

Verbesserung eines Voiceportals mit Hilfe akustischer Klassifikation von Emotion

Richard Huber, Florian Gallwitz, Volker Warnke

Sympalog Voice Solutions GmbH
Karl-Zucker-Straße 10
91052 Erlangen
[huber,gallwitz,warnke]@sympalog.de

Abstract: Es wird ein Verfahren vorgestellt, den emotionalen Zustand eines Sprechers zu überwachen. Dabei wird für jede Äußerung bewertet, ob sich der Sprecher eher in einem neutralen oder ärgerlichen Zustand befindet. Integriert in ein automatisches Dialogsystem kann dieses Verfahren dazu beitragen, zu verhindern, dass Anrufer bei Verständnisproblemen einfach auflegen, z.B. indem das System vorher den Anruf an einen Call-Center-Agenten weiterleitet.

1 Einleitung

Die Qualität von automatischen Dialogsystemen wurde in den letzten Jahren stetig verbessert, so dass mittlerweile eine Vielzahl von Systemen kommerziell im Einsatz sind. Trotz der fortgeschrittenen Technologie arbeitet kein System völlig fehlerfrei. Es kommt immer wieder vor, dass ein System Probleme hat, einen Anrufer richtig zu verstehen und den gewünschte Service zu liefern. Stattdessen kommt es wiederholt zu Missverständnissen und als Folge davon wird der Anrufer zunehmend „genervt“ und verärgert. Im schlimmsten Fall ist der Anrufer derart verärgert, dass er einfach auflegt und das System nie wieder anrufen wird, so dass dem Betreiber ein Kunde verloren geht.

Eine Möglichkeit zu verhindern, dass der Anrufer frustriert das Gespräch beendet, ist, den emotionalen Zustand des Anrufers zu überwachen. Wird festgestellt, dass der Benutzer verärgert ist, kann das Dialogsystem speziell darauf reagieren, z.B. sich beim Benutzer dafür entschuldigen, dass die Kommunikation nicht abläuft wie gewünscht. Im Extremfall könnte das System den Anruf auch zu einem Call-Center-Mitarbeiter durchstellen, sofern dies in der vorliegenden Anwendung möglich ist.

In diesem Beitrag beschreiben wir einen Ansatz zum Bewerten des emotionalen Zustands unter Verwendung akustisch-prosodischer Merkmale und eines Normalverteilungsklassifikators (NVK).

Der Beitrag ist wie folgt gegliedert. Im nächsten Kapitel wird kurz auf die Systemarchitektur, in der das vorgestellte *Emotionsmodul* verwendet wird, eingegangen. Danach werden

die verwendeten Merkmale und die Bewertung des emotionalen Zustands beschrieben. Im nächsten Abschnitt werden die für Training und Test verwendeten Datenmengen und die erzielten Ergebnisse vorgestellt. Eine kurze Zusammenfassung beendet den Beitrag.

2 Systemarchitektur

Der in diesem Beitrag vorgestellte Ansatz zum Bewerten des emotionalen Zustands eines Anrufers ist generell anwendungsunabhängig. Die Bewertung des Zustands wird in einem eigenständigen Modul durchgeführt, wobei lediglich das Sprachsignal zu einer Äußerung als Eingabe benötigt wird. Aus dem Sprachsignal wird eine Reihe von prosodisch-akustischen Merkmalen berechnet, die dann mit einem NVK. bewertet werden. Diese Bewertung wird dann als Ergebnis zu diesem Sprachsignal geliefert (s. Abschnitt 3). Das Dialogsystem, das das Emotionsmodul verwendet, kann dann anhand der Bewertung entscheiden, welche weiteren Schritte einzuleiten sind.

Das Modul selbst ist Teil eines Voiceportals, das von der Firma T-Systems International GmbH entwickelt wurde und sich in der Pilotphase befindet. In diesem Voiceportal wird der emotionale Zustand eines Anrufers zusätzlich von einem weiteren Modul bewertet. Dieses Modul verwendet dabei eine rein semantische Analyse, z.B. sucht es nach Schimpfwörtern. Ein „Interaction Manager“ erhält dann beide Bewertungen und entscheidet, welcher Dialogschritt als Nächstes auszuführen ist [BSB05].

3 Akustisch-prosodische Klassifikation

Emotionen werden auf viele verschiedene Arten ausgedrückt, z.B. über Mimik, Gestik oder Stimme und Sprache. Beim stimmlichen bzw. sprachlichen Ausdruck von Emotionen kann man noch unterscheiden, ob der Ausdruck über prosodische Elemente, wie z.B. Änderungen in der Tonhöhe, und/oder über die Wortwahl geschieht. In der Literatur werden eine Vielzahl von verschiedenen Emotionen genannt, z.B. Ärger, Trauer, Freude [Hub02]. Da das Emotionsmodul innerhalb eines Voiceportals eingesetzt wird, beschränken wir uns auf die Unterscheidung zwischen einem *neutralen* und einem *ärgerlichen* Zustand des Anrufers, die im Folgenden mit den Klassen *N* und *A* bezeichnet werden.

Zu den wichtigsten prosodischen Ereignissen werden in der Literatur Änderungen in der Grundfrequenz, der Energie, Wort- und Lautdauern, Pausensetzung und Stimmqualität angegeben [MA93], [Fri85].

Daher werden in diesem Ansatz akustisch-prosodische Merkmale aus der Grundfrequenzkurve, der Energiekurve und Lautdauern berechnet. Der in dem Voiceportal verwendete Spracherkenner liefert als Ergebnis nur die erkannte Wortkette ohne eine zeitlichen Zuordnung der gesprochenen Wörter. Um dennoch Dauermerkmale verwenden zu können, wird das Sprachsignal einem zweiten Spracherkenner zur Verfügung gestellt, der jedoch aus Aufwandsgründen statt Wortfolgen nur Phonemfolgen mit zeitlicher Zuordnung berech-

net. Aus der Phonemfolge werden dann die Dauermerkmale berechnet. Folgende Merkmale werden berechnet:

Grundfrequenz: Minimum, Maximum, Mittelwert, Median, Standardabweichung, Mittelwert im *ersten/letzten* stimmhaften (SH) Block, Anzahl der SH-Blöcke mit *positiver/negativer* Steigung der Regressionsgeraden, Mittelwert der *positiven/negativen* Steigung der Regressionsgeraden, Maximum der *positiven/negativen* Steigung der Regressionsgeraden, Steigung der Regressionsgeraden im letzten SH-Block

Energie: Minimum, Maximum, Mittelwert, Median, Verhältnis zwischen Integral in allen SH-Frames zu den stimmlosen (SL) Frames, Mittelwert im *ersten/letzten* SH Block, Anzahl der SH-Blöcke mit *positiver/negativer* Steigung der Regressionsgeraden, Mittelwert der *positiven/negativen* Steigung der Regressionsgeraden, Maximum der *positiven/negativen* Steigung der Regressionsgeraden, Steigung der Regressionsgeraden im letzten SH-Block

Dauer: Mittlere Länge der SH-Blöcke in Frames, Verhältnis zwischen Anzahl SH und SL Frames, Anzahl SH Frames, Mittlere Länge der *Vokale/Nichtvokale*, Anzahl der Pausen, Durchschnitt der Pausenlängen, Verhältnis der Länge der Vokale zu den Nichtvokalen, Summe der Länge von Vokalen und Nichtvokalen

Insgesamt werden pro Äußerung 38 akustisch-prosodische Merkmale berechnet und zu einem Merkmalvektor zusammengefasst. Da der Ausdruck von Emotionen stark sprecherabhängig ist, werden zusätzlich zu den Absolutwerten der Merkmale sogenannte Delta-Merkmale berechnet. Ausgehend von der Annahme, dass ein Anrufer am Anfang des Dialogs mit dem Voiceportal sich in einem neutralen Zustand befindet, wird der Merkmalvektor für die erste Äußerung als Referenzvektor verwendet. Für jede weitere Äußerung wird ein Merkmalvektor und der Differenzvektor zwischen diesem Merkmalvektor und dem Referenzvektor berechnet, so dass insgesamt 76 akustisch-prosodische Merkmale vorliegen. Beim Training des NVK und dem späteren Einsatz des Emotionsmoduls kann eine Untermenge dieser 76 Merkmale verwendet werden.

Statt einen Merkmalvektor zu klassifizieren, d.h. als Ergebnis zu liefern, dass er zur Klasse N bzw. zur Klasse A gehört, werden mit Hilfe eines NVK zwei Bewertungen berechnet. Eine Bewertung für die Klasse N (S_N) und eine Bewertung für die Klasse A (S_A). Diese beiden Bewertungen werden mit folgenden Gleichungen normiert und als Ergebnis der Emotionsbewertung werden die beiden Werte S'_N und S'_A zurück geliefert.

$$S'_N = \frac{\min(S_N, S_A)}{S_N}, \quad S'_A = \frac{\min(S_N, S_A)}{S_A} \quad (1)$$

Durch diese Normierung hat immer einer der beiden Werte den Wert 1.0 und der andere einen Wert < 1 . Auf diese Weise kann der „Interaction Manager“ selbst noch Einfluss auf die Entscheidung nehmen, ob es sich um eine ärgerliche Äußerung handelt oder nicht. Im einfachsten Fall entscheidet man sich für die Klasse mit der Bewertung 1.0.

Je nach Anwendungsszenario kann es jedoch wichtiger sein, dass die neutralen Äußerungen besser erkannt werden, zu Ungunsten der Erkennung der ärgerlichen. Durch die Normierung der Bewertungen des NVK kann der „Interaction Manager“ selbst noch Schwellwerte für die eigentliche Klassifikation setzen, z.B. immer dann auf neutral zu entscheiden,

wenn $S'_N \geq 0.90$ gilt.

4 Datenmengen und Evaluation

Ein großes Problem bei der Entwicklung eines Emotionsmoduls ist das Fehlen von Daten mit „echter“ Emotion, d.h. Aufnahmen von Gesprächen, bei denen ärgerliche Äußerungen vorkommen. Für das Training des NVK wurden vier verschiedene Datensammlungen verwendet. Drei davon beinhalten ausschließlich Aufnahmen mit gespielter Emotion, d.h. die Testpersonen wurden angehalten, so zu sprechen, als ob sie verärgert sind. Die vierte Datenmenge beinhaltet Aufnahmen von Dialogen zwischen Personen und einem automatischen Dialogsystem zur Flugbuchung, die während einer Evaluierungsphase des Systems aufgenommen wurden. Alle Aufnahmen wurden über Telefon durchgeführt. Tabelle 1 zeigt die Verteilung der Trainingsmenge auf die beiden Klassen N und A . Als

Klasse	Dauer (hh:mm:ss)	# Äußerungen
N	08:46:44	9004
A	00:24:35	416
insg.	09:11:19	9420

Tabelle 1: Datenmenge für das Training des NVK.

unabhängige Testmenge wurden Aufnahmen von einem Vermittlungssystem verwendet, die während der Optimierungsphase des Systems mitgeschnitten wurden. Auch diese Aufnahmen beinhalten ausschließlich Telefondaten. Da nicht nur der Ausdruck von Emotionen, sondern auch die Annotation der Emotion stark sprecherabhängig sind, wurde die Testmenge von insgesamt sechs Personen annotiert. Beim Auswerten des Emotionsmoduls wurde einer Äußerung als Referenz immer diejenige Klasse zugeordnet, die von der Mehrzahl der Annotierer ausgewählt wurde, bei Gleichstand die Klasse N . Tabelle 2 zeigt die Aufteilung der Testmenge in die beiden Klassen N und A .

