AUTOMATISCHE KLASSIFIKATION DES ERSTEINDRUCKS DER GESPRÄCHSWAHRNEHMUNG BEI UNGESCHULTEN HÖRERN

Mathias Walther[1], *Frances Theres Beier*[2], *Taïeb Mellouli*[1], *Baldur Neuber*[2]

[1]*Lehrstuhl für Wirtschaftsinformatik und Operations Research*
[2]*Seminar für Sprechwissenschaft und Phonetik*
[1,2]*Martin-Luther-Universität Halle-Wittenberg*
mathias.walther@wiwi.uni-halle.de

Kurzfassung: Unsere Untersuchung setzt sich zum Ziel, die Wahrnehmung ungeschulter Hörer ohne die Vorgabe von Wirkungskategorien zu analysieren. Im Experiment werden die Spontanbewertungen der Laien mittels Klassifikationsverfahren automatisch prognostiziert. Es wird untersucht, wie gut sich für einzelne Probanden die Gesprächsbewertung voraussagen lässt. Durch die auf Entscheidungsbäumen basierenden zweistufigen Klassifikationsmodelle können Bewertungen der Hörer im Zeitverlauf durch stimmliche Merkmale erklärt und begründet werden.

1 Einleitung

Die Automatisierung der qualitativen Bewertung von Callcentergesprächen ist seit mehreren Jahren ein Untersuchungsgegenstand unserer Forschungsgruppe. Die bisherige Forschung im Bereich der Callcentertelefonie konnte bereits fundierte Ergebnisse hervorbringen, die beispielsweise auf auditiven Expertenanalysen, der Messung phonetischer Parameter oder der Ermittlung von Kundeneindrücken mittels Fragebogen und Bewertungsskala basieren [1]. Der oftmals angewandte Ansatz unter Verwendung von Fragebögen schränkt die Probanden in ihrer Bewertung ein, da sie nur Aussagen zu vorgegebenen Kriterien treffen können. Häufig werden vor allem bei Untersuchungen zur automatischen Klassifikation antonyme Wertepaare oder nominale Skalen vorgegeben.

Neben der Identifikation von Bewertungskategorien ist auch die Segmentierung der Signale entscheidend für eine gute Prognosefähigkeit. In den bisher vorgestellten Modellen wird ein manuell vorsegmentiertes Korpus verwendet [2]. Dies ist aber für einen praktischen Einsatz an einem unbekannten Gespräch nicht möglich. Daher ist es Ziel der Experimente zu untersuchen, ob sich die Bewertung von Laienhörern auf Basis automatisch segmentierter Gespräche durch Klassifikationsverfahren zuverlässig erkennen lässt. Die Klassifikation erfolgt dabei mit einer Methode, die mit ihrer zweistufigen Struktur nachvollziehbare Entscheidungen treffen kann [2, 3].

2 Datenerhebung

2.1 CRDI-Verfahren

Es werden vier Gespräche aus dem Outbound-Bereich einer Krankenkasse untersucht, die von drei weiblichen sowie einem männlichen Agenten geführt werden. Die Gespräche haben eine Länge von 123 s, 127 s, 182 s und 332 s bei einer Gesamtdauer von 764 s. Bei der Bewertung der Gespräche kam ein CRDI-Gerät (Continuous Response Digital Interface) zum Einsatz, das Reaktionen auf ein dem Testhörer vorgespieltes Gespräch sofort in Echtzeit erfassen und den

Verlauf der Reaktion protokollieren kann. Beim Hören eines Gesprächs können die Testpersonen durch Drehen an einem Stellrad, dessen Position digital ausgelesen wird, den zeitlichen Verlauf einer Bewertung nachzeichnen.

