Fusion von Klassifikationsmodellen zur automatischen Erkennung von Stimmeigenschaften in der Qualitätsbewertung von Callcentergesprächen

Mathias Walther[1]*, Taïeb Mellouli*[1] *und Oliver Jokisch*[2]

[1]*Institut für Wirtschaftsinformatik und Operations Research, Martin-Luther-Universität Halle-Wittenberg*
[2]*Institut für Kommunikationstechnik, Hochschule für Telekommunikation Leipzig*
mathias.walther@wiwi.uni-halle.de

Kurzfassung: Der vorliegende Beitrag diskutiert einen verbesserten Modellierungsansatz zur Erkennung der Gesprächsqualität in Callcenterdialogen durch Verfahren der Mustererkennung mit dem Ziel, ein neuartiges Fusionssystem zu entwickeln, das die menschliche Sprachwahrnehmung nachbilden kann. Das Fusionssystem verwendet zuvor trainierte Basisklassifikatoren, die Stimmattribute des Sprechers auf Basis von Sprachsignalmerkmalen erkennen können und führt diese in erklärbaren und introspektierbaren Entscheidungsbäumen auf einer sprachlichen „High Level"-Ebene zur Qualitätsbewertung zusammen.

1 Einleitung

Die Steigerung der Kundenzufriedenheit in einem Servicecenter durch Verbesserung der Gesprächsqualität ist ein zentrales Anliegen der Arbeit von Trainern und Teamleitern und wird in zunehmendem Maße von Auftraggebern gefordert. Bei Callcentergesprächen sind nach wissenschaftlichen Kriterien einige grundsätzliche Defizite erkennbar. Diese liegen u. a. in fehlendem Wissen zur Sprechwirkung und den mangelnden gesprächsrhetorischen Fähigkeiten der Callcentermitarbeiter. Es existieren wenige Forschungsarbeiten, die sich mit Faktoren der Gesprächsqualität in der professionellen Telefonie auseinandersetzen. In der Praxis werden hauptsächlich Best Practice-Methoden angewendet, die sich auf Erfahrungswerte beziehen und vorwiegend rhetorische Ziele verfolgen [1]. Eine automatisierte Gesprächsbewertung auf Rhetorik- und Ausdrucksebene ist bisher nicht möglich, da es keine einsatzbereiten Systeme gibt, die aus einem Sprachstrom Gesprächsqualitätsmerkmale zuverlässig erkennen und Hilfestellungen zur Verbesserung geben können.

Die wissenschaftliche Untersuchung dieser Problemstellung ist Ziel eines mit dem Seminar für Sprechwissenschaft und Phonetik der Universität Halle-Wittenberg und anderen Partnern durchgeführten Projekts zur Erforschung und Verbesserung der Qualität von Servicegesprächen im Callcenter. Der Fokus liegt dabei auf der sprecherischen und rhetorischen Gesprächsgestaltung durch den Agenten, da diese als Untersuchungsgegenstand sehr gut abgegrenzt werden kann und durch gezielte Trainingsmaßnahmen steuerbar ist. Eine umfassende Analyse der komplexen Interaktionsprozesse auf der Kundenseite hinsichtlich der Gesprächs- und Dialoggestaltung kann innerhalb des Projekts nicht erfolgen. Im Rahmen des Projekts werden Modellierungsansätze sowie Modelle zur automatischen Messung und Bewertung der stimmlichen und rhetorischen Gesprächsqualität des Agenten entwickelt, die in diesem Artikel vorgestellt werden. Die Grundannahme folgt der These, dass sich Gesprächsqualität durch perzeptive Merkmale beschreiben und sowohl von menschlichen Experten als auch durch Klassifikationssysteme erkennen und in erklärbaren Modellen darstellen lässt.

