EINE DATENBANK FÜR UNTERSUCHUNGEN ZUR SPRECHERIDENTIFIKATION UND SPRECHERVERIFIKATION

Stephan Euler[*], *Christian Pilz*[**]

[*]*Fachhochschule Gießen-Friedberg, Fachbereich MND*
[**]*VOICE.TRUST AG*
Stephan.Euler@mnd.fh-friedberg.de

Abstract: In diesem Beitrag wird die Datenbank SUVI – Sprachaufnahmen für Untersuchungen zur Sprecher-Verifikation und Identifikation – vorgestellt. Ein Schwerpunkt liegt dabei auf der Auswahl geeigneter Wörter und Texte. Als Kriterium zur Auswahl der Wörter wird ein Maß für die phonetische Eignung eingeführt. Im Beitrag werden dann der Aufbau der Datenbank und das Aufnahmeverfahren beschrieben. In einer ersten Auswertung werden Art und Anzahl von Lesefehlern sowie die Längen der einzelnen Aufnahmen untersucht. In einem ersten Kontrollexperiment wird ein einfaches System zur Sprecher-Identifikation erprobt.

1 Einleitung

Biometrische Verfahren zur Identifikation und Verifikation von Personen haben in den letzten Jahren verstärkt an Bedeutung gewonnen. Im Bereich der Spracherkennung werden in diesem Anwendungsbereich überwiegend Verfahren eingesetzt, die auf dem Vergleich zwischen Referenz- und Testäußerungen eines oder mehrerer vorgegebener Texte beruhen. Neben der Sicherheit der Erkennung spielt für den praktischen Einsatz auch die Auswahl geeigneter Texte eine große Rolle. In Hinblick auf eine hohe Akzeptanz ist es wichtig, Texte zu finden die bei möglichst kurzer Länge ausreichende Informationen für eine zuverlässige Erkennung beinhalten.

Für den Bereich der Sprecher-Verifikation und Identifikation existieren bereits eine Anzahl von Datenbank wie YOHO *Voice Verification Corpus* (Englisch), SIVA (Italienisch) [1] und VeriDat (Deutsch) [2]. Ergänzend zu diesen Sammlungen wurde die Datenbank SUVI – **S**prachaufnahmen für **U**ntersuchungen zur Sprecher-**V**erifikation und **I**dentifikation - entwickelt. Die Datenbank soll einerseits zur Evaluierung von entsprechenden Verfahren dienen. Gleichzeitig soll auch die Brauchbarkeit verschiedener Wörter oder Sätze für derartige Anwendungen untersucht werden.

Zur Auswahl geeigneter Wörter wurde ein Maß für die phonetische Eignung eingeführt. Dieses Maß basiert auf der bekannten Erfahrung, dass eine Teilmenge aller Phoneme die wesentlichen Informationen über die Identität einer Sprecherin oder eines Sprechers enthält Das Maß bewertet die Anzahl solcher besonders geeigneter Phoneme zur Gesamtzahl der Phoneme in einem Wort oder einer Phrase.

Als erster Test wurde auf der Basis der gesammelten Sprachdaten ein Simulationsexperiment zur Sprechererkennung durchgeführt, bei dem jede Sprecherin und jeder Sprecher durch ein Codebuch von Merkmalsvektoren modelliert wurde. Das Experiment sollte einen ersten Eindruck von den Eigenschaften der Sprachdaten liefern und eventuelle Probleme frühzeitig aufzeigen. Daher wurden die ersten Messungen frühzeitig gestartet. In diesem Beitrag werden Ergebnisse mit den ersten 150 Sprecherinnen und Sprecher berichtet.

2 Phonetische Eignung

Erfahrungsgemäß eignen sich die verschiedenen Laute unterschiedlich gut für die Verifikation. Basierend auf empirischen Untersuchungen zur Erkennungsleistung einzelner Lauteinheiten in der Sprechererkennung, [3, 4], wurde die Statistik in Tabelle 1 zusammengestellt. In der Tabelle sind die Gleichfehlerraten (Equal Error Rate, EER) für einige Phoneme eingetragen. Da sich die Erkennungsgüte von Laut zu Laut in der Tat stark unterscheidet, liegt es nahe, überhaupt nur Lauteinheiten mit guter Erkennungs-Eignung zu verwenden [5]. Wörter, die überwiegend die geeigneten Laute enthalten, sollten bevorzugt eingesetzt werden. Ferner kann die Silbenanzahl insofern einbezogen werden, als dass die Wörter mit dem maximalen Silbenvorkommen bevorzugt werden.

SAMPA Symbol	EER (Gleichfehlerrate)	SAMPA Symbol	EER (Gleichfehlerrate)
a:	8,2%	u:	18,1%
m	8,5%	s	18,6%
N	9,7%	f	21,0%
i:	10,5%	k	23,7%
E	10,6%	v	24,7%
o:	11,4%	t	25,3%
S	15,7%	n	31,0%

Tabelle 1 - Auszug der Gleichfehlerraten einzelner Phoneme in der Spracherkennung

Um diese Hypothese zu testen, wurden 8 verschiedene Eingaben in einem System zur textabhängigen Sprecherverifikation auf ihre Eignung verglichen. Tabelle 2 zeigt die Eingaben – vereinfachend im Folgenden als Passwörter bezeichnet – des VoiceTrust internen CommonCriteria2005 Korpus mit jeweiliger phonetischer Transkription (SAMPA Format). Zu jedem Passwort ist der nach dem beschriebenen Verfahren bestimmte Bewertungskoeffizient eingetragen. Zur Vereinfachung wurde dabei angenommen, dass nur die Phoneme /a, a:/ /e, e:/, /E, E:/, /I, i:/, /O, o:/, /Y, y:/, /6/, /m/, /N/, /j/, /S/ gut geeignