WIEDERERKENNBARKEIT VON SPRECHERN BEI SCHMAL- UND BREITBANDIGER TELEFONÜBERTRAGUNG

Sebastian Möller[1], Laura Fernández Gallardo[1,2], Michael Wagner[1,2,3]

[1] *Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin, Deutschland*
[2] *Fac. of Education, Science, Technology and Mathematics, University of Canberra, Australia*
[3] *College of Engineering and Computer Science, Australian National University, Australia*

sebastian.moeller@telekom.de, laura.fernandez-gallardo@telekom.de, michael.wagner@canberra.edu.au

Abstract: In diesem Beitrag wird diskutiert, wie sich die Erkennbarkeit von Sprechern beim Übergang von schmalbandiger auf breitbandige oder super-breitbandige Telefonübertragung verbessert. Dazu wurde zunächst ein Versuchsparadigma definiert, bei dem innerhalb einer Gruppe bekannte Sprecher von Versuchspersonen auf Basis von Segmenten unterschiedlicher Länge identifiziert werden sollten. Es zeigt sich, dass sich die Erkennbarkeit mit der Kanalbandbreite signifikant verbessert. Allerdings können diese Verbesserungen teilweise durch inhärente Kanalbeeinträchtigungen (Kodierer, Endgeräte) wieder reduziert werden. In einer zweiten Versuchsreihe wurde untersucht, wie sich die maschinelle Sprechererkennung durch den verbreiterten Kanal ändert. Es zeigt sich wiederum eine Reduktion der *Equal Error Rate*, die je nach Erkennungsparadigma unterschiedlich deutlich ausfällt. Die Ergebnisse werden in einen Zusammenhang mit der erzielbaren Verbesserung der Gesamtqualität, der Sprachverständlichkeit und der Erkennbarkeit paralinguistischer Merkmale gebracht, um somit ein umfassendes Bild der erzielbaren Verbesserungen zu bekommen.

1 Einleitung und Motivation

Die Übertragung von Sprachsignalen erfolgt in zunehmendem Maße über Kanäle, die über eine gegenüber dem Standard-Telefonband (300-3400 Hz) erweiterte Bandbreite verfügen. Grund dafür ist neben der einfachen Realisierbarkeit bei IP-gebundenen Kanälen insbesondere die nachgewiesenermaßen bessere Sprachqualität. So geht man bei breitbandiger (50-7000 Hz) Sprachübertragung von einem mittleren Qualitätszuwachs von ca. 29% aus, bei super-breitbandiger (20-14000 Hz) Übertragung erzielt man möglicherweise noch höhere Qualität [1]. Neben der Gesamtqualität wird aber vermutlich auch die Verständlichkeit bei Konsonanten, insbesondere Frikativen, verbessert, sowie die (Wieder-) Erkennbarkeit von bekannten Sprechern. Letzterer Fragestellung widmet sich die Dissertation von Fernández Gallardo [2], die die Basis des vorliegenden Beitrags bildet.

Im Rahmen der Dissertation wurden zunächst Untersuchungen zur auditiven Unterscheidbarkeit von bekannten Sprechern bei Übertragungen über unterschiedliche Kanäle durchgeführt [3][4]. Hierzu wurde ein experimentelles Paradigma verwendet, bei dem miteinander bekannte Sprecher eines Instituts kurze Segmente einer/m von 8 möglichen Sprecherinnen oder Sprechern zuordnen sollten. Bei kurzen und Satz-langen Stimuli konnte ein signifikanter Anstieg der Erkennungsleistung und eine Beschleunigung des Entscheidungsvorganges beim Übergang von schmalbandiger zu breitbandiger Übertragung nachgewiesen werden. Auch der Einfluss unterschiedlicher Sprachkodierer und Endgeräte wurde untersucht.

Aufbauend auf diesen Ergebnissen wurde in einer zweiten Versuchsreihe der Einfluss der Kanalbandbreite und anderer Kanaleigenschaften auf automatische State-of-the-Art Sprecher-Verifikations-Algorithmen untersucht [5]. Es wurden unterschiedliche Techniken auf unterschiedlichen Datenbanken miteinander verglichen, wobei die Daten männlicher und

weiblicher Sprecher getrennt analysiert wurden. Dabei zeigte sich wiederum eine signifikante Reduktion der *Equal Error Rate*.

Die Ergebnisse werden in den beiden folgenden Kapiteln im Detail vorgestellt. In Kapitel 4 werden sie insbesondere im Hinblick auf das Verhältnis von Sprechererkennbarkeit zur wahrgenommenen Sprachqualität [6] diskutiert. Daraus werden Schlüsse für mögliche Folgearbeiten gezogen.

2 Auditive Unterscheidbarkeit von Sprechern

Methode: Die Erkennbarkeit von Sprechern in einer Telefonsituation hängt von der Bekanntheit von Sprecher und Hörer untereinander sowie vom Kontext (Erwartungshaltung, etc.) ab. Prinzipiell können auch unbekannte Sprecher identifiziert werden, sofern dem Hörer eine Trainingsphase angeboten wird. Dies entspricht allerdings nicht der normalen Telefonsituation, bei dem die Erkennbarkeit insbesondere für bekannte Sprecher von Relevanz ist.

Es wurden daher Erkennungstests innerhalb einer miteinander bekannten Gruppe von Personen durchgeführt. Hierzu wurde von 8 Mitarbeiterinnen und 8 Mitarbeitern der Telekom Innovation Labs (T-Labs) Sprachaufnahmen angefertigt, anhand derer andere Mitglieder der T-Labs die Sprecherinnen und Sprecher identifizieren sollten. Es wurde darauf geachtet, dass sowohl Sprecher als auch Hörer Muttersprachler waren, dass die Sprecher keinen auffälligen Dialekt aufwiesen, und dass alle Sprecher in etwa gleichem Maße (auch aus Telefonsituationen heraus) bekannt waren. Als Sprachmaterial wurden 24 phonemisch ausbalancierte EUROM-Textpassagen aufgezeichnet, die typische Telefoninhalte darstellen könnten. Aus diesen wurden anschließend auch kürzere Segmente (Sätze bzw. Wörter in Versuchsreihe 1a, den Satzbeginn „Könnten Sie mir“ in Versuchsreihe 1b) herausgeschnitten, um den Hörern Aufgaben unterschiedlicher Schwierigkeit zu geben. Die Aufnahmen wurden in einer akustisch isolierten Kabine mit einem qualitativ hochwertigen Mikrophon und Soundkarte vorgenommen und anschließend durch Kanalsimulationen gezielt gestört. Details zum Versuch finden sich in [3] und [4].

Versuchsreihe 1a: In einer ersten Versuchsreihe wurde zunächst der Effekt der Kanalbandbreite (Schmalband vs. Breitband) sowie unterschiedlicher für diese Kanäle typische Kodierer untersucht. Hierzu wurden die drei Wörter „auch“, „immer“ und „können“ als kurze Stimuli (3-5 Phoneme), ein Satz („Wir erwarten das Taxi genau um 5 Uhr 30.“; 14 Silben) als mittellanger Stimulus (∅=2,7s, sd=0,3s), sowie eine Textpassage (60 Silben) als langer Stimulus (∅=11,9s, sd=1,4s) verwendet. Die Stimuli wurden auf gleichen aktiven Sprachpegel (-26 dB ovl) gebracht und entweder durch schmalbandige Kanäle mit Standard-Kanalfilter nach ITU-T Rec. G.712 und anschließende Kodierung mit log. PCM (ITU-T Rec. G.711, 64 kbit/s), GSM-EFR-Kodierung (12,2 kbit/s) oder AMR-NB (4,75 kbit/s) oder durch breitbandige Kanäle mit Kanalfilter nach ITU-T Rec. P.341 und Kodierung nach ITU-T Rec. G.722 (64 kbit/s) oder AMR-WB (23,05 kbit/s) prozessiert. Die Stimuli wurden in einem ruhigen Büroraum diotisch über Kopfhörer präsentiert.

26 Versuchspersonen (19m, 7w) nahmen am Test teil. Sie hörten zunächst einen Satz (Breitband, ohne Störungen und Kanaleinschränkungen) jedes Sprechers, um sich mit den Stimmen noch einmal vertraut zu machen. Anschließend hörten sie zunächst die Wörter, dann die Sätze und abschließend die Textpassagen in randomisierter Sprecher- und Kanal-Reihenfolge. Die Aufgabe bestand in der Auswahl eines Sprechers oder einer Sprecherin aus 16 möglichen (*Forced Choice*). Die Sprecher wurden mit Namen und Bild auf einer grafischen Versuchsoberfläche zum Anklicken präsentiert, und die Antwortzeit (Klick-Zeit vom Beginn des Abspielens) bei den Textabschnitten wurde aufgezeichnet.

Versuchsreihe 1b: In einem zweiten Versuch wurde der Einfluss des sendeseitigen und empfangsseitigen Endgerätes sowie von Kanalstörungen in Form von Paketverlusten näher

beleuchtet. Hierbei wurden zwei Versionen eines Satzanfangs („Könnten Sie mir...“) als Stimulusmaterial verwendet. Dieses Material wurde in einem Büroraum über einen festen (SNOM 870) oder einen mobilen Handapparat (Sony Xperia T), einen Freisprecher (Polycom IP 7000) oder ein Headset (Beyerdynamic DT 790) abgespielt, mit einem Kunstkopf (HeadAcoustics HMS 3 bzw. B&K 4128C) wieder aufgezeichnet, und anschließend über verschiedene Kanalsimulationen prozessiert. Letztere bestanden entweder aus einem schmalbandigen Kanal mit G.711-Kodierung (64 kbit/s), einem breitbandigen Kanal mit G.722-Kodierung (64 kbit/s), oder einem super-breitbandigen Kanal mit Kodierung nach ITU-T Rec. G.722.1 (32 und 48 kbit/s). Beim mobilen Endgerät wurden stattdessen die Kodierungen nach AMR-NB (Schmalband, 12,2 kbit/s) oder AMR-WB (Breitband, 12,65 kbit/s) eingesetzt. Zusätzlich wurden bei einigen Kanälen Paketverluste mittels eines Asterisk-Servers (Verlustraten 0-15% random) simuliert. Die Präsentation der Stimuli erfolgte entweder über den festen Handapparat, den Freisprecher, das Headset oder einen hochqualitativen Kopfhörer (AKG K601 bei unterschiedlichen Endgeräten in Senderichtung), wiederum in einer ruhigen Büroumgebung.

Der Versuchsablauf folgte im Wesentlichen Versuchsreihe 1a, allerdings waren aufgrund der hohen Stimuluszahl 2 Sessions pro Versuchsperson notwendig, die an unterschiedlichen Tagen durchgeführt wurden. 20 Versuchspersonen (16m, 4w) nahmen an diesem Test teil; sie wurden nach denselben Kriterien wie bei Versuchsreihe 1a ausgewählt.

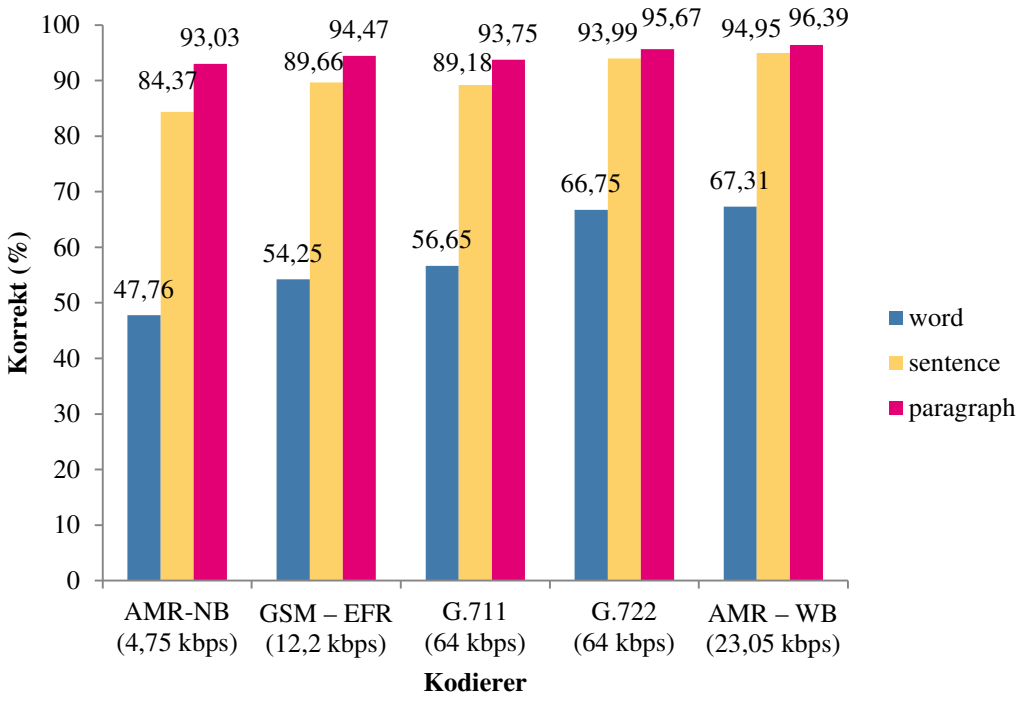
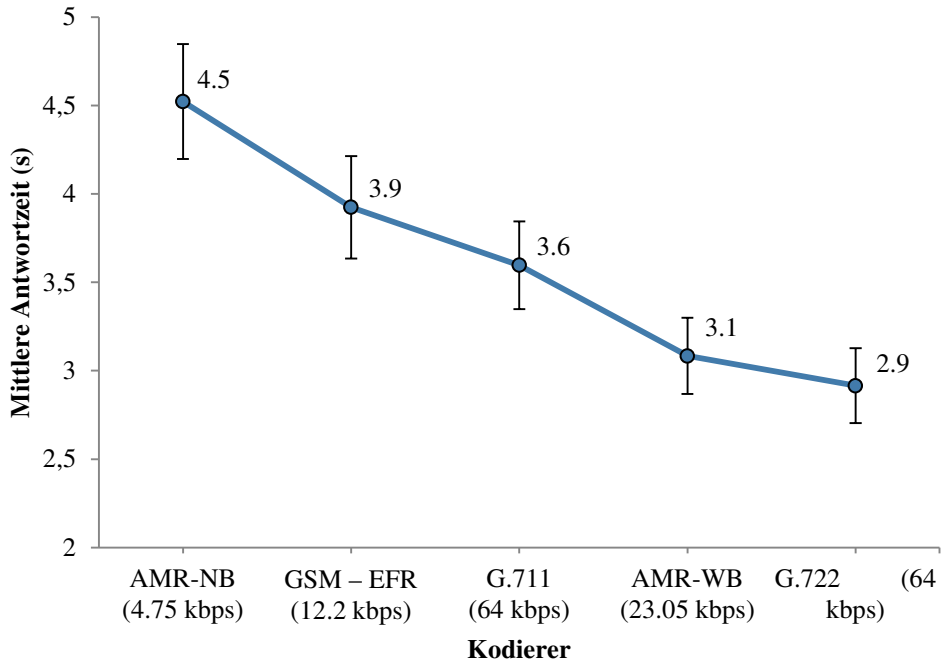


Abbildung 1: Korrektheit der Sprecheridentifizierung in Versuchsreihe 1a.

Ergebnisse: In Abb. 1 sind die mittleren Ergebnisse der Sprecheridentifizierung aus Versuchsreihe 1a dargestellt. Es zeigt sich ein signifikanter Anstieg der Korrektheit bei Wörtern und Sätzen zwischen den beiden Schmalbandkanälen GSM-EFR bzw. G.711 und den Breitbandkanälen G.722 bzw. AMR-WB (McNemar-Test, $p<0{,}001$ für Wörter, $p<0{,}05$ für Sätze). Der niederbitratige Kanal AMR-NB zeigte wiederum eine signifikant geringere Korrektheit gegenüber allen anderen Kanälen bei Wörtern und Sätzen. Bei den längeren Stimuli lässt sich ein ähnlicher Trend beobachten, allerdings scheint die Erkennungsaufgabe hier sehr einfach zu sein, sodass dieser Test in die Sättigung gerät; die Unterschiede zwischen den Kanälen sind daher nicht signifikant. Bei diesem Stimulusmaterial ist allerdings die Dauer

bis zur Erkennung (Antwortzeit) von Interesse. Diese ist in Abb. 2 dargestellt. Es zeigt sich, dass bei den schmalbandigen Kanälen G.711 und GSM-EFR eine signifikant längere Antwortzeit benötigt wird als bei den beiden breitbandigen Kanälen (Mann-Whitney U-Test,



p<0,001). Auch beim AMR-NB-Kanal ist die Antwortzeit wiederum signifikant länger als bei allen anderen Kanälen (p<0,001).

Bei den verschiedenen Endgeräten zeigt sich wiederum der Vorteil von schmalbandiger gegenüber breitbandiger Übertragung. Abb. 3 zeigt die Erkennungsraten aus Versuchsreihe 1b für unterschiedliche Endgeräte in Senderichtung. Der Unterschied zwischen Schmalband und Breitband ist für alle Endgeräte signifikant (McNemar-Test, p<0,05 beim festen Handapparat, p<0,001 bei allen anderen Endgeräten).