2 Sprachkorpus

Die Experimente wurden mit einem umfangreichen Korpus durchgeführt, das im Rahmen des Projekts nach sprechwissenschaftlichen Anforderungen erstellt wurde. Die Annotation des Korpus erfolgte anhand eines Kriterienkatalogs, der das Ergebnis der ersten sprechwissenschaftlichen Untersuchungen war [2]. Der Kriterienkatalog enthält, wie in Abbildung 1 dargestellt, sechs Gesprächsqualitätsfaktoren: Gesprächsgestaltung, Emotionalität, Verständlichkeit, Gesprächspartnerorientierung, Persönlichkeit/Authentizität und Situationsangemessenheit [2]. Im

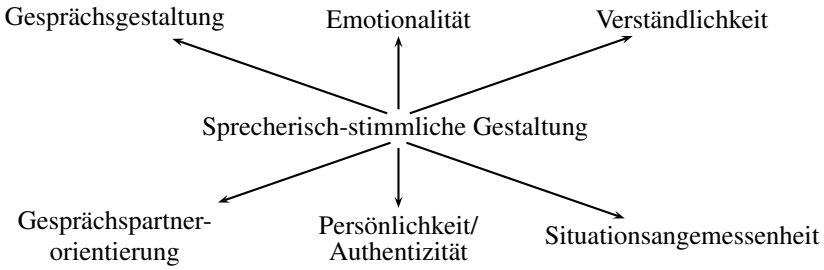

Abbildung 1 - Faktoren der Gesprächsqualität (nach [2])

Zentrum der Betrachtung steht die sprecherisch-stimmliche Gestaltung, da diese – qualitativ sowie z. T. durch akustische Korrelate messbar – entscheidend verantwortlich für die Sprechwirkung ist. Die Qualitätsmerkmale und die stimmliche Gestaltung wurden in Bewertungskriterien unterteilt. Die Stimmmerkmale umfassen dabei u. a. die Bewertungskriterien Stimmhöhe, Lautheit, Stimmklang, Sprechgeschwindigkeit und Artikulation sowie die Merkmalskomplexe Akzentuierung, Gliederung, Rhythmus sowie Sprechspannung. Die Merkmale werden an dieser Stelle gemäß der sprechwissenschaftlichen Terminologie als „Sprechausdruck“ bezeichnet [3].

Basis des Korpus sind reale Verkaufsgespräche aus drei Outbound-Kampagnen, die von verschiedenen Callcentern zur Verfügung gestellt wurden [4]. Für die Experimente wurden die Kanäle getrennt und nur der Kanal des Agenten berücksichtigt. Da eine vollständige Annotation der Gespräche aufgrund der begrenzten Ressourcen nicht möglich war, wurden aus mehreren Hundert Gesprächen im Ausgangsbestand 218 für das endgültige Korpus ausgewählt, die als repräsentativ bezüglich der sprechwissenschaftlichen Kriterien galten [2]. Die Gespräche wurden im Anschluss von vier Studenten der Sprechwissenschaft hinsichtlich Auffälligkeiten nach dem in [4] und [2] vorgestellten Katalog mit einer sechswertigen Skala annotiert. Für die beschriebenen Experimente wurden diese ordinalskalierten Daten auf zwei bzw. drei Klassen transformiert, um Klassifikationsalgorithmen anwenden zu können. Weiterhin wurde mit Random-Downsampling eine Gleichverteilung der Klassen pro Kriterium künstlich erzeugt.

Aus den in der ersten Version des Kataloges erfassten 48 Bewertungskriterien werden für die hier beschriebenen Experimente 19 Kriterien verwendet. Neben den 13 Kriterien des Sprechausdrucks, die in der Tabelle 1 aufgeführt sind, werden auch 6 wichtige Kriterien der Gesprächsqualität gemäß Tabelle 2 verwendet. Für jedes Kriterium sind in der zweiten Spalte die Klassen sowie die Gesamtzahl der Instanzen je Klasse angegeben. Abgesehen von der Akzentuierungsform mit drei Klassen variieren die Kriterien in zwei Klassen.

