EIN DATENSET ZUR UNTERSUCHUNG EMOTIONALER SPRACHE IN KUNDENBINDUNGSDIALOGEN

Ingo Siegert[1], David Philippou-Hübner[1], Michael Tornow[1], Ralph Heinemann[1], Andreas Wendemuth[1], Kerstin Ohnemus[2], Sarah Fischer[2], Gerald Schreiber[2]

[1]*Lehrstuhl Kognitive Systeme, Otto-von-Guericke Universität Magdeburg*
[2]*davero Gruppe Erlangen*
ingo.siegert@ovgu.de

Kurzfassung: Für die sprachbasierte Emotionserkennung stellen Sprachdaten eine wichtige Ressource dar. Um optimale Klassifizierer zu trainieren, wäre es wünschenswert, wenn die Daten hochqualitativ sind, möglichst lange und reichhaltige Interaktionen umfassen und die Annotation möglichst viele Emotionen abdeckt und dabei reliabel ist. Weiterhin sollte der Datensatz zusätzliche Informationen über die Sprecher, wie Alter, Geschlecht, Dialekt oder Persönlichkeitsmerkmale enthalten.

In diesem Beitrag wird ein neuer Datensatz natürlicher emotionaler Interaktionen vorgestellt, der die meisten Punkte adressiert. Dieser Datensatz ist in einem Forschungsprojekt entstanden, dessen Ziel es ist, ein technisches System zu entwickeln, das einen Call-Center Mitarbeiter dabei unterstützen soll, situationsangemessen auf einen telefonischen Gesprächspartner mit dessen aktuellen Emotionen zu reagieren. Zusätzlich zur emotionalen Annotation wird Crosstalk noch als weiteres Interaktionsmerkmal untersucht, um sowohl das Turntaking-Verhalten als auch die Dominanz der Sprecher zu charakterisieren.

1 Einleitung

Sprachdaten stellen für die sprachbasierte Emotionserkennung eine wichtige Ressource dar. Sie ermöglichen zum einen einen empirischen Erkenntnisgewinn beobachteter Phänomene, zum anderen dienen sie auch als Datenmaterial für das Training von Modellen zur Erkennung dieser Phänomene. Damit die aufgenommenen Daten später auch für das Training von automatischen Erkennern genutzt werden können, müssen die Daten von hoher Qualität sein. Dies muss durch entsprechende Aufnahmetechnik sichergestellt werden. Weiterhin sollte sich das Aufnahmeszenario so nahe wie möglich am späteren Einsatzszenario orientieren, damit die beobachteten Phänomene auch im Anwendungsfall auftreten. Auch sollten die technischen Spezifikationen von Aufnahme und späterem Einsatz übereinstimmen, um die Daten nicht durch unerwünschte Effekte zu verfälschen.

Nach erfolgter Aufnahme ist eine Aufbereitung der Daten sowie eine Vorverarbeitung nötig. Die Datenaufbereitung überprüft die technische Korrektheit der Aufnahmen, also ob die Daten vollständig aufgenommen wurden und bestimmte Qualitätsmerkmale, wie Lautstärkepegel oder Hintergrundgeräusche eingehalten worden. In der Vorverarbeitung werden die Sprachdaten in sinnvolle logische Einheiten wie Dialoge, Turns, Sätze oder auch Wörter, Silben und einzelne Phoneme unterteilt. Danach schließt sich eine meist manuelle Informationsanreicherung der Daten an. Da für die Emotionserkennung weniger der Sprachinhalt sondern eher der Sprachausdruck von Interesse ist, wird eine emotionale Annotation der einzelnen Einheiten durchgeführt. Die aufbereiteten Daten werden anschließend zusammen mit der Transkription und Annotation

in einem Datensatz gesammelt und stehen dann sowohl für eine empirische Analyse als auch ein Training von Sprach- oder Emotionserkennern zur Verfügung.

In den letzten Jahren hat sich in der Emotionserkennung die Art der verwendeten Daten geändert. Wurden anfangs hochqualitative Aufnahmen mit von Schauspielern inszenierten Emotionsausdrücken genutzt [1], so wird heutzutage vor allem darauf Wert gelegt möglichst natürliche Emotionsausdrücke zu erhalten [7, 8]. Um hohe Erkennungsergebnisse sicherzustellen, sollten die aufgenommenen Daten hochqualitativ sein sowie lange und reichhaltige Interaktionen umfassen. Ferner sollte die Annotation möglichst viele Emotionen abdecken und dabei reliabel sein [10]. Weiterhin sollte der Datensatz zusätzliche Informationen über die Sprecher, wie Alter, Geschlecht, Dialekt und Persönlichkeitsmerkmale enthalten. Bisher sind nur wenige Datensätze verfügbar, die diese Kriterien erfüllen.

Dieser Beitrag stellt einen neuen Datensatz natürlicher emotionaler Interaktionen vor, der die meisten der zuvor erwähnten Punkte erfüllt. Der Datensatz ist in einem Forschungsprojekt entstanden, dessen Ziel es ist, ein technisches System zu entwickeln, das einen Call-Center Mitarbeiter dabei unterstützt, situationsangemessen auf die aktuelle Emotionslage eines Gesprächspartners zu reagieren.

2 Erstellung des Datensatzes

Der Datensatz wurde in einem Call-Center aufgenommen, wobei reale Telefongespräche genutzt wurden. Die Anrufe haben verschiedene Themen, von einfachen Informationsgesprächen, über Kundendatenänderungsmitteilungen bis hin zu Beschwerdeanrufen. Daher nehmen wir an, dass der emotionale Inhalt sowohl negative als auch positive Emotionen umfasst.

Um eine umfassende Analyse des emotionalen Materials zu ermöglichen, wurden vier Agenten ausgewählt, deren Gespräche tageweise aufgezeichnet wurden. Hierbei wurde der Agent audiovisuell und der Anrufer nur akustisch aufgezeichnet.

2.1 Aufnahmesetup

(a) Sicht auf Versuchsarbeitsplatz (b) Perspektivisch aus Agentensicht

Abbildung 1 - Fotos des Versuchsarbeitsplatzes aus zwei verschiedenen Perspektiven

Damit realistische Aufnahmen mit hoher Qualität möglich sind, wurde im Call-Center ein Versuchsarbeitsplatz eingerichtet, der die zu benutzende Aufnahmetechnik mit einem Agentenarbeitsplatz integriert. Der Versuchsarbeitsplatz ist etwas abseits von anderen Arbeitsplätzen, um zu verhindern, dass störende Nebengeräusche auftreten. Weiterhin kann für die Videoaufnahmen durch eine Abschirmung ein einheitlicher Hintergrund geschaffen werden und die Aufnah-

me Unbeteiligter ausgeschlossen werden. Für eine gleichmäßige und flimmerfreie Ausleuchtung des Arbeitsplatzes wurden LED-Panele genutzt. Abbildung 1 zeigt den Versuchsarbeitsplatz aus zwei verschiedenen Perspektiven. Die Aufnahmen umfassen den Sprachkanal sowohl vom Call-Center Agenten als auch vom Anrufer. Zusätzlich wird der Agent mit drei hochauflösenden Kameras gefilmt, um dessen Gestik und Mimik während des Gespräches analysieren zu können. Abbildung 2 zeigt eine schematische Darstellung des Aufnahmesetups.

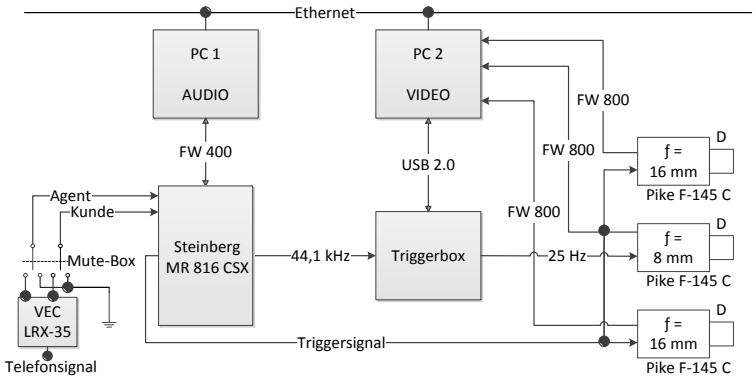


Abbildung 2 - Schematische Darstellung des Aufnahmesetups

Das Telefonsignal wird mittels des Telefon Record Adapters VEC LRX-35 direkt beim Agenten abgezweigt und zusammen mit dem hardware-synchronisierten Videosignal gespeichert. Das Audiosignal hat eine Abtastrate von 8KHz, die durch Upsampling im LRX-35 auf 44,1kHz erhöht wurde. Aufgenommen wird das Audiosignal mittels des Audio-Interfaces Steinberg MR816 CSX und dann als 32bit Wav-Datei abgespeichert.

Für die Videoaufnahmen werden für den Dauereinsatz ausgelegte Industrie-Farbkameras AVT Pike F-145C mit einer Auflösung von 1388x1038 Pixel eingesetzt. Die Videos werden als MJPEG kodiert, um so Bewegungs-Artefakte zu minimieren. Die linke und rechte Kamera haben ein 16 mm Objektiv und fokussieren auf die Mimik des Agenten. Die mittlere Kamera erfasst mit einem 8mm Objektiv sowohl Mimik als auch Gestik. Vor Beginn der Aufnahmen werden die Kameras hinsichtlich ihrer inneren und äußeren Orientierung kalibiert. Zur synchronen Aufnahme wird das 25 Hz Triggersignal in der Trigger-Box aus dem Audiotakt generiert und bei Start der Aufnahme an die Videokameras übertragen. Gleichzeitig wird das Triggersignal auch durch das Steinberg MR816 CSX synchron zu den beiden Telefonspuren aufgenommen. Dadurch kann die zeitliche Lage der Audio- und Video-Signale zueinander rekonstruiert werden.

2.2 Vorverarbeitung

Für die Erstellung des Datensatzes standen vier Agenten zur Verfügung, die durch entsprechende Kürzel pseudonomisiert wurden[1]. In Summe wurde so an 49 Tagen ein Agent jeweils über ca. 7 Stunden aufgenommen. Die aufgenommenen Telefongespräche sind reale Kundendialoge und wurden daher als Erstes anonymisiert. Alle Passagen, in denen persönliche Informationen genannt wurden, sind durch entsprechende Stille-Passagen ersetzt worden, um die Synchronität mit den Videoaufnahmen des Agenten weiterhin zu gewährleisten.

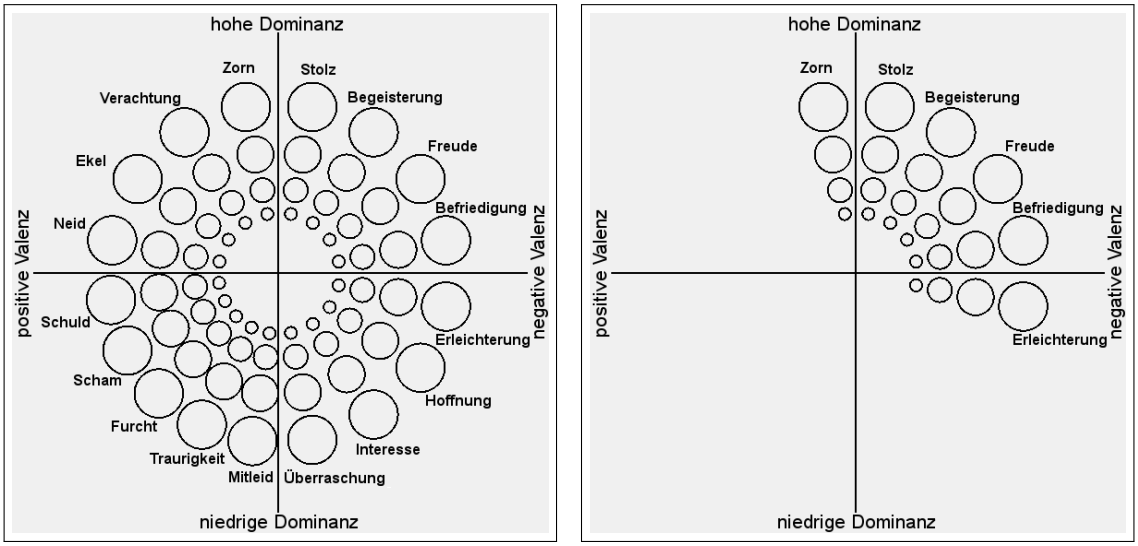
Für die weitere Auswertung ist nur die reine Dialogzeit in den Telefongesprächen von Interesse. Daher wurden auf Basis der Audioaufnahmen Start- und Endzeit der Dialoge markiert. Weiter-

