_n^{ML}\right\}$ ein Maß für die Güte des Evaluierers Υ_k dar. Ein Schätzwert für diese Güte lässt sich berechnen, wenn man alle Aussagen eines Evaluierers Υ_k in der Folge $\{\hat{\mathbf{x}}_{n,k}\}_{n=1,\ldots,N}$ zusammengefasst betrachtet und mit dem Erwartungswerten $\{\hat{\mathbf{x}}_n^{ML}\}_{n=1,\ldots,N}$ vergleicht.
Dieser Mehrgewinn an Information über Υ_k lässt sich damit begründen, dass nicht nur *eine*, sondern N Sprachproben beurteilt werden.
Als Ähnlichkeitsmaß dient der für jede Vektorkomponente separat berechnete Pearsonsche Korrelationskoeffizient [13]

$$r_k = \frac{\sum_{n=1}^{N}\left(\hat{x}_{n,k} - \mu_k\right)\left(\hat{x}_n^{ML} - \mu^{ML}\right)}{\sqrt{\sum_{n=1}^{N}\left(\hat{x}_{n,k} - \mu_k\right)^2}\sqrt{\sum_{n=1}^{N}\left(\hat{x}_n^{ML} - \mu^{ML}\right)^2}}, \tag{5}$$

mit den über die gesamte Datenbank betrachteten Mittelwerten

$$\mu_k = \frac{1}{N}\sum_{n=1}^{N}\hat{x}_{n,k} \text{ und} \tag{6}$$

$$\mu^{ML} = \frac{1}{N}\sum_{n=1}^{N}\hat{x}_n^{ML}. \tag{7}$$

Da für einen sehr schlechten Evaluierer auch ein Wert $r_k < 0$ entstehen kann, wird für diesen Fall zusätzlich eine untere Schranke $r_k = 0$ eingeführt.
Mit Hilfe dieses Gütemaßes kann nun ein *Gewichteter Schätzer* $\hat{\mathbf{x}}_n^{GS}$ definiert werden, dessen Vektorelemente nach

$$\hat{x}_n^{GS} = \frac{1}{\sum_{k=1}^{K} r_k}\sum_{k=1}^{K} r_k\hat{x}_{n,k} \tag{8}$$

berechnet werden.
Für den Fall, dass alle Evaluierer gleich gut sind, geht der Gewichtete Schätzer in den Maximum-Likelihood-Schätzer über:

$$\hat{x}_n^{GS} \longrightarrow \hat{x}_n^{ML} \text{ für } r_1 = \ldots = r_K. \tag{9}$$

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. Die Signale sind mit 16 kHz abgetastet. Sie wurden durch manuelle Segmentierung des *Lego-Korpus* von Kehrein [14] erstellt, das eine Aufzeichnung emotionaler Sprache im Dialog zweier sich vertrauter Menschen enthält. Es handelt sich bei den auftretenden Emotionen durchweg um natürliche Emotionen, wie aus der detaillierten Beschreibung des Versuchsaufbaus [14] ersichtlich wird.
Diese Sprachdatenbank wurde von $K = 13$ Evaluierern nach ihrem Emotionsgehalt mit Hilfe der SAMs und einer Angabe *überrascht/nicht überrascht* beurteilt. Die Auswahl der SAMs wurde für jede Dimension bijektiv auf eine der natürlichen Zahlen $\{1, \ldots, 5\}$ abgebildet:

$$\textit{Auswahl des i. SAM von links in Abb. 1} \Longleftrightarrow \hat{x}_{n,k} = i. \tag{10}$$

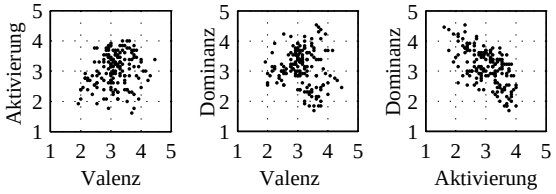

Abbildung 2: Korrelation der Basisdimensionen. Jeder Punkt entspricht einer auftretenden Kombination von SAM-Emotionskomponenten in der Datenbank.