3 Klassifikationsmodelle auf Signalebene

Für die Berechnung der Signalmerkmale wird ein mit openEAR [5] berechneter Merkmalssatz mit 2106 Merkmalen verwendet. Die Konfiguration beruht auf der für die „Paralinguistic Challenge“ der Interspeech-Konferenz 2010 genutzten Merkmalsmenge [6] und wurde um die ersten fünf Formanten sowie statistische Funktionale ergänzt. Für eine detaillierte Beschreibung

sei auf [7] verwiesen. Zusätzlich zu den Signalmerkmalen wurde das Geschlecht des Agenten manuell erfasst.

Die Klassifikation erfolgt mit acht häufig verwendeten Algorithmen in ihren Implementierungsvarianten von Weka [8]: Naive-Bayes (NB), Bayes-Netz (BN), Logistic-Model-Tree (LMT), Ripper (JRip), Support-Vektormaschine (SMO), Ada-Boost (Ada), C4.5-Entscheidungsbaum (J48) und Multilayer-Perzeptron (MLP). Algorithmische Details werden in [9] und [10] diskutiert. Als Maß für die Erkennungsleistung wurde die mittlere Erkennungsrate (Recognition Rate, RR) mit Hilfe einer zehnfachen Kreuzvalidierung ermittelt.

Die Tabellen 1 und 2 zeigen im Überblick, absteigend nach der Erkennungsrate sortiert, die besten Klassifikationsalgorithmen bezüglich der ausgewählten Kriterien.

Wie Tabelle 1 zu entnehmen ist, können alle dichotomen Kriterien des Sprechausdrucks mit einer mittleren Erkennungsrate von über 65 % zuverlässig klassifiziert werden. Die besten Erkennungsraten erzielen Lautheit und Sprechstimmlage. Hier war die gute Erkennung zu erwarten, da diese Merkmale durch ihre akustischen Korrelate Grundfrequenz, Dauer bzw. Intensität messbar sind [11]. Die Kriterien Pausenart, Melodiesprung und Akzentuierungsfrequenz können ebenfalls gut erkannt werden. Der Datenumfang ist jedoch gering, so dass die erzielten Ergebnisse nicht als valide angesehen werden können. Für Akzentuierungsform wird eine Erkennungsrate von 48 % erreicht, die deutlich über dem Erwartungswert von 33 % bei drei Klassen liegt.

Die in Tabelle 2 aufgeführten Erkennungsleistungen für die Gesprächsqualität liegen zwischen 55 % und 78 % und sind somit besser als der statistische Erwartungswert von 50 %. Die Support-Vektormaschine (SMO) erreicht bei Kompetenz die höchste Erkennungsrate von ca. 78 %, was den siebtbesten Wert aller 13 Kriterien darstellt. Im Mittel lassen sich die Qualitätsfaktoren jedoch schlechter als der Sprechausdruck erkennen. Die Ergebnisse zeigen weiterhin, dass alle Klassifikationsalgorithmen bei mindestens einem Kriterium die maximale Erkennungsrate liefern. Das Multilayer-Perzeptron kann in sechs Fällen die maximalen mittleren Erkennungsraten für die untersuchten Daten erreichen, gefolgt vom Bayes-Netz mit vier ersten Rankingplätzen. Somit kann ad hoc ohne statistische Analysen kein global bester Klassifikationsalgorithmus identifiziert werden.

Aufgrund unterschiedlicher Vorgehensweisen bei den Korpora, Merkmalssätzen u. a. ist der direkte Vergleich mit Arbeiten zur paralinguistischen Sprachverarbeitung schwierig und kann nur grob durchgeführt werden. Bei der „Interspeech Speaker Trait Challenge 2012“ [12] wurden beispielsweise als Benchmark für Persönlichkeitseigenschaften des Sprechers (OCEAN bzw. Big-Five) ca. 70 % Erkennungsrate angegeben. Diese Werte können für Kompetenz, Sicherheit und Freundlichkeit in den hier vorliegenden Experimenten ebenfalls erreicht werden (siehe Tabelle 2).