[1] Da die Aufnahmen ursprünglich mit sechs Agenten geplant wurden, sind die Kürzel von A1 bis A6 vergeben worden. Während der zweimonatigen Aufnahmezeit sind jedoch zwei Agenten ausgefallen, die Kürzel aber nicht neu verteilt worden.

hin wurden auch die jeweiligen Äußerungen den Anrufern und Agenten zugeordnet und Anfang und Ende markiert. Bereiche, in denen beide gleichzeitig sprechen wurden als „Crosstalk“ markiert, um eine gesonderte Analyse durchzuführen (siehe Abschnitt 3.2).

2.3 Annotation

Nach der Unterteilung der Dialoge in einzelne Turns, werden die Äußerungen von Agent und Anrufer emotional annotiert. Die Emotionsannotation ist an das Genfer Emotionsrad von K. R. Scherer angelehnt [9]. Es ist ein empirisch getestetes Instrument zur Bewertung von Emotionen. Die für diesen Datensatz genutzte Version umfasst 17 Emotionsfamilien, die auf einem Kreis entlang der Achsen „Dominanz“ und „Valenz“ angeordnet sind, siehe Abbildung 3(a).



(a) Das komplette Genfer Emotionsrad (b) Auswahl der 6 finalen Emotionen

Abbildung 3 - Screenshots der benutzten Version des Genfer Emotionsrades von „ikannotate“

Zur Unterstützung der Annotation wird das Programm „ikannotate“ eingesetzt [2, 3]. Dieses erleichtert die Emotionsannotation durch einen dreistufigen Prozess:

- Im ersten Schritt entscheidet der Annotator ob hohe oder niedrige Dominanz vorliegt.
- Im zweiten Schritt entscheidet der Annotator zwischen positiver und negativer Valenz.
- Im dritten Schritt wird der resultierende Quadrant mit den Emotionen angezeigt und der Annotator wählt eine der Emotionsfamilien und die Intensität aus.

Dieses Vorgehen reduziert die Komplexität der Annotation. Der Annotator hat zunächst zwei entkoppelte binäre Entscheidungen zu treffen. Erst im dritten Schritt werden die Emotionen angezeigt. So braucht der Annotator nicht mehr zwischen 17 Emotionen sondern nur zwischen 6 Emotionen zu entscheiden, siehe Abbildung 3(b). Da es gerade bei Grenzfällen entlang der Achsen schwierig wird, eine Korrekte Entscheidung zu treffen, werden im Annotationsprogramm zusätzlich die jeweils zwei angrenzenden Emotionen des gewählten Quadranten angezeigt. Die emotionale Annotation wurde von vier Studenten der Psychologie durchgeführt, die zusätzlich entsprechend geschult worden.

Zusätzlich zur emotionalen Annotation wird noch der Crosstalk als ein Interaktionsmerkmal untersucht. Da angenommen wird, dass dieses einen wertvollen Beitrag zur Bewertung der Dominanz der Gesprächspartner in einer dyadischen Interaktion leisten kann [4]. „Crosstalk“ bezeichnet den Fall, wenn beide Sprecher gleichzeitig reden. Dieser wird bisher in Interaktionsstudien meist vernachlässigt.

3 Datensatzüberblick

Bisher wurden 36 der 49 Aufnahmetage anonymisiert sowie die Gesprächsanteile markiert. Dies resultiert in 1600 Dialoge mit 27.000 einzelnen Äußerungen. Wenn der Datensatz komplett bearbeitet ist, wird er einen Umfang von ca. 2.000 Dialogen mit 30.000 Äußerungen umfassen. Die Dialoge haben eine mittlere Länge von ca. 5 Minuten mit einer Standardabweichung von $\pm$ 2 Minuten. Weiterhin zeigt sich ein deutlicher Unterschied bei den Dialogen zwischen den vier Agenten, siehe Abbildung 4. Besonders Agent A2 hat eine statistisch signifikant längere Dialogzeit als die anderen Agenten.

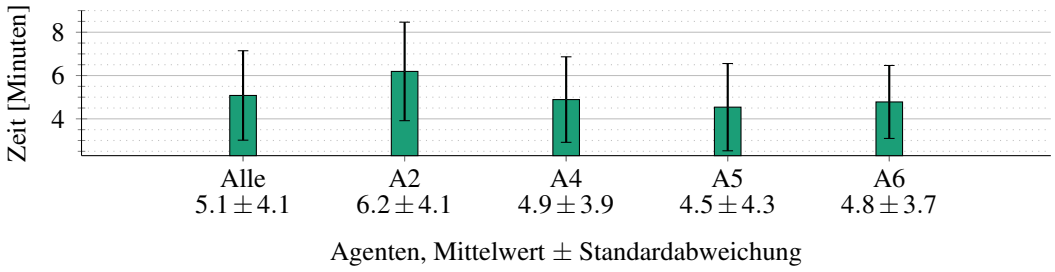


Abbildung 4 - Verteilung der Dialoglängen für die einzelnen Agenten

Bezüglich der einzelnen Äußerungen zeigt sich kein sehr großer Unterschied. Die Längen der Äußerungen sowohl des Agenten als auch des Anrufers sind relativ ähnlich. Einzig Agent A4 hat etwas längere Äußerungen, die aber mit einer hohen Standardabweichung einhergeht.

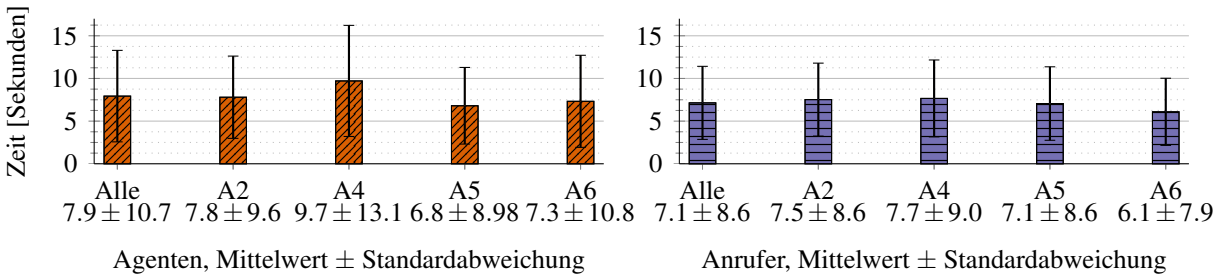


Abbildung 5 - Verteilung der Längen der Äußerungen für die einzelnen Agenten

3.1 Ergebnisse der Emotionalen Annotation

Für die emotionale Annotation stehen bisher 8 Tage mit insgesamt 5.221 Äußerungen Zur Verfügung. Da dies nur einen kleinen Teil des Datensatzes darstellt, wird in diesem Abschnitt auf eine detaillierte Analyse der einzelnen Agenten verzichtet. Aufgrund der Annonymisierung sind 160 Äußerungen unvollständig und können nicht bewertet werden.

Ein wichtiges Maß für die Verlässlichkeit der Annotation ist die „Interrater Reliabilität“. Sie misst das Ausmaß der Übereinstimmungen der annotierten Label bei unterschiedlichen Beobachtern. Hierdurch kann angegeben werden, inwieweit die Ergebnisse vom Beobachter unabhängig sind [10]. Die individuelle Erfahrung des Bewerters hat jedoch Enfluss auf die Bewertung von Emotionen, daher liegt die Reliabilität meist nur im Bereich von 0,1 bis 0,4. Zur Berechnung wurde Krippendorff’s Alpha [6] benutzt. Für die Auswertung der Annotation wurden nur diejenigen Emotionslabel berücksichtigt, die aufgrund einer echten Mehrheitsentscheidung zustanden gekommen sind, also in den Fällen, in denen mindestens 3 der Annotatoren das gleiche Label vergeben haben. Für die ersten Auswertungen des Datensatzes, haben wir uns

auf die beiden Achsen des Emotionsrades (Dominanz und Valenz), die daraus resultierenden Quadranten sowie die vergebenen Intensitätswerte konzentriert. Die Reliabilität und die Quote der Mehrheitsentscheidungen ist in Tabelle 1 angegeben.

Tabelle 1 - Errechnete Qualitätsmaße für die bisher annotierten Daten

	Dominanz	**Valenz**	**Quadranten**	**Intensität**
Reliabilität	0.19	0.34	0.23	0.22
# Mehrheitsentscheidungen [%]	73	75	48	46

Die Verteilung auf die 4 Quadranten (siehe Abbildung 6) zeigt, dass dieser Datensatz jeden der vier Quadranten abdeckt, jedoch mit einer leichten Verschiebung in Richtung negativer Valenz. Die Werte sind: 0,33% für niedrige Dominanz und negative Valenz, 0,23% für niedrige Dominanz und positive Valenz, 0,27% für hohe Dominanz und negative Valenz sowie 0,17% für hohe Dominanz und positive Valenz. Zu beachten ist hierbei jedoch, dass von allen zur Verfügung stehenden Äußerungen nur für 48% eine Mehrheitsentscheidung zustande gekommen ist.

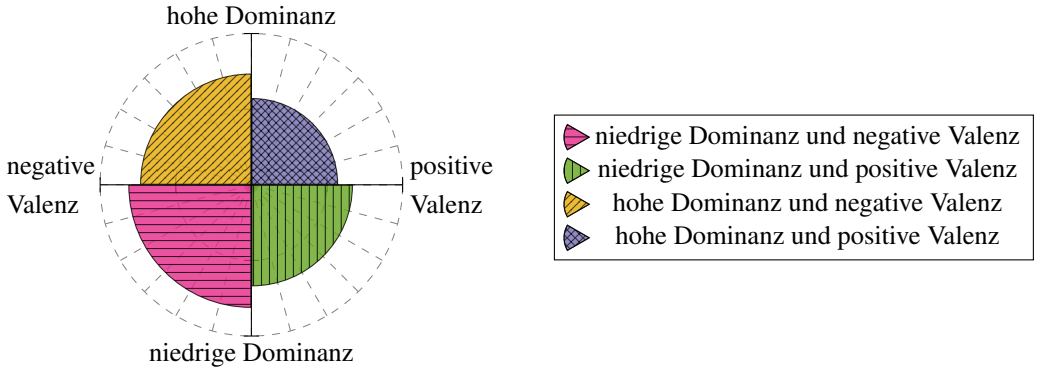


Abbildung 6 - Prozentuale Verteilung der Emotionslabel auf die vier Quadranten

Betrachtet man die Annotationsergebnisse auf beiden Dimensionen unabhängig voneinander (siehe Abbildung 7), so zeigt sich hier ein fast ausgeglichenes Bild. Bei den durch Mehrheitsentscheidung gewonnenen Labels hat hohe Dominanz einen Anteil von 59% und positive Valenz einen Anteil von 53%. Im Gegensatz zu den Quadranten ist in diesem Fall für über 70% der Äußerungen eine Mehrheitsentscheidung zustande gekommen. Bezüglich der Intensitätsverteilung ist klar zu erkennen, dass die meisten Äußerungen (72%) nur eine geringe Intensität haben und eine Intensität der Stufe 4 praktisch nicht vorkommt. Das deckt sich jedoch mit anderen Untersuchungen, die auf Call-Center Daten durchgeführt worden [5].

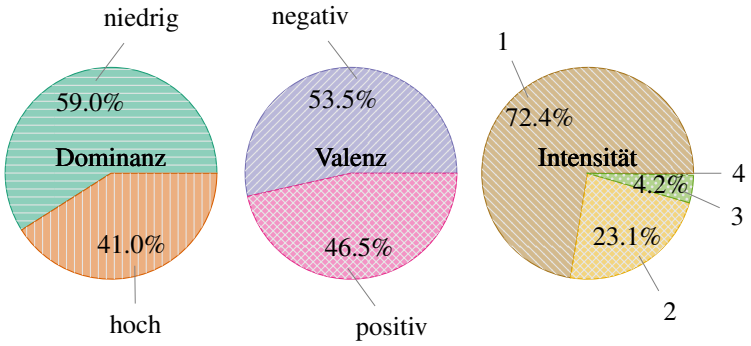


Abbildung 7 - Verteilung der Label für Dominanz und Kontrolle, sowie der Intensität

3.2 Ergebnisse der Annotation des Crosstalks

Für die akustische Analyse des Turnwechselverhaltens wird der Crosstalk untersucht. Unserer Meinung nach kann dieser viel darüber verraten, wer das dyadische Gespräch dominiert. Crosstalk tritt in fast jedem Gespräch auf, die Frage hierbei ist nur, aus welchen Gründen. Bevor ein Zusammenhang zwischen der Kontrolle im Gespräch und dem auftretenden Crosstalk untersucht werden soll, ist zunächst zu klären, ob man verschiedene Typen des Crosstalks unterscheiden muss. Hierzu wird eine Höhrerbewertung durchgeführt, die den Crosstalk in folgende vier Kriterien unterteilen soll:

1. einfaches Bestätigungssignal,
2. Zwischenrede
3. vorzeitige Turnübernahme
4. Gleichzeitiges Anfangen nach Pause

Die Bewertung der Kategorien wurde wieder mittels manueller Annotation gewonnen. Von den aktuell zur Verfügung stehenden 27.000 Äußerungen sind 5100 Äußerungen (18%) als Crosstalk markiert und entsprechend der 4 Kategorien annotiert (siehe Abbildung 8).

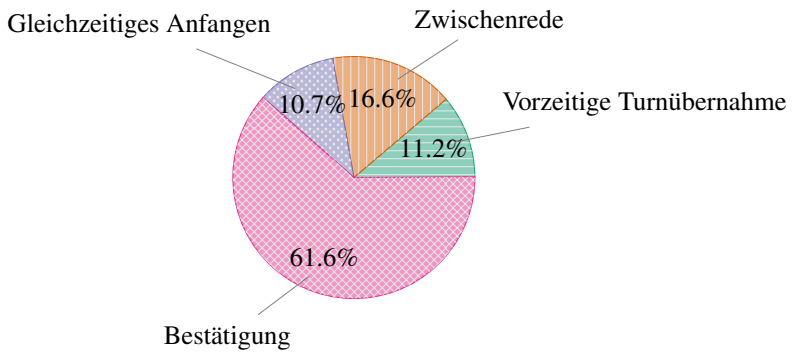


Abbildung 8 - Prozentuale Verteilung der Crosstalk-Bewertungen auf die 4 Kategorien

4 Ausblick

In diesem Beitrag wird ein neuer Typ Datensatz für die emotionale Bewertung von telefonbasierten Dialogen vorgestellt. Der Datensatz hat einen Umfang von ca. 2.000 Dialogen mit einer mittleren Dauer von 5 Minuten. Neben der Akustik wurde auch die Mimik und Gestik des Agenten durch hardware-synchronisierte Kameras aufgenommen. Für die Emotionsannotation wird das Genfer Emotionsrad von K. Scherer genutzt. Dies erlaubt sowohl eine kategoriale als auch eine dimensionale Auswertung der Emotionen. Durch die Markierung des Crosstalks und der Einteilung in 4 verschiedene Kategorien erlaubt es dieser Datensatz, auch dieses Phänomen erstmals genauer zu untersuchen.

Die Aufbereitung des in diesem Beitrag vorgestellten Datensatzes ist noch nicht abgeschlossen, gerade die Annotation der Emotionen und die sich daran anschließende Korrelation mit den Crosstalk-Kategorien ist noch geplant. Hierbei vermuten wir, dass gerade die Kategorien 2. und 3. mit einer Erhöhung der Dominanz einhergehen. Diese haben einen Anteil von ca. 28% aller Crosstalks und treten somit häufig genug auf, um eine Korrelationsanalyse durchzuführen.

Zusätzlich bietet dieser Datensatz, wenn genügend Gespräche annotiert sind, die Möglichkeit emotionale Verläufe der Gespräche zu analysieren, da zum einen mit ca. 5 Minuten Dauer ausreichend lange Gespräche vorliegen und durch die Nutzung des Genfer Emotionsrades auch ein Modell vorliegt, das eine räumliche Anordnung von Emotionskategorien kennt.

Damit die emotionale Annotation des Datensatzes mit weniger Zeitaufwand durchgeführt werden kann, ist geplant auf eine semi-automatische Annotation zurückzugreifen. Hierbei wird mit den schon bewerteten Äußerungen ein Klassifizierer trainiert und es werden nur noch die Äußerungen manuell annotiert, die unterhalb eines Schwellwerts liegen.

Literatur

[1] BURKHARDT, F., A. PAESCHKE, M. ROLFES, W. SENDLMEIER und B. WEISS: *A database of German emotional speech*. In: *Proc. of the INTERSPEECH-2005*, S. 1517–1520, Lisbon, Portugal, 2005.

[2] BÖCK, R., I. SIEGERT, M. HAASE, J. LANGE und A. WENDEMUTH: *ikannotate – A Tool for Labelling, Transcription, and Annotation of Emotionally Coloured Speech*. In: D'MELLO, S., A. GRAESSER, B. SCHULLER und J.-C. MARTIN (Hrsg.): *Affective Computing and Intelligent Interaction*, Bd. 6974 d. Reihe *LNCS*, S. 25–34. Springer, Berlin, Heidelberg, Germany, 2011.

[3] BÖCK, R., I. SIEGERT, B. VLASENKO, A. WENDEMUTH, M. HAASE und J. LANGE: *A Processing Tool for Emotionally Coloured Speech*. In: *Proc. of the 2011 IEEE ICME*, S. s.p., Barcelona, Spain, 2011.

[4] CUPERMAN, R. und W. ICKES: *Big Five Predictors of Behavior and Perceptions in Initial Dyadic Interactions: Personality Similarity Helps Extraverts and Introverts, but Hurts "Disagreeables"*. J Pers Soc Psychol, 97(4):667–684, 2009.

[5] DEVILLERS, L. und I. VIDRASCU: *Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs*. In: *Proc. of the INTERSPEECH-2006*, S. 801–804, Pittsburgh, USA, 2006.

[6] KRIPPENDORFF, K.: *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, Thousand Oaks, USA, 3 Aufl., 2012.

[7] MCKEOWN, G., M. F. VALSTAR, R. COWIE und M. PANTIC: *The SEMAINE corpus of emotionally coloured character interactions*. In: *Proc. of the 2010 IEEE ICME*, S. 1079–1084, Singapore, 2010.

[8] PRYLIPKO, D., D. RÖSNER, I. SIEGERT, S. GÜNTHER, R. FRIESEN, M. HAASE, B. VLASENKO und A. WENDEMUTH: *Analysis of significant dialog events in realistic human–computer interaction*. Journal on Multimodal User Interfaces, 8:75–86, 2014.

[9] SCHERER, K. R.: *What are emotions? And how can they be measured?*. Soc Sci Inform, 44:695–729, 2005.

[10] SIEGERT, I., R. BÖCK und A. WENDEMUTH: *Inter-Rater Reliability for Emotion Annotation in Human-Computer Interaction – Comparison and Methodological Improvements*. Journal of Multimodal User Interfaces, 8:17–28, 2014.