Das CRDI wird bereits seit mehreren Jahren in der Musikforschung eingesetzt und erzeugt verlässliche Vergleichswerte [4]. Abbildung 1 stellt schematisch die Versuchsanordnung der

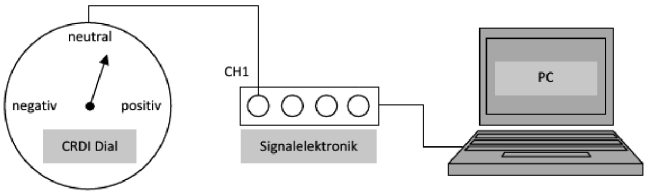

Abbildung 1 - Schematischer Aufbau des CRDI-Experiments

durchgeführten CRDI-Untersuchung dar. Sie zeigt ein Stellrad, die Signalelektronik, über welche die Verbindung zum PC hergestellt wird, und den für die Erfassung der Messwerte benötigten PC. An das Gerät können insgesamt acht Stellräder gleichzeitig angeschlossen und ausgewertet werden.

Die Gespräche wurden durch 31 Probanden bewertet. Bei der Auswahl der Teilnehmer wurde darauf geachtet, dass die Probanden gesetzlich krankenversichert und somit potentielle Kunden der Krankenkasse waren. Weiterhin sollte das Alter der Versuchsperson mindestens 28 Jahre betragen. Erfahrungen mit Callcenterkommunikation, etwa durch die eigene Krankenkasse, waren wünschenswert, jedoch nicht zwingend erforderlich.

Aufgabe der Probanden war es, während des gesamten zu hörenden Gesprächs ihre Wahrnehmung zur Wirkung des Agenten zu äußern, indem sie ihr momentanes Empfinden spontan über den manuell zu betätigenden Drehregler des CRDI einstellen. Somit bewerteten die Probanden implizit nach ihren individuellen Kriterien, die ein gutes Gespräch definieren. Der Wertebereich der aufgezeichneten Signale reicht von 0–255. Das Stellrad besitzt jedoch keine Zahlenskala, da auf die konkrete Vorgabe von Kategorien verzichtet wurde. Bei der Einweisung der Probanden wurde auf eine Einteilung in Bereiche verwiesen, die auf dem Stellrad durch „positiv“ (Zahlenwert 0), „negativ“ (Zahlenwert 255) und „neutral“ (Zahlenwert 127) gekennzeichnet waren. Die Wahl der Begriffe, welche die Wertungsbereiche festlegen, ist bewusst sehr allgemein gehalten worden, um eine Beeinflussung der Probanden auszuschließen. Durch die Messung per CRDI entsteht eine quasi-kontinuierliche Reihe von Messwerten pro Hörer und Gespräch. Die Abtastrate beträgt $1/s$, d. h. es wird jede Sekunde der aktuell eingestellte Wert gespeichert.

2.2 Gesprächsbeispiel

Für die Auswertung der erfassten Daten konnten nur 29 Probanden berücksichtigt werden, da zwei von ihnen bei einem Gespräch keine Bewertung abgaben. Abbildung 2 zeigt exemplarisch für ein ausgewähltes Gespräch die Einzelbewertung der 29 Probanden sowie den Mittelwert aller Bewertungen je Zeiteinheit und das Gesamtmittel. Wie zu erkennen ist, variieren die Bewertungen der Probanden sehr stark. Für die Analyse relevanter Gesprächsstellen ist der Verlauf des Mittelwertes von Bedeutung. Die im arithmetischen Mittel zusammengefassten einzelnen Bewertungen der Probanden weisen eine starke Glättung und geringere Ausschläge als die ursprünglichen Verläufe auf. Die Kurve des Mittelwertes zeigt drei ausgeprägte negative Ausschläge bei ca. 10 s, 60 s und 85 s sowie zwei auffällige Maxima bei 110 s und 120 s. Hieran ist die im Allgemeinen negative Einschätzung des Agenten durch die Probanden zu erkennen. Mit einem Gesamtmittelwert von 113,49 liegt die Gesprächsbewertung unterhalb einer neutralen

Bewertung von 127. Es ist zu erkennen, dass einige der Probanden nach einer gewissen Orientierungsphase zu Gesprächsbeginn eine allgemeine Einstellung zum Gespräch entwickeln. Manche Hörer bewerten demnach beispielsweise allgemein negativ und schwanken lediglich im negativen Bewertungsbereich um einen gewissen Wert. Andere Hörer kehren nach einer abgegebenen negativen Bewertung wieder in die neutrale Ausgangslage zurück.

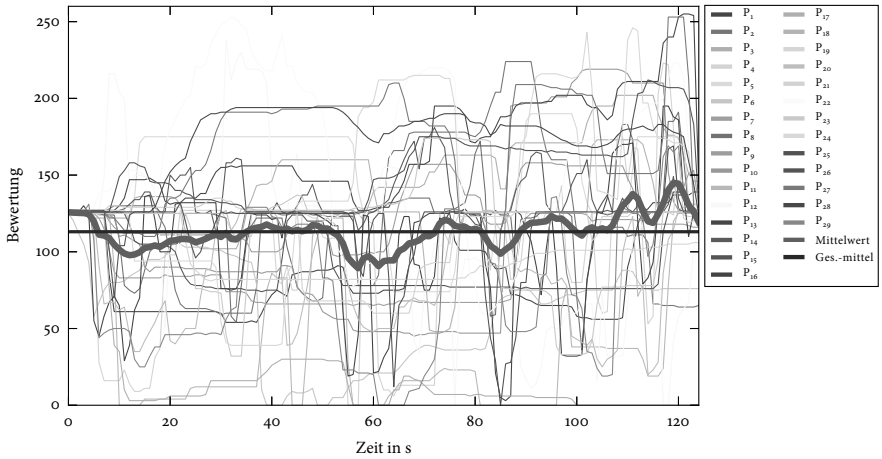

Abbildung 2 - Zeitlicher Verlauf der Bewertung eines Gesprächs durch 29 Probanden

Eine Korrelationsanalyse soll prüfen, ob die Probanden in allen Gesprächen uneinheitliche und teilweise widersprüchliche Bewertungen abgeben. Da die Korrelationsmatrix mit 29×29 Feldern zu groß für eine geeignete Darstellung ist, werden die einzelnen Pearson-Korrelationskoeffizienten r in einem Histogramm in Abbildung 3 abgetragen. Der minimale Wert für r wird mit -0,575 zwischen den Probanden P_2 und P_{20} erreicht. Der höchste Wert von 0,737 liegt beim Probendenpaar P_7–P_{17}. Es ist zu erkennen, dass die Korrelation im Mittel (Mittelwert = 0,162, Median = 0,175) leicht positiv ist. Dies bedeutet, dass sich trotz der verschiedenen Bewertungsverläufe ein kohärenter Eindruck herausbildet, der z. B. durch den Mittelwert repräsentiert werden kann.

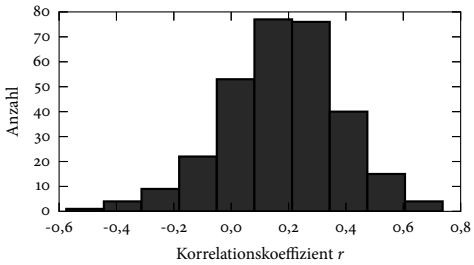

Abbildung 3 - Histogramm der Korrelationen der Probanden P_1–P_{29} über alle vier Gespräche

Das CRDI-Messverfahren liefert eine quantitative Aussage zur Sprechwirkung des Agenten auf die Laienhörer, ohne jedoch spezifische Begründungen für die jeweiligen Bewertungen zu geben. Die anhand der Mittelwertbildung identifizierten auffällig positiven oder negativen Gesprächsabschnitte wurden im Anschluss auf sprachliche Auffälligkeiten detailliert analysiert. Dieses Vorgehen soll die Ergebnisse des CRDI validieren und konkret darüber Auskunft geben,

was von den Versuchspersonen *wie* wahrgenommen wird. Dazu hörten diese jedes Gesprächsbeispiel ein zweites Mal und hatten dabei die Aufgabe, in einem verschrifteten Manuskript des Gesprächsverlaufs entsprechende Anmerkungen zum Verhalten des Agenten zu notieren. Aus der Vielzahl von gegebenen Beschreibungen wurden dann Kategorien erstellt und die meist kommentierten Gesprächsabschnitte selektiert, um diese genauer auf sprachliche Auffälligkeiten detailliert zu analysieren. Für eine umfassende Beschreibung der auditiven Analyse sei auf [5] verwiesen.