4 Entwicklung eines Fusionssystems

Neben der guten Klassifikationsrate, die bei einigen Kriterien erzielt wird, ist für das beschriebene Anwendungsszenario von entscheidender Bedeutung, Klassifikationsentscheidungen erklären zu können. Von allen Gesprächsqualitätsmerkmalen wird Kompetenz durch eine Support-Vektormaschine am besten erkannt. Jedoch besitzt dieser Klassifikationsalgorithmus im Vergleich zu Entscheidungsbaumverfahren keine gute Erklärungsfähigkeit. Generell ist die Erklärungsfähigkeit von Modellen der Signalebene aufgrund der großen Merkmalsvektoren relativ gering. Ferner zeigen die beiden Ergebnistabellen, dass sich Stimmmerkmale im Sprechausdruck im Mittel besser erkennen lassen als Persönlichkeitseigenschaften bzw. Gesprächsqualität.

Gestützt durch diese Erkenntnisse wurde ein Fusionssystem entwickelt, das gute Erkennbarkeit

Tabelle 1 - Übersicht über die Erkennungsgüte für Kriterien des Sprechausdrucks

Kriterium	**Klassen (Anzahl Instanzen)**	**Algorithmus**	**RR (%)**
Lautheit	laut (57), leise (57)	J48	96,89
Sprechstimmlage	hoch (146), tief (146)	MLP	94,84
Pausenart	Grenzpausen (9), Binnenpausen (9)	LMT	95,00
Melodiesprung	stark (6), schwach (6)	NB	90,00
Pausenfrequenz	viel (37), wenig (37)	BN	86,43
Endmelodieverlauf	terminal (57), interrogativ (57)	SMO	78,79
Tonhöhenverlauf	bewegt (120), monoton (120)	MLP	77,08
Pausendauer	lang (45), kurz (44)	MLP	75,42
Akzentuierungsfrequenz	viele Akzente (8), wenig Akzente (8)	MLP	75,00
Sprechgeschwindigkeit	schnell (44), langsam(44)	MLP	75,00
Sprechspannung	gespannt (98), ungespannt (98)	BN	69,97
Stimmklang	angenehm (74), unangenehm (74)	BN	65,43
Akzentuierungsform	dynamisch (71), temporal (71), melodisch (71)	Ada	48,00

Tabelle 2 - Übersicht über die Erkennungsgüte für Kriterien der Gesprächsqualität

Kriterium	**Klassen (Anzahl Instanzen)**	**Algorithmus**	**RR (%)**
Kompetenz	inkompetent (71), kompetent (71)	SMO	78,76
Sicherheit	unsicher (55), sicher(55)	MLP	72,73
Freundlichkeit	freundlich (73), unfreundlich (73)	NB	67,95
Natürlichkeit	unnatürlich (140), natürlich (140)	JRip	64,29
Glaubwürdigkeit	unglaubwürdig (85), glaubwürdig (85)	BN	58,24
Kooperativität	unkooperativ (82), kooperativ (82)	SMO	54,78

der Stimmmerkmale nutzt und mit diesen in einer mehrstufigen Struktur auf einer High Level-Ebene die wahrgenommene Kompetenz als Gesprächsqualitätskriterium sowohl hinreichend gut bewerten als auch erklären kann.

Generell betrachtet sind Fusionssysteme Klassifikationssysteme zur Mustererkennung, die auf der Kombination von mehreren unterschiedlichen Teil-Klassifikationssystemen basieren [13]. Fusionssysteme bestehen aus mehreren eigenständig arbeitenden Teilsystemen, sogenannten Basisklassifikatoren, die getrennt voneinander Teilschritte der Mustererkennung ausführen. Die Teilsysteme können entweder verschiedene Instanzen des gleichen Klassifikationsalgorithmus sein oder aus unterschiedlichen Algorithmen aufgebaut werden [13]. Abbildung 2 zeigt die Struktur der Fusionsmodelle am Beispiel des Kriteriums Kompetenz.

Bei der konkreten Implementierung wird zunächst jede einzelne Instanz der Klasse Kompetenz mit den auf den Signalmerkmalen operierenden Basisklassifikatoren für die 13 Kriterien des Sprechausdrucks klassifiziert. Dadurch entsteht eine neue Instanz mit 13 berechneten Merkmalen und dem ursprünglichen Klassenattribut, im Beispiel „kompetent“. Abbildung 3 zeigt beispielhaft und in Anlehnung an das arff-Format von Weka eine solche Instanz.