Klasse	Dauer (hh:mm:ss)	# Äußerungen
N	00:30:03	605
A	00:03:32	72
insg.	00:33:35	677

Tabelle 2: Datenmenge für die Evaluation des NVK.

Bei der Evaluation wurde jede Äußerung der Testmenge vom Emotionsmodul bewertet. Je nachdem welcher Wert von S'_N als Minimum verwendet wird, um eine Äußerung als *neutral* zu klassifizieren, ergeben sich unterschiedliche Ergebnisse. Tabelle 3 zeigt die Evaluationsergebnisse für $S'_N \geq 0.90$, $S'_N \geq 0.95$ und $S'_N = 1.00$. Wie zu erwarten war, steigt der Recall für die Klasse A , je höher der Schwellwert für N gewählt wird (von

	$S'_N \geq 0.90$		$S'_N \geq 0.95$		$S'_N = 1.00$	
Ref/Klass	#N/%	#A/%	#N/%	#A/%	#N/%	#A/%
N	593/98.0	12/ 2.0	562/92.9	43/ 7.1	500/82.6	105/17.4
A	64/88.9	8/11.1	58/80.6	14/19.4	53/73.6	19/26.4

Tabelle 3: Klassifikationsergebnisse für unterschiedliche Werte von S'_N (38 Delta-Merkmale).

11.1% auf 26.4%), wobei im Gegenzug der Recall für die Klasse N fällt (von 98% auf 82.6%). Da in der Datenmenge deutlich mehr Äußerungen der Klasse N enthalten sind (605 gegenüber 72 aus Klasse A), ist die Precision für die Klasse A deutlich geringer als für die Klasse N . Durch ein Erhöhen des Schwellwertes für die Klassifikation von N werden mehr Äußerungen als A klassifiziert (Ansteigen des Recalls für A), jedoch treten auch deutlich mehr Fehlentscheidungen für die Klasse A auf. Dadurch sinkt die Precision für die Klasse A von 40% auf 15.2%. Die Gesamterkennungsrate schwankt dabei zwischen 76.2% für $S'_N = 1.00$ und 88.8% für $S'_N = 0.90$.

5 Zusammenfassung

In diesem Beitrag wurde ein Ansatz zur Bewertung des emotionalen Zustands eines Sprechers vorgestellt, wobei nur eine Unterscheidung zwischen einem neutralen und einem ärgerlichen Zustand berücksichtigt wird. Dabei werden ausschließlich akustisch-prosodisch Merkmale verwendet, die aus dem Sprachsignal berechnet werden. Die Merkmale werden mit einem NVK bewertet. Als Trainingsmenge für den Klassifikator dienen Aufnahmen mit gespielter Emotion. Die Evaluation wurde mit Aufnahmen aus einem Vermittlungssystem realisiert. Die Vorteile einer Bewertung von neutral und ärgerlich statt einer echten Klassifizierung durch das Emotionsmodul beim Einsatz in einem Voiceportal werden in [BSB05] dargestellt.

Literatur

- [BSB05] F. Burkhardt, J. Stegmann und M. Van Ballegooy. A Voiceportal Enhanced by Semantic Processing and Affect Awareness. In *Lecture Notes in Informatics (LNI)*. Köllen Verlag, erscheint 2005.
- [Fri85] R. W. Frick. Communicating Emotion: The Role of Prosodic Features. *Psychological Bulletin*, 97:412–429, 1985.
- [Hub02] R. Huber. *Prosodisch-linguistische Klassifikation von Emotion*. Number 8 in Studien zur Mustererkennung. Logos Verlag, Berlin, 2002.
- [MA93] I.R. Murray und J.L. Arnott. Towards the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion. In *Journal of the Acoustic Society of America*, Jgg. 2, Seiten 1097–1108, 1993.