Um einen Überblick über das Auftreten der unterschiedlichen Emotionen in der Datenbank zu bekommen, sind in Abb. 2 die Mittelwerte für die Evaluation jedes Sprachsignals in jeweils zwei Dimensionen gegeneinander aufgetragen. Es bestätigt sich damit, dass die von den SAMs erfassten Dimensionen weitgehend unkorreliert sind. Nur zwischen Aktivierung und Dominanz besteht eine etwas größere Korrelation, was sich damit erklären lässt, dass die fehlenden Kombinationen in dem verwendeten Dialog nicht aufgetreten sind.

Beurteilung der Ergebnisse

Zur Beurteilung der Datenbank wird nun für beide Schätzer die Standardabweichung d_n nach (4) berechnet. Dies wird für jedes Sprachfile und für jede Dimension separat durchgeführt. Je kleiner die Standardabweichung ist, desto größer ist die Übereinstimmung der Evaluierer bei der Beurteilung des Emotionsgehaltes.
Der Wert $d_n^{opt} = 0$ wird nur erreicht, wenn alle Evaluierer das gleiche SAM ausgewählt haben. Es ist jedoch zu berücksichtigen, dass die wahre Emotion einer beurteilten Sprachprobe nicht unbedingt mit einem der diskreten Punkte zusammenfallen muss. Da den Evaluierern jedoch nur die diskreten Werte im Emotionsraum als Auswahl zur Verfügung stehen, wird selbst bei maximaler Übereinstimmung die Standardabweichung in diesen Fällen größer als der genannte Wert $d_n^{opt} = 0$ sein. Eine obere Schranke für die maximale optimale Standardabweichung d_n^{opt} lässt sich für den Fall berechnen, dass die wahre Emotion genau zwischen zwei möglichen diskreten Werten im Emotionsraum liegt und alle Evaluierer gleichverteilt die nächstliegenden SAMs auswählen. Es gelte in diesem Fall also für die wahre Emotionskomponente $x_n = x_{n0}$:

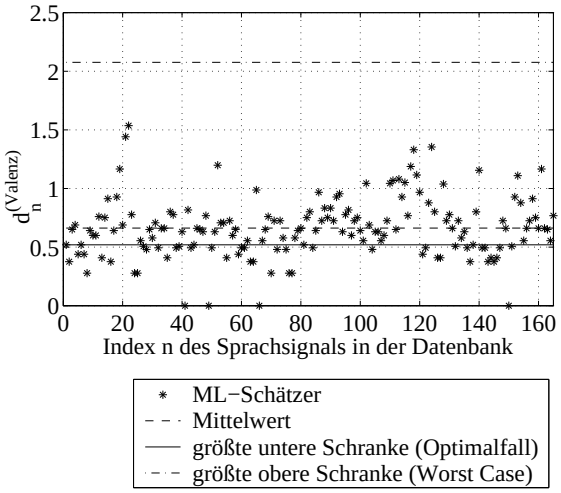

Abbildung 3: Standardabweichung der Evaluierung für jedes Sprachfile in der Datenbank (Valenz-Komponente). Der Vergleich mit den Schranken für den optimalen und den schlechtesten Fall zeigt, wie gut die Evaluierung insgesamt einzuordnen ist.

$$x_{n0} \in \{1{,}5,\ 2{,}5,\ 3{,}5,\ 4{,}5\}. \tag{11}$$

Mit einer geraden Anzahl an Evaluierern K gelte außerdem o.B.d.A.

$$\hat{x}_{n0,k} = \begin{cases} x_{n0} - 0{,}5, & k = 1, \ldots, K/2 \\ x_{n0} + 0{,}5, & k = K/2+1, \ldots, K \end{cases}. \tag{12}$$

Nach (3) und (4) gilt somit

$$\hat{x}_{n0}^{ML} = x_{n0} \ \text{ und } \ d_{n0} = \frac{1}{2}\sqrt{\frac{K}{K-1}}. \tag{13}$$

Für ungerade K lässt sich die Bedinung (12) entsprechend umformulieren. Es ergibt sich ein ähnliches Ergebnis, was sich zusammenfassen lässt zu

$$d_n^{opt} \leq \frac{1}{2}\sqrt{\frac{\xi}{\xi-1}}, \ \text{ mit } \xi = \begin{cases} K & \text{für gerade } K \\ K+1 & \text{für ungerade } K \end{cases}. \tag{14}$$

Durch diese Betrachtung ist also der Bereich definiert, in dem die Standardabweichungen einer "gut evaluierbaren" Sprachprobe liegen. Natürlich lässt ein ermittelter Wert $0 < d_n \leq \max d_n^{opt}$ keinen Umkehrschluss zu, dass eine optimale Evaluierung stattgefunden haben *muss*.

Es lässt sich auch die Standardabweichung für den schlechtestmöglichen Fall betrachten. Dieser Fall tritt ein, wenn die wahre Emotion in der Mitte der Skala liegt,

$$x_{n0} = 3, \tag{15}$$

die Evaluierer aber gleichverteilt die Skalenränder wählen. Mit K gerade gelte also o.B.d.A.

$$\hat{x}_{n0,k} = \begin{cases} 1, & k = 1, \ldots, K/2 \\ 3, & k = K/2+1, \ldots, K \end{cases} . \tag{16}$$

Nach (3) und (4) sowie der Zusammenfassung mit dem Fall K *ungerade*, der sich wieder entsprechend formulieren lässt, ergibt sich unter diesen Bedingungen die Schranke

$$d_n \leq 2\sqrt{\frac{\xi}{\xi - 1}}, \quad \text{mit } \xi = \begin{cases} K & \text{für gerade } K \\ K+1 & \text{für ungerade } K \end{cases} . \tag{17}$$

Die Zahlenwerte für die verwendete Datenbank belaufen sich mit $K = 13$ damit auf $d_n^{opt} \leq 0,52$ bzw. $d_n \leq 2,08$. Abb. 3 zeigt die Standardabweichungen für die Valenz-Komponente zusammen mit den genannten Schranken. Es lässt sich ablesen, dass ein Großteil der Sprachfiles mit einer Standardabweichung im Bereich der maximalen Optimalschwelle evaluiert wird; bei manchen konnte sogar der Optimalfall $d_n = 0$ erreicht werden. Der schlechtestmögliche Fall wird dagegen bei keiner der evaluierten Äußerungen erreicht.
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 $\overline{d}_n^{\,opt}$ herankommt. Diese mittlere Optimalschwelle lässt sich unter der Annahme der Gleichverteilung der wahren Emotionen zu

$$\overline{d}_n^{\,opt} = \frac{1}{2}\left(d_{n,\min}^{opt} + d_{n,\max}^{opt}\right) = \frac{1}{2}(0 + 0{,}52) = 0{,}26 \tag{18}$$

berechnen. Interpretiert man dieses Ergebnis mit Hilfe der SAMs, so kann man sagen, dass die Übereinstimmung der Evaluierer von der Genauigkeit her ungefähr dem halben Abstand zwischen zwei SAMs entspricht.

Dieses Ergebnis lässt sich mit Einsatz des Gewichteten Schätzers noch weiter verbessern. Für die $K = 13$ Evaluierer zeigen sich sehr unterschiedliche Korrelationswerte untereinander und zu dem Mittelwert, siehe Abb. 4. Die Gewichtungsfaktoren, die sich für die Valenz-Komponente alle im Bereich $0,22 \leq r_k \leq 0,84$ bewegen, lassen sich aus der letzten Bildspalte ablesen.

Setzt man diese Gewichtungsfaktoren in (8) ein und berechnet erneut die Standardabweichungen für jedes Sprachfile in der Datenbank, so wird ersichtlich, dass die Standardabweichung in den meisten Fällen reduziert werden konnte. Dies ist damit zu erklären, dass die Evaluierer, deren Korrelationskoeffizienten zu geringen Gewichtungen führen, ja gleichzeitig die Bewertungen abgeben, die am meisten vom Mittelwert entfert sind und so den