Beim Vergleich der Ergebnisse der beiden Untersuchungen konnten Übereinstimmungen zwischen den mit CRDI und den mittels Transkript erhobenen Daten festgestellt werden. Werden pro Gespräch die meist kommentierten Gesprächsstellen zu Grunde gelegt, so finden sich die gleichen Stellen mit einem Anteil von 70 % ebenfalls in der CRDI-Bewertung der Versuchspersonen wieder. Gesprächsstellen, welche die Probanden laut der CDRI-Messung positiv bewerteten, wurden gleichfalls auch bei der reflektierten Bewertung oft kommentiert und als positiv auffällig gekennzeichnet.

Die Ergebnisse der beiden Teiluntersuchungen sind somit entsprechend der gewählten Auswertung zu einem großen Anteil konform. Somit kann aus den Ergebnissen der beiden Experimente gefolgert werden, dass auch ungeschulte Hörer ihren Höreindruck konsistent bewerten können.

3 Automatische Klassifikation

3.1 Vorverarbeitung

Für die automatische Klassifikation müssen Sprachdaten segmentiert und gelabelt, d. h. mit Klassen versehen werden. Diese Teilaufgaben sind Herausforderungen für die robuste Erkennung von paralinguistischen Eigenschaften, bspw. Emotionen in Echtzeit [6]. Die Segmentierung kann auf Basis von zeitlich variablen Einheiten der Sprache (Lauten, Silben, Sätzen, etc.) oder in festgelegten zeitlichen Intervallen erfolgen. Bei kontinuierlichem Sprachstrom ist die automatische Segmentierung in konstante Zeitabschnitte eine bewährte Strategie, die durch das CRDI bereits vorgegeben ist. Die Literatur bestätigt, dass eine gute Erkennung bei einer Segmentlänge von einer Sekunde erreicht werden kann [7].

Zur Anwendung der Klassifikationsverfahren wird die von den Probanden erzeugten Messwerte von der Bewertungsskala (0–255) auf eine zweiwertige Klassenstruktur abgebildet. Wie die Analyse der Bewertungsverläufe gezeigt hat, kann angenommen werden, dass die Probanden beim Hören der Gespräche einen individuellen Wert für neutrale Gesprächsanteile bilden, die sie weder als besonders positiv noch als besonders negativ einschätzen. Dies bedeutet, dass die Bewertungen um einen Mittelwert schwanken und positive und negative Anteile als Abweichungen von diesem gesehen werden können.

Für jedes Segment der Gespräche zum Zeitpunkt t wird die Klasse k_t wie folgt ermittelt:

$$k_t = \begin{cases} -1 & \text{für } y_t \leq \bar{y} \\ +1 & \text{für } y_t > \bar{y}. \end{cases} \tag{1}$$

Der erfasste Messwert zum Zeitpunkt t wird mit y_t bezeichnet. Die Größe $\bar{y}$ ist der Mittelwert aller Bewertungen des Probanden $y = \{y_1, \ldots, y_T\}$ im Gesprächsverlauf über die Dauer $1 \ldots T$ Sekunden.

Ergebnis der Vormodellierung ist eine Instanz pro Gespräch und Bewerter und Sekunde. So stehen für die Experimente insgesamt pro Proband 764 bewertete Instanzen als Trainings- bzw. Testmenge zur Verfügung.

Zusätzlich wird ein Datensatz erstellt, der auf dem Mittelwert aller Probanden für jede Sekunde beruht. Zur Ermittlung der Klassen wird dazu Gleichung (1) wie folgt abgewandelt: y_t wird

durch $\overline{y}_t$, den Mittelwert aller Bewertungen zum Zeitpunkt t ersetzt, und $\overline{y}$ wird durch $\hat{y}$, dem Mittelwert von $\overline{y}$ über alle Testhörer ersetzt.