Der zweite Schritt ist die Erstellung eines Klassifikationsmodells auf Basis der neuen Instanzen. Generell können alle Klassifikationsverfahren angewendet werden. Da die Nachvollziehbarkeit eine zentrale Anforderung an diese Modelle ist, werden aktuell Bäume mit dem C4.5-Algorithmus automatisch gelernt [14].

Abbildung 4 zeigt einen trainierten Entscheidungsbaum für Kompetenz. Bei den High Level-

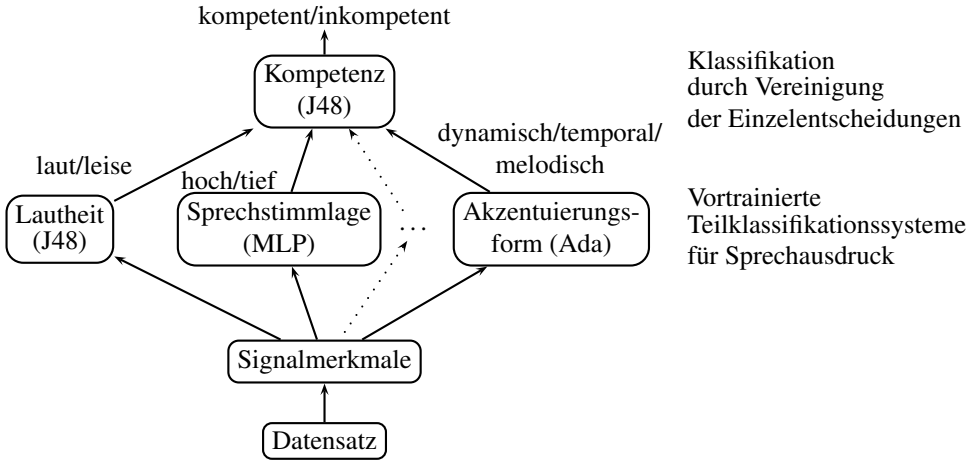

Abbildung 2 - Struktur eines Fusionsmodells am Beispiel des Qualitätskriteriums Kompetenz

```
[...]
@data
laut, tief, Binnenpausen, stark, viel, interrogativ, bewegt,
kurz, viele, schnell, gespannt, unangenehm, dynamisch, kompetent
[...]
```

Abbildung 3 - Trainingsinstanz für den Kompetenzklassifikator auf Fusionsebene

Modellen der Fusionsebene werden ebenfalls mit zehnfacher Kreuzvalidierung Performanzmaße ermittelt, mit denen die Klassifikationsgüte abgeschätzt werden kann. Die beschriebenen Modelle für Kompetenz erreichen eine Erkennungsrate von 73,2 %. Gegenüber den Basismodellen sind die Fusionssyteme somit im Mittel 8 % schlechter. Sie sind jedoch introspektierbar und können nachvollziehbare Klassifikationsentscheidungen treffen. Entscheidungsbäume ermöglichen durch ihre deklarative Struktur eine Bewertung von einzelnen Entscheidungsregeln. Im Baum aus der Abbildung 4 ist ersichtlich, dass Regel 8 nur auf den Attributen Pausenlänge und Pausenart beruht, deren Basisklassifikatoren zwar zahlenmäßig eine gute Leistung erbringen, aber nur auf kleinen Trainingsmengen basieren. Im weiteren Vorgehen wird untersucht, wie die einzelnen Regeln die Klassifikationsleistung des Gesamtbaums beeinflussen.

5 Diskussion

Die mit dem Fusionssystem für „Kompetenz“ erreichte Erkennungsrate zeigt eine grundsätzliche Realisierbarkeit von Verfahren zur automatischen Erkennung und Erklärung der Gesprächsqualität in realen Callcentergesprächen.

Die mit Standardlernverfahren erstellten Entscheidungsbäume liefern allerdings etwas schlechtere sowie weniger robuste (instabile) Klassifikationsleistungen im Vergleich zu den einstufigen Klassifikationsmodellen und bedürfen einer weiteren Optimierung. Bei Änderungen der Trainingsdaten wird damit u. a. die Baumstruktur verändert. Im Rahmen einer ersten Analyse wurde bereits ein Entscheidungsbaum mit 12 Sprechausdrucksmerkmalen veröffentlicht, der eine andere Baumstruktur zeigt, aber eine ähnliche Klassifikationsleistung erzielt [15]. Unterschiedliche Baumstrukturen und damit unterschiedliche Klassifikationsregeln können so interpretiert werden, dass die gefundenen Regeln nicht universell sind, sondern von mehreren Faktoren beeinflusst werden – z. B. durch das Korpus oder die Auswahlkriterien für Trainungs- und Testdaten.

Bei den Fusionsmodellen auf der Ebene der Gesprächsqualität wird die Änderung der Trainings-

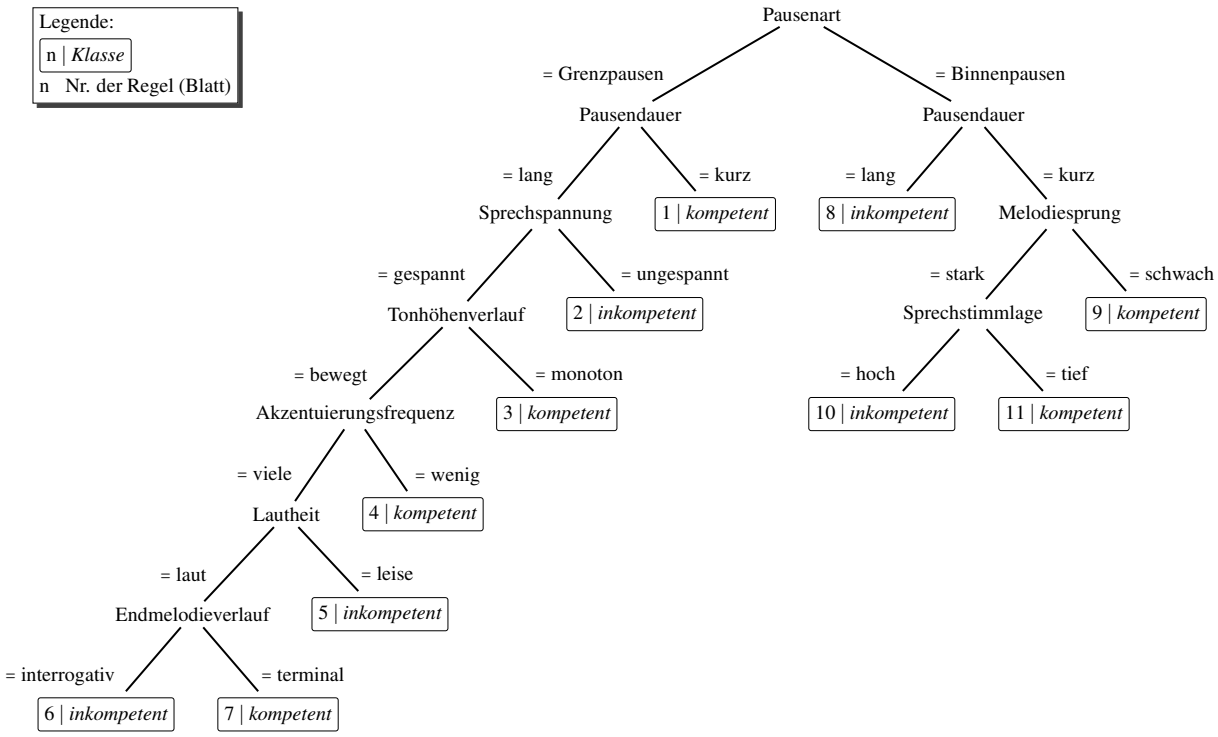

Abbildung 4 - C4.5-Entscheidungsbaum als Fusionsmodell für „Kompetenz“

bzw. Testmenge in erster Linie durch die Auswahl des Basisklassifikators, d. h. durch dessen Entscheidungen, verursacht. Eine nähere Analyse und Lösung dieses Problems ist Gegenstand der Folgeuntersuchungen. Eine Lösungsmöglichkeit betrifft die weitere Steigerung der Erkennungsleistung einzelner Kriterien. Hierbei ergeben sich zwei Herausforderungen – einerseits die Identifikation des optimalen Algorithmus und andererseits die Anpassung einzelner Lernalgorithmen durch Adaptation der Prozessparameter.