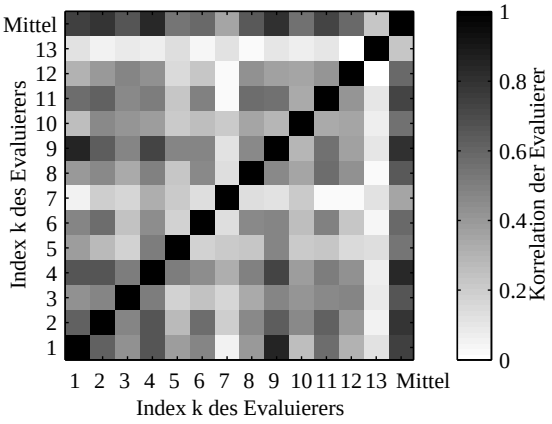

Abbildung 4: Korrelation zwischen den Aussagen der Evaluierer (bzw. deren Mittelwert) im Gesamten. Helle Bildbereiche entsprechen geringer Ähnlichkeit. Nur die Valenz-Komponente ist dargestellt.

Wert der Standardabweichung vergrößern.
Abb. 5 zeigt schließlich den Vergleich für alle drei Dimensionen. Für jede Emotionskomponente konnte die mittlere Standardabweichung reduziert werden. Die größte Verbesserung ergibt sich bei der Valenz-Komponente: Die Veränderung des Wertes von 0,66 auf 0,58 entspricht einer relativen Verbesserung von 12,1%. Für die anderen Dimensionen fallen die Verbesserungen mit 7,3% für die Aktivierung bzw. 3,6% für die Dominanz etwas geringer aus.

Diskussion

Grundsätzlich ist es wünschenswert, dass alle Standardabweichungen der Evaluationsergebnisse in den Bereich unterhalb der maximalen Optimalschwelle fallen. Es zeigt sich jedoch, dass dies bei der betrachteten Datenbank nicht der Fall ist, auch nicht bei Verwendung des gewichteten Schätzers. Dies ist darauf zurück zu führen, dass manche Äußerungen intrinsisch schlecht zu beurteilen sind. Gerade in der Valenz-Komponente ist zu berücksichtigen, dass sich die Datensätze mit hoher Standardabweichung im Sprachausdruck durch eine gewisse ironische Zweideutigkeit auszeichnen. Bei diesen Ausreißern ist die Sprecherin über einen unerfreulichen Vorgang amüsiert, indem sie von ihrer Gesprächspartnerin offensichtlich falsch verstanden wurde. Nach der Auswertung der Evaluierungsergebnisse können diese Ausreißer aus der Datenbank entfernt werden, bevor ein Erkennungssystem damit trainiert wird.

Dieses Verhalten zeigt aber auch eine spezifische Schwierigkeit in der Evaluation natürlicher Emotionen. Während bei gespielten Emotionsausdrücken die erwünschte Emotion bekannt und eindeutig ist, sind diese beiden Voraussetzungen bei natürlichen Emotionen nicht gegeben. Verwendet man für die Lösung dieses Problems die vorgestellten Schätzer, so kann die bekannte Emotionsabsicht im gespielten Fall durch die Hypothese der Schätzeinrichtungen ersetzt werden. Da bei natürlichen

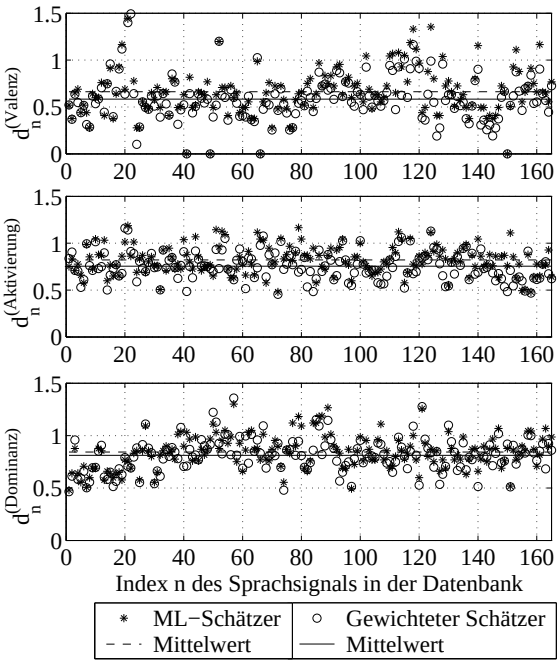

Abbildung 5: Standardabweichung des ML- und des Gewichteten Schätzers für jedes Sprachfile in der Datenbank.

Emotionen jedoch stets eine semantische Ebene im Sprachsignal enthalten ist, wird es nicht zu vermeiden sein, in einem zweiten Schritt solche Äußerungen aus der Datenbank auszusortieren, deren Emotionsgehalt in Semantik und Prosodie widersprüchlich zu einander sind.

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, natürlich immer noch stark vereinfachenden Emotionsraum zu erfassen.

Zusammenfassung

In diesem Beitrag wurde die Evaluation von natürlichen Emotionen im Sprachsignal untersucht. Es hat sich gezeigt, dass die Self Assessment Manikins ein geeignetes Evaluationswerkzeug darstellen, da sie mit den Komponenten Valenz, Aktivierung und Dominanz eine schnelle, mehrdimensionale Beurteilung eines Emotionseindrucks ermöglichen ohne auf linguistische Beschreibungen zurückgreifen zu müssen.
Mit Hilfe des Maximum-Likelihood-Schätzers bzw. des um die evaluiererspezifischen Gütemaße erweiterten Gewichteten Schätzers kann zuverlässig eine Schätzung für die einzelnen Emotionskomponenten berechnet werden. Die Ergebnisse haben gezeigt, dass sich die meisten Sprachfiles mit einer Standardabweichung evaluieren lassen, die im Bereich der Optimalschwelle liegt. Einzelne Ausreißer lassen für den nächsten Schritt der Auswertungskette zudem eine Aussage über die Komplexität der enthaltenen Emotion zu.
Insgesamt ist die Kombination von Self Assessment Manikins und Schätzeinrichtung ein gut geeignetes Instrument, um natürliche Emotionen in Sprache zu evaluieren.

In der weiteren Arbeit ist nun zu untersuchen, wie diese detaillierte, mehrdimensionale Information über den Emotionsgehalt einer natürlichen Sprachprobe ideal eingesetzt werden kann, um ein Erkennungssystem zu trainieren.

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, Philadelphia (PA), USA, Nr. 3, S. 1970-1973, 1996

[2] R.W. Picard, *Toward computers that recognize and respond to user emotion.* IBM Systems Journal, Vol 39, Nr. 3-4, S. 705-719, 2000.

[3] S. Yildirim et al., *An Acoustic Study of Emotions Expressed in Speech.* Proc. ICSLP, Jeju Island, Korea, S. 2193-2196, 2004.

[4] C.M. Lee et al., *Emotion Recognition Based on Phoneme Classes.* Proc. ICSLP, Jeju Island, Korea, S. 889-892, 2004.

[5] R. Kehrein, *Linguistische und psychologische Aspekte der Erforschung des prosodischen Emotionsausdrucks.* Germanistische Linguistik 157-158, S. 91-123, 2001

[6] E. Douglas-Cowie, R. Cowie, M. Schröder, *A New Emotion Database - Considerations, Sources and Scope.* Proc. ISCA ITRW on Speech and Emotion, Newcastle, UK, S. 39-44, 2000

[7] E. Douglas-Cowie, R. Cowie, M. Schröder, *The description of naturally occurring emotional speech.* 15th Int. Conf. of Phonetic Sciences, Barcelona, Spanien, S. 2877-2880, 2003

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

[9] M. Schröder, *Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions.* Proc. Workshop on Affective Dialogue Systems, S. 209-220, Kloster Irsee, 2004

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

[11] Lang, P.J., *Behavioral treatment and bio-behavioral assessment: Computer applications.* In J.B. Sidowski, J.H. Johnson & T.A. Williams (Hrsg.), Technology in

mental health care delivery systems, S. 119-137. Ablex Publishing, Norwood (NJ), USA, 1980

[12] K. Kroschel, *Statistische Informationstheorie.* Springer Verlag, Berlin, 2004

[13] J. Hartung, *Statistik: Lehr- und Handbuch der angewandten Statistik.* Oldenbourg Verlag, München, 2002

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