3.2 Einstufige Klassifikation

Für die Berechnung der Signalmerkmale wird ein mit openEAR [8] berechneter Merkmalssatz mit 2106 Merkmalen verwendet. Die Konfiguration beruht auf der für die „Paralinguistic Challenge“ der Interspeech-Konferenz 2010 genutzten Merkmalsmenge [9] und wurde um die ersten fünf Formanten sowie statistische Funktionale ergänzt. Für eine detaillierte Beschreibung sei auf [10] verwiesen. Die Modelle für das erste Experiment werden mit der Implementierung der Support-Vektor-Maschine aus Weka (SMO) trainiert [11]. Die Erkennungsgüte je Proband wurde mit einer 4-fachen Kreuzvalidierung auf Basis der Gespräche ermittelt, die eine Abwandlung der Leave-One-Corpus-Out-Strategie ist [12]. Für die Experimente werden dabei die Modelle jeweils auf den Instanzen aus drei Gesprächen trainiert. Mit den Instanzen des vierten Gesprächs wird die Validierung des Modells ausgeführt. Dieses Vorgehen hat den Vorteil, dass die Bewertung gesprächsorientiert erfolgt und daher durch Zufallseinflüsse keine Konstellationen der Instanzen auftreten, die nicht zusammen in einem Gespräch vorkommen. Die Anzahl der zum Lernen verwendeten Instanzen werden vor dem Training für jeden Datensatz und jede Konstellation der Gespräche mit Random-Undersampling angeglichen.

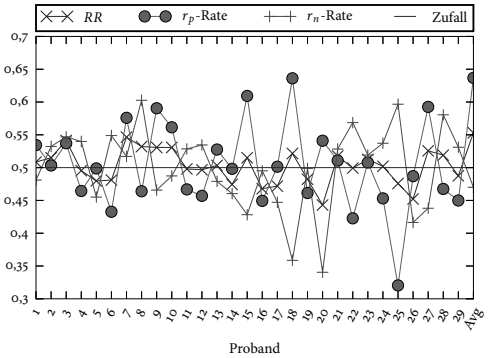

Abbildung 4 - Erkennungsleistung für Modelle der ersten Stufe

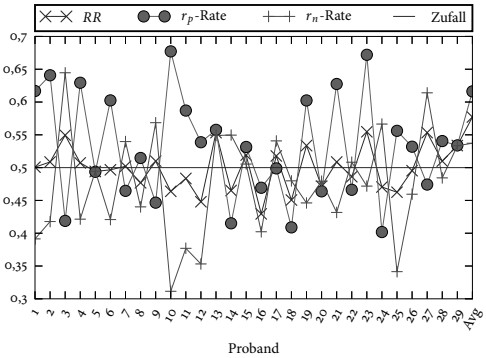

Abbildung 5 - Erkennungsleistung für Modelle der zweiten Stufe

In Abbildung 4 sind die Ergebnisse der Modelle des ersten Experiments, das direkt auf den Signalmerkmalen arbeitet, verzeichnet. Neben der Erkennungsrate *RR* sind auch die r_p-Rate und r_n-Rate eingezeichnet, die angeben wie gut die einzelnen Klassen erkannt werden. Die Klassifikationsleistungen der Modelle sind von Schwankungen um den Erwartungswert von 0,5 gekennzeichnet. Besonders auffällig ist, dass bei vielen Datensätzen die Erkennung der einzelnen Klassen unterschiedlich gut ist. Am besten ist die Klassifikationsrate mit 54 % für P_3 und P_7. Hier können beide Klassen ähnlich gut erkannt werden. Der letzte Datensatz mit den Mittelwerten hat eine hohe *RR*, aber einen schlechten Wert für r_n und kann somit nur die positive Klasse gut erkennen.

3.3 Zweistufige Klassifikation

Das zweite Experiment wurde mit den in [2] beschriebenen zweistufigen Klassifikationsmodellen durchgeführt. Deren Grundidee ist es, Klassifikationen von Gesprächsbewertungen mit

vortrainierten Modellen von Stimmeigenschaften durchzuführen. Diese Modelle sind an einem von Experten gelabelten Korpus auf die Erkennung von 13 sprecherisch-stimmlichen Merkmalen wie Sprechstimmlage, Sprechgeschwindigkeit, Melodieverlauf, Pausenverhalten, Art der Akzentuierung u. a. trainiert. Die Klassifikationsmodelle unterscheiden zwischen Gegensatzpaaren, bspw. *hoch–tief* und *schnell–langsam*, die durch geschulte Hörer auditiv erfassbar sind. Für die zweistufige Klassifikation wird zunächst jede einzelne Instanz des Datensatzes mit den auf den Signalmerkmalen operierenden Basisklassifikatoren für die 13 Kriterien des Sprechausdrucks klassifiziert. Dadurch entstehen neue Instanzen mit 13 berechneten Merkmalen und dem ursprünglichen Klassenattribut, im Beispiel +1, bzw -1. Abbildung 6 zeigt beispielhaft zwei Instanzen im ARFF-Format von Weka. Diese Transformation von der Signalrepräsentation in eine

```
@data
unangenehm, tief, leise, bewegt, terminal, stark, ungespannt, langsam,
Binnenpausen, lang, viel, temporal, wenig, -1
unangenehm, hoch, leise, monoton, interrogativ, stark, ungespannt,
langsam, Binnenpausen, lang, viel, dynamisch, wenig, +1
```

Abbildung 6 - Trainingsinstanz für das Modell der zweiten Stufe

symbolische Darstellung hat das Ziel, für die Modelle der zweiten Stufe Entscheidungsbaumverfahren nutzbar zu machen, so kann der Algorithmus J48 aus Weka verwendet werden. Anhand der Regeln der Entscheidungsbäume, die nur Sprechausdruckskriterien auswerten, können Experten die Klassifikationsentscheidungen nachvollziehen. Die erlernten Regeln können im praktischen Einsatz flexibel an neue Daten angepasst werden [13].

Die Resultate für die Erkennung der zweiten Stufe sind in Abbildung 5 markiert. Es kann festgestellt werden, dass der Schwankungsbereich zwischen r_p und r_n relativ groß ist. Tabelle 1 zeigt im Detail den Probanden mit der besten Erkennungsrate, den mit der schlechtesten und ein Beispiel für einen als durchschnittliche angesehenen Probanden. Die Bewertung von Proband P_{13} kann am besten prognostiziert werden. Hier werden beide Klassen mit jeweils ca. 55 % gut erkannt. Bei P_{10} wird mit 0,677 die höchste Erkennungsrate für eine einzelne Klasse erreicht, gleichzeitig wird aber mit 0,312 die andere Klasse sehr schlecht erkannt. Die Erkennungsleistung für die Entscheidungen von Proband P_{16} ist für beide Klassen schlecht. Die besten Ergebnisse mit $RR = 57,71\,\%$ werden mit dem Datensatz *Avg* erzielt, der die Klassenzuordnung durch den Mittelwert aller Probanden bestimmt.

Tabelle 1 - Resultate der zweiten Klassifikationsstufe bei ausgewählten Probanden

Proband	**RR**(%)	$\mathbf{r_p}$**-Rate**	$\mathbf{r_n}$**-Rate**	**Anz. Regeln**
P_{10}	46,38	0,6772	0,3115	23
P_{13}	55,34	0,5576	0,5491	30
P_{16}	42,95	0,4693	0,4022	51
Avg	57,71	0,6162	0,5372	22

Der zugehörige Entscheidungsbaum mit 22 Regeln ist in Abbildung 7 gezeigt. Die maximale Tiefe beträgt neun bei den Regeln 14 und 15. Für jede Regel kann bestimmt werden, für wie viele Instanzen sie gilt und wie viele korrekte Entscheidungen sie trifft. Aus der Betrachtung der Regeln ist ersichtlich, dass die meisten Regeln, aber insbesondere die mit großer Tiefe, für nur wenige Instanzen ($< 5\,\%$) gelten. Dagegen gilt die eher kurze Regel 18 für 302 Instanzen (40 %) und kann davon 63,2 % korrekt erkennen. Sie lässt sich wie folgt formulieren: „Monotoner Melodieverlauf *und* langsame Sprechgeschwindigkeit *und* tiefe Stimmlage *führen zu* negativer Bewertung (-1)". Anhand des Vergleichs der Regelanzahl der Bäume kann geschlossen werden, dass weniger Regeln tendenziell besser klassifizieren können.

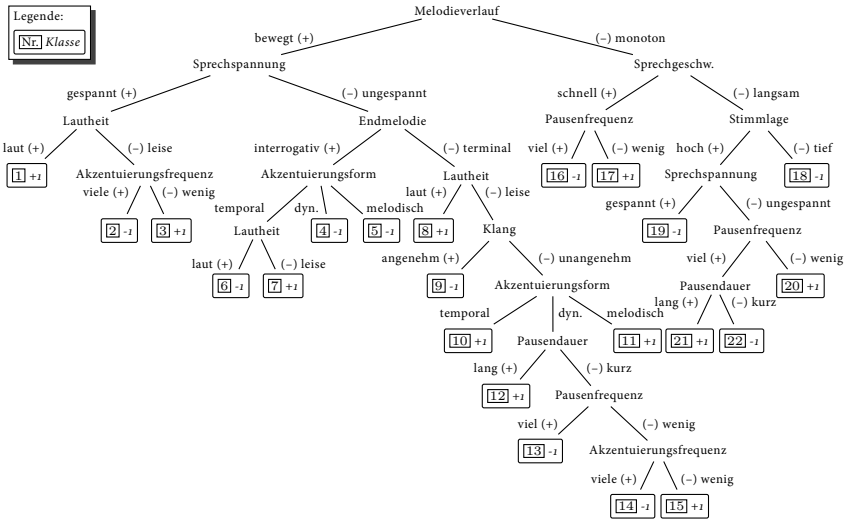

Abbildung 7 - Entscheidungsbaum der zweiten Stufe für den mittelwertbasierten Datensatz (*Avg*)

4 Fazit

Während in den bisherigen Arbeiten die Gespräche nur von Experten bewertet wurden [2], konzentriert sich diese Untersuchung darauf, spontane Eindrücke von Laienhörern zu analysieren und deren Erkennung zu automatisieren. Die Untersuchung erfolgte mit Laien, da der Eindruck des ungeschulten Hörers mit der Perspektive des Kunden im Callcentergespräch vergleichbar ist. Die erste Untersuchung hat gezeigt, dass die direkte und intuitive Bewertung von Gesprächsqualität in Callcenterdialogen durch ungeschulte Laienhörer gesprächsindividuell mit auditiven Merkmalen begründet werden kann. Durch die vom CRDI vorgegebene zeitliche Gliederung kann das Problem der automatischen Segmentierung umgangen werden, das kritisch für die Erkennung von spontaner Sprache ohne Vorverarbeitung ist. In der zweiten Untersuchung wurde festgestellt, das die sekundengenaue Segmentierung trotz der Verwendung von statischen Klassifikationsalgorithmen eine semi-dynamische Prognose von Sprechwirkung ermöglicht. Obwohl sich die Bewertungen probandenspezifisch unterscheiden, wurde bei einem mittelwertbasierten Datensatz eine Erkennungsrate erzielt, die höher ist als der statistische Erwartungswert. Weiterhin wurden zweistufige Entscheidungsbäume trainiert, die ihre Klassifikation durch für Experten verständliche stimmliche Merkmale begründen können. Diese erzielen ebenfalls Erkennungsraten von über 55 % sowohl probandenspezifisch als auch auf den Mittelwert aller Probanden bezogen.

Die gewonnenen Erkenntnisse sind ein weiterer Schritt zur Entwicklung eines Systems, das in realen Callcentergesprächen die Qualität des Dialogs erkennen kann und dem Agenten anhand von beeinflussbaren stimmlichen Merkmalen Verbesserungsmöglichkeiten zeigt.