Die Analyse der Klassifikationsergebnisse in Abschnitt 3 zeigt, dass es keinen universellen Klassifikationsalgorithmus gibt, der in allen Kriterien andere Algorithmen dominiert. Dieses als „No Free Lunch“-Theorem [16] bekannte Phänomen bedingt eine intensive statistische Analyse der Algorithmen und ihrer Klassifikationsraten [17].

Neben der aufwändigen Analyse und Anpassung einzelner Algorithmen können zur Verbesserung der Klassifkationsgüte bereits als Basisklassifikatoren Systeme auf Basis mehrerer Modelle eingesetzt werden. Dies können z. B. Voting-Verfahren sein, die Klassifikationsentscheidungen zusammenfassen und gewichten können [18].

Ein zukünftiger Untersuchungsschwerpunkt betrifft die Konstruktion der Entscheidungsbäume auf der Fusionsebene. Bei den herkömmlichen Lernalgorithmen werden Splits auf Basis von Attributeigenschaften durchgeführt, beispielsweise über die Berechnung des Informationsgehalts der Attribute beim ID3-Algorithmus [19] oder über den Informationsgewinn [14]. Dieses Vorgehen basiert auf der Annahme, dass die Attribute a priori vorliegen und nicht veränderlich sind. Da die hier genutzten Attribute aber selbst Ergebnis einer Berechnung sind und einer Unsicherheit unterliegen, ist die Erweiterung der Split-Funktion durch die Klassifikationsgüte des Basisklassifikators denkbar, d. h. dass Splits bei Attributen bevorzugt werden, deren Klassifikatoren eine hohe Treffsicherheit haben. Hierzu muss die Erkennungsleistung intensiv statistisch analysiert und gewichtet werden.

Zur Beurteilung von Klassifikationsgüte und Sensitivität der Fusionsmodelle sollte der Einfluss des Sprachkorpus sowie der Sprachexperten untersucht werden. Die für einen Praxiseinsatz elementar wichtige Frage ist, ob die entwickelten Modelle mit unbekannten Korpora ebenfalls gute

Erkennungsraten erzielen. Hierbei könnte eine Verbesserung der Fusionssysteme erreicht werden, in dem die Entscheidungsbäume manuell oder semi-manuell durch einen Experten erzeugt werden und so dessen Wahrnehmung nachbilden [20].
Ein in der vorliegenden Untersuchung gänzlich unberücksichtigter Aspekt ist die zeitliche Komponente. Temporale Strukturen spielen eine wichtige Rolle bei der Wahrnehmung und können z. B. durch Hidden Markov-Modelle abgebildet werden [21].

6 Fazit

Im ersten Teil des Artikels wurde gezeigt, dass sich sowohl stimmliche Qualitätsmerkmale als auch Kriterien der Gesprächsqualität mit Klassifikationsalgorithmen erlernen lassen. Es wurden dabei Erkennungsraten von 55 % bis 97 % erreicht, wobei die stimmlichen Merkmale im Mittel über 70 % erzielten und somit besser erkennbar waren als die Kriterien der Gesprächsqualität.
Mit zuvor trainierten Basismodellen wurde auf Basis von Entscheidungsbaumverfahren ein Fusionssystem erstellt. Die einzelnen Fusionssystemvarianten haben im Vergleich zu Modellen auf der Signalebene eine schlechtere Klassifikationsleistung, sind aber im Gegensatz zu diesen introspektierbar. Bisherige Experimente zeigen, dass die aktuell eingesetzten Verfahren zum Lernen von Entscheidungsbäumen eine hohe Sensitivität in Bezug auf die Auswahl der Basisklassifikatoren sowie bezüglich der konkreten Lern- und Testdaten haben. Verschiedene Lösungsstrategien sind Gegenstand zukünftiger Untersuchungen.

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] MEISSNER, S. und J. PIETSCHMANN: *Rhetorische und phonetische Einflussfaktoren auf die Qualität von Telefonverkaufsgesprächen*. In: HIRSCHFELD, U. und B. NEUBER (Herausgeber): *Erforschung und Optimierung der Callcenterkommunikation.*, Seiten 215–248. Frank & Timme, Berlin, 2011.

[3] BOSE, I.: *dóch da sin ja' nur mûster: Kindlicher Sprechausdruck im sozialen Rollenspiel*. Peter Lang, Frankfurt, 1 Auflage, 2003.

[4] MEISSNER, S., J. PIETSCHMANN, M. WALTHER und L. NÖBEL: *Innovative IT-gestützte Ansätze zur Bewertung der Gesprächsqualität in Telefonverkaufsgesprächen*. In: HIRSCHFELD, U. und B. NEUBER (Herausgeber): *Erforschung und Optimierung der Callcenterkommunikation.*, Seiten 195–214. Frank & Timme, Berlin, 2011.

[5] 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.

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

[7] 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.

[8] 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.

[9] WITTEN, I. H., E. FRANK und M. A. HALL: *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 3 Auflage, 2011.

[10] WU, X., V. KUMAR, J. R. QUINLAN, J. GHOSH, Q. YANG, H. MOTODA, G. J. MCLACHLAN, A. NG, B. LIU und P. S. YU: *Top 10 algorithms in data mining*. Knowledge and Information Systems, 14(1):1–37, 2008.

[11] PAESCHKE, ASTRID: *Prosodische Analyse emotionaler Sprechweise*. Nummer 1 in *Mündliche Kommunikation*. Logos, Berlin, 2003.

[12] SCHULLER, B., S. STEIDL, A. BATLINER, E. NÖTH, A. VINCIARELLI, F. BURKHARDT, R. VAN SON, F. WENINGER, F. EYBEN, T. BOCKLET, G. MOHAMMADI und B. WEISS: *The INTERSPEECH 2012 Speaker Trait Challenge*. In: *Proceedings INTERSPEECH*, 2012.

[13] HERTLEIN, H.: *Fusion von Klassifikationssystemen für die automatische Sprechererkennung*. Logos, Berlin, 2010.

[14] QUINLAN, J. R.: *C4.5: programs for machine learning*. Morgan Kaufmann, San Mateo, 1993.

[15] WALTHER, M., T. MELLOULI und L. NÖBEL: *KI-basierte Modellierungsansätze für die Qualitätsbewertung von Servicegesprächen mittels Sprach- und Stimmanalyse*. In: BÖHMANN, T., R. KNACKSTEDT, J. M. LEIMEISTER, M. NÜTTGENS und O. THOMAS (Herausgeber): *Service Engineering & Management: Proceedings der Teilkonferenz im Rahmen der Multi-Konferenz Wirtschaftsinformatik*. Books on Demand, 2012.

[16] WOLPERT, D. H.: *The lack of a priori distinctions between learning algorithms*. Neural Computations, 8(7):1341–1390, 1996.

[17] SALZBERG, S.: *On comparing classifiers: Pitfalls to avoid and a recommended approach*. Data Mining and knowledge discovery, 1(3):317–328, 1997.

[18] DIETTERICH, T.: *Ensemble methods in machine learning*. International Workshop on Multiple Classifier Systems, Seiten 1–15, 2000.

[19] RUSSELL, S. und P. NORVIG: *Künstliche Intelligenz*. Pearson Studium, München, 2004.

[20] WARE, M., E. FRANK, G. HOLMES, M. HALL und I.H. WITTEN: *Interactive machine learning: letting users build classifiers*. International Journal of Human-Computer Studies, 55(3):281–292, 2001.

[21] 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, 1 Auflage, 2011.