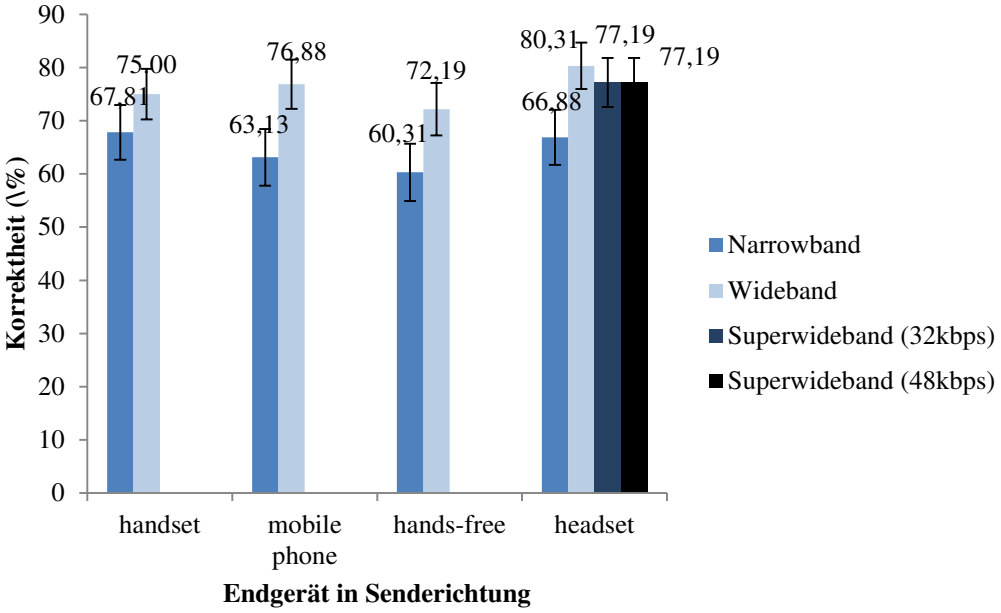


Abbildung 3: Korrektheit bei unterschiedlichen Endgeräten in Versuchsreihe 1b.

Interessanterweise steigt die Erkennungsleitung aber beim Übergang zu super-breitbandiger Übertragung nicht weiter an, sondern fällt gegenüber Breitband etwas ab. Dies könnte entweder darin begründet liegen, dass die Hörer diese Art der Sprachübertragung bislang noch nicht kennen und die Stimmen daher ungewohnt klingen, oder dass der bei der super-breitbandigen Übertragung verwendete G.722.1C-Kodierer besondere Artefakte aufweist, die die Sprechererkennung erschweren. Der Effekt ist bei beiden Bitraten des G.722.1C zu beobachten. Ähnliche Ergebnisse zeigen sich auch bei den unterschiedlichen Endgeräten in Empfangsrichtung, vgl. [2].

Erwartungsgemäß beeinflusst ein Paketverlust die Erkennungsrate. Dieser Effekt ist in Abb. 4 dargestellt. Es zeigt sich, dass der Abfall der Erkennungsleistung beim breitbandigen G.722-Kodierer geringer ist als beim schmalbandigen G.711-Kodierer. Für jede Verlustrate ist der Unterschied zwischen Schmalband und Breitband jedoch weiterhin signifikant ($p<0,01$).

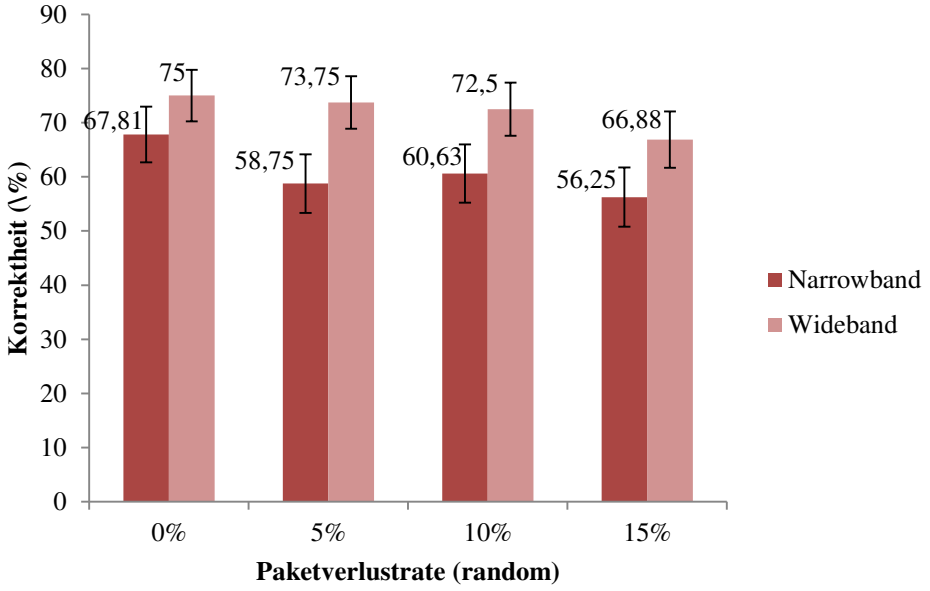


Abbildung 4: Korrektheit bei Paketverlusten in Versuchsreihe 1b.

3 Maschinelle Sprecheridentifikation

In modernen Telekommunikationsdiensten kommen auch Sprecheridentifikations- und Verifikationssysteme zum Einsatz. Diese verwenden ähnliche Eigenschaften des Sprachsignals bzw. des Sprachspektrums, wie sie auch ein menschlicher Hörer verwendet, um einen Sprecher aus *n* möglichen zu identifizieren (1-aus-*n*-Aufgabe) bzw. einen bekannten Sprecher zu verifizieren (1-zu-1-Aufgabe). Die Frage ist, ob diese Systeme in ähnlicher Weise von einer vergrößerten Bandbreite profitieren wie menschliche Hörer.

Die Leistung eines maschinellen Sprechererkenners hängt zunächst von den verwendeten Merkmalen und dem Erkennungsalgorithmus sowie dem zur Verfügung stehenden Trainings- und Testmaterial ab. Daher sind die in diesem Kapitel angegebenen Erkennungsergebnisse relativ in Bezug zum Algorithmus und zu den Trainingsdatenbanken zu verstehen. Daneben können Erkenner durch Wahl der Rückweisungsschwelle so eingestellt werden, dass sie eher zur Zurückweisung (dann auch eigentlich legitimer Sprecher) oder eher zur Akzeptanz (dann auch nicht legitimer Sprecher) neigen. Die Performanz eines Sprecherkenners wird daher üblicherweise als *Equal Error Rate* (EER) angegeben, d.h. als diejenige Rate, bei der falsche Akzeptanz und falsche Zurückweisung gleichauf liegen.

Versuchsreihe 2: Für die Untersuchung des Einflusses des Übertragungskanals standen drei unterschiedliche Systeme zur Sprecherverifikation (1-zu-1) zur Verfügung: Ein Gaussian-Mixtur-Model-Erkenner, der über ein Universal Background Model verfügt, um Sprecher zurückzuweisen (GMM-UBM); ein System, das mit Joint Factor Analysis arbeitet (JFA); und ein System, welches mit i-Vektoren arbeitet, um Sprecher- von Kanaleinflüssen zu trennen (i-vector). Diese Systeme wurden mit standardisierter Datenbanken trainiert und getestet. Im Folgenden beschränken wir uns auf die Darstellung der Ergebnisse für das GMM-UBM-System, welches mit 20 MFCC-Merkmalen (inkl. Delta- und Delta-Delta-Koeffizienten) und einem UBM mit 1024 Gauss'schen Komponenten arbeitet. Dabei wurde die AusTalk-Datenbank als Trainings- und Testdatenbank (15,5 Stunden Sprache zum Training des UBM, 25 Min. Sprache für das *Enrolement* jeden Sprechers, 10 *ClientModels*, Einzelwörter zum Testen) verwendet, und es wurde angenommen, dass Training- und Testdaten dieselben Störungen aufwiesen (*no mismatch*). Details zum Aufbau, Training und Test des Erkenners finden sich in [2], und weitere Ergebnisse für i-Vektoren sind in [5] veröffentlicht.

Ergebnisse: Die Erkennungsergebnisse unterscheiden sich signifikant zwischen männlichen und weiblichen Stimmen, deshalb werden Sprecherinnen und Sprecher separat betrachtet. Tabelle 1 zeigt die Ergebnisse bei unterschiedlichen Übertragungskanälen.

Tabelle 1: Erkennungsergebnisse (EER) beim GMM-UBM und unterschiedlichen Übertragungskanälen in Versuchsreihe 2

	EER (%)	
Übertragungskanal	Sprecher	Sprecherinnen
Kein Kodierer, 4 kHz	2,36	4,21
G.711, 64 kbit/s	2,95	4,68
AMR-NB, 12,2 kbit/s	3,82	6,63
GSM-EFR, 12,2 kbit/s	4,23	6,11
G.723.1, 5,3 kbit/s	6,47	9,15
Speex NB, 24,6 kbit/s	3,17	6,42
Kein Kodierer, 8 kHz	1,36	1,55
G.722, 64 kbit/s	1,23	1,64
AMR-WB, 12,65 kbit/s	1,82	2,35
Speex WB, 42,2 kbit/s	1,23	1,92
Kein Kodierer, 16 kHz	1,17	1,05
G.722.1C, 48 kbit/s	1,14	1,10

Es zeigt sich zunächst eine signifikant bessere Erkennungsleistung (niedrigere EER) bei Breitband gegenüber Schmalband. Dies deutet darauf hin, dass (wie angenommen) Frequenzen außerhalb des schmalen Telefonbandes wichtige Informationen auch für die maschinelle Sprechererkennung liefern können. Sowohl bei Sprechern als auch bei Sprecherinnen lässt sich eine weitere Reduktion der EER bei Übergang von breitbandiger zu super-breitbandiger Übertragung erzielen, allerdings ist der Rückgang nur bei Sprecherinnen signifikant. Dieses Ergebnis steht im Kontrast zur Beobachtung bei der menschlichen Sprecherkennung, die nicht von der super-breitbandigen Übertragung profitieren konnte (bei gleichem Kodierer G.722.1C).

Neben der Kanalbandbreite hat aber auch die Kodierung einen signifikanten Einfluss auf die Erkennungsleistung. So ist die EER bei schmalbandiger G.723.1-Kodierung mehr als doppelt so hoch wie bei Standard-G.711-Kodierung; bei breitbandiger Übertragung ist zumindest ein deutlicher Vorteil von G.722 (64 kbit/s) gegenüber AMR-WB (12,65 kbit/s) zu erkennen. Im Super-Breitband-Fall scheint der G.722.1C-Kodierer keine negativen Auswirkungen auf die Erkennungsleistung zu haben im Vergleich zum Fall ohne Kodierung. Ähnliche Ergebnisse zeigen sich auch bei JFA und i-Vektoren.

4 Diskussion und Ausblick

Die oben beschriebenen Versuchsreihen belegen deutlich, dass die Sprechererkennbarkeit bei breitbandiger gegenüber schmalbandiger Sprachübertragung signifikant gesteigert werden kann. Bei der auditiven Unterscheidung durch menschliche Hörer konnten sowohl bei Wort-langen als auch bei Satz-langen Stimuli signifikante Verbesserungen der Erkennungsrate für bekannte Sprecherinnen und Sprecher gezeigt werden. Bei Textpassagen konnte eine schnellere Erkennung (verringerte Antwortzeit) gezeigt werden. Der Vorteil hängt aber stark vom verwendeten Kodieralgorithmus ab, der jeder Übertragung inhärent ist; er bleibt bei unterschiedlichen Endgeräten erhalten, wobei sowohl Endgerät als auch Kanalstörungen (in diesem Fall Paketverluste) ebenfalls deutlichen Einfluss auf die Erkennungsleistung zeigten. Eine darüber hinausgehende Erweiterung des Kanals bis 14 kHz brachte demgegenüber keine weitere Verbesserung der auditiven Erkennungsleistung.

Auch bei der maschinellen Sprecherverifikation bestätigte sich der Vorteil breitbandiger Sprachübertragung. Hier konnte darüber hinaus auch ein weiterer Vorteil für super-breitbandige Übertragung gezeigt werden. Auch hier hat der Kodierer wiederum einen deutlichen Einfluss auf die EER.

Der Vorteil bei der auditiven Erkennungsleistung (durchschnittlich 14% höhere Erkennungsleistung bei Wörtern und 7% bei Sätzen, sowie 1s schnellere Antwortzeit bei Textpassagen) ist in Beziehung zu setzen zum Qualitätsvorsprung, den breitbandige gegenüber schmalbandiger Übertragung liefern [6]. Dieser kann bspw. mit Hilfe von Netzwerkplanungsmodellen wie dem E-Modell bestimmt werden, welche Nutzerurteile zur Gesamtqualität einer Telefonübertragung auf Basis von Eigenschaften des Netzes schätzen. Das bei der International Telecommunication Union (ITU-T) standardisierte E-Modell nimmt beispielsweise eine Qualitätsverbesserung von 29% beim Übergang von Schmalband zu Breitband an [7]. Für eine super-breitbandige Verbindung liegen noch keine standardisierten Werte vor, jedoch wurden von Wältermann [8] Verbesserungen bis 70% gegenüber Schmalband gezeigt.

Die Telefon-Übertragungsqualität ist allerdings keine eindimensionale Größe. So wurde in den Ansätzen von Heute et al., Wältermann, Côté und Scholz versucht, Übertragungsqualität multidimensional zu messen und vorherzusagen [9][10][11]. Für schmal- und breitbandige Verbindungen und unterschiedliche Störungsklassen wurden vier perzeptive Dimensionen von Beeinträchtigungen bestimmt: Klangverfärbung, Diskontinuierlichkeit, Rauschhaftigkeit, sowie nicht-optimale Lautheit. Dabei unterscheiden sich schmal- und breitbandige Verbindungen prinzipiell in der Klangverfärbung. Es könnte sein, dass diese perzeptive Dimension auch hauptsächlich für die Verbesserung der Sprechererkennbarkeit verantwortlich ist; erste Ergebnisse von Fernández Gallardo, die ebenfalls in der Dissertation [2] vorgelegt wurden, deuten darauf hin.

Daneben wird natürlich auch die Verständlichkeit durch die größere Kanalbandbreite positiv beeinflusst. Hier sind insbesondere die Frikative zu nennen, die einen hohen Energiegehalt im erweiterten Frequenzband aufweisen. Interessanterweise gibt es offenbar kaum Untersuchungen, die diesen Effekt unter realistischen (Kodierungs-) Bedingungen kontrolliert quantifizieren, vgl. den Überblick in [2]. Hierzu wären weitere Ergebnisse sehr wünschenswert.

Neben der Identität von Sprechern sind aber auch weitere paralinguistische Merkmale von Sprache relevant für die Kommunikation. Bspw. wäre es wichtig, die Erkennbarkeit des Alters, der Emotionen oder der Persönlichkeit eines Sprechers bei unterschiedlichen Kanalbandbreiten kontrolliert zu untersuchen.

Wie angedeutet wird auch die maschinelle Erkennung von Sprechern positiv durch die Verbreiterung des Telefonkanals beeinflusst. Diese Ergebnisse sind in Beziehung zu setzen zur Erkennbarkeit des linguistischen Inhaltes (maschinelle Spracherkennung), aber auch

anderer paralinguistischer Merkmale (Alter, Geschlecht, Emotionen, Persönlichkeit). Leider hängt die Performanz dieser Erkenner stark vom verwendeten Erkennungsalgorithmus sowie dem Trainings- und Testmaterial ab. Daher wird es schwierig sein, hier allgemein gültige und (für die Planung solcher Systeme) hilfreiche Vorhersagen zu treffen.

Insgesamt erscheint die Erweiterung des Übertragungsbandes jedoch an vielen Stellen positive Auswirkungen zu zeigen, die dem Kommunikationsprozess förderlich sein könnten. Diese Vorteile könnten aber durch die gleichzeitige (bessere) Übertragung von Hintergrundgeräuschen wieder zunichte gemacht werden. Auch hierzu sind weitere Untersuchungen wünschenswert.

Literatur

[1] Möller, S., Raake, A., Kitawaki, N., Takahashi, A., Wältermann, M. (2006). *Impairment Factor Framework for Wideband Speech Codecs*, IEEE Trans. Audio, Speech and Language Processing 14(6), 1969- 1976.

[2] Fernández Gallardo, L. (2014). *Human and Automatic Speaker Recognition over Telecommunication Channels*, Doctoral dissertation, University of Canberra, submitted.

[3] Fernández Gallardo, L., Möller, S., Wagner, M. (2013). *Human Speaker Identification of Known Voices Transmitted Through Different User Interfaces and Transmission Channels*, in: Proc. 2013 IEEE Int. Conf. Acoust. Speech and Signal Processing (ICASSP 2013), CA-Vancouver, 26-31 May.

[4] Fernández Gallardo, L., Möller, S. Wagner, M. (2012). Comparison of Human Speaker Identification of Known Voices Transmitted Through Narrowband and Wideband Communication Systems, in: *Informationstechnische Gesellschaft im VDE (ITG) Conference on Speech Communication*, pp. 219-222.

[5] Fernández Gallardo, L., Wagner, M., Möller, S. (2014). *I-vector Speaker Verification for Speech Degraded by Narrowband and Wideband Channels*, in: ITG Conference on Speech Communication, DE-Erlangen, pp. 4 pages.

[6] Möller, S., Köster, F., Fernández Gallardo, L., Wagner, M. (2014). *Comparison of Transmission Quality Dimensions of Narrowband, Wideband, and Super-Wideband Speech Channels*, in: Proc. 8th Int. Conf. on Signal Processing and Communication Systems (ICSPCS 2014), AU-Gold Cost, Dec. 15-17.

[7] ITU-T Recommendation G.107.1 (2011). *Wideband E-Model*, International Telecommunication Union, CH-Geneva, November 2011.

[8] Wältermann, M., Raake, A., Möller, S. (2010) *Extension of the E-Model Towards Super-Wideband Speech Transmission*, in: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2010), US-Dallas TX, 14-19 Mar.

[9] Heute, U., Möller, S., Raake, A., Scholz, K., Wältermann, M. (2005). *Integral and Diagnostic Speech-Quality Measurement: State of the Art, Problems, and New Approaches*, in: Proc. 4th European Congress on Acoustics (Forum Acusticum Budapest 2005), HU-Budapest, 1695-1700.

[10] Wältermann, M. (2012). *Dimension-based Quality Modeling of Transmitted Speech*, Springer, DE-Heidelberg.

[11] Côté, N. (2010). *Integral and Diagnostic Intrusive Prediction of Speech Quality*, Springer, DE-Heidelberg.