Literatur

[1] HIRSCHFELD, U. und B. NEUBER: *Optimierungsmöglichkeiten der Telekommunikation aus Sicht der Sprechwissenschaft – Überblick über Fragestellungen und Untersuchungsansätze*. In: HIRSCHFELD, U. und B. NEUBER (Herausgeber): *Erforschung und Optimierung der Callcenterkommunikation.*, Seiten 9–28. Frank & Timme, Berlin, 2011.

[2] WALTHER, M., T MELLOULI und O. JOKISCH: *Fusion von Klassifikationsmodellen zur automatischen Erkennung von Stimmeigenschaften in der Qualitätsbewertung von Callcentergesprächen*. In: WIRSCHING, G. (Herausgeber): *Konferenzband Elektronische Sprachsignalverarbeitung (ESSV) 2015*, Band 78 der Reihe *Studientexte zur Sprachkommunikation*, Seiten 188–195, Dresden, 2015. TUDpress.

[3] WALTHER, M.: *Automatische Bewertung der Gesprächsqualität auf Basis mehrstufiger Modelle stimmlicher Merkmale – Zwischenbericht über ein Forschungsprojekt*. In: BOSE, I. und B. NEUBER (Herausgeber): *Interpersonelle Kommunikation: Analyse und Optimierung*, Band 39 der Reihe *Hallesche Schriften zur Sprechwissenschaft und Phonetik*, Seiten 313–320. Peter Lang, Frankfurt, 2011.

[4] GREGORY, D.: *The continuous response digital interface: An analysis of reliability measures*. Psychomusicology: Music, Mind and Brain, 14(1-2):197–208, 1995.

[5] BEIER, F. T.: *Untersuchung zum Ersteindruck von Sprechereigenschaften – Eine CDRI gestützte Sprechwirkungsstudie an Telefongesprächen aus dem Gesundheitswesen*. Masterarbeit, Martin-Luther-Universität Halle-Wittenberg, Seminar für Sprechwissenschaft und Phonetik, 2015.

[6] LEFTER, I., P. WIGGERS und L. J. M. ROTHKRANTZ: *EmoReSp: an online emotion recognizer based on speech*. In: *Proceedings of the 11th International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing on International Conference on Computer Systems and Technologies*, CompSysTech '10, Seiten 287–292, New York, 2010. ACM.

[7] MANSOORIZADEH, M. und N. CHARKARI: *Speech Emotion Recognition: Comparison of Speech Segmentation Approaches*. Proc of IKT, 2007.

[8] EYBEN, F., M. WÖLLMER und B. SCHULLER: *openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit*. In: *Proc. 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction 2009 (ACII 2009)*, Band I, Seiten 576–581. IEEE, 2009.

[9] SCHULLER, B., S. STEIDL, A. BATLINER, F. BURKHARDT, L. DEVILLERS, C. MUELLER und S. NARAYANAN: *The Interspeech 2010 Paralinguistic Challenge*. In: *INTERSPEECH 2010*, Seiten 2795–2798, 2010.

[10] EYBEN, F., M. WÖLLMER und B. SCHULLER: *openSMILE – the Munich open Speech and Music Interpretation by Large Space Extraction toolkit*. München, 2010. Programmdokumentation, Version 1.0.1, 23.05.2010.

[11] HALL, M., E. FRANK, G. HOLMES, B. PFAHRINGER, P. REUTEMANN und I. WITTEN: *The WEKA data mining software: an update*. SIGKDD Explor. Newsl., 11(1):10–18, 2009.

[12] ZHANG, Z., F. WENINGER, M. WOLLMER und B. SCHULLER: *Unsupervised learning in cross-corpus acoustic emotion recognition*. In: *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, Seiten 523–528. IEEE, 2011.

[13] WALTHER, M., B. NEUBER, O. JOKISCH und T. MELLOULI: *Towards a Conversational Expert System for Rhetorical and Vocal Quality Assessment in Call Center Talks*. In: *SLaTE 2015 – Sixth Workshop on Speech and Language Technology in Education*, Seiten 29–34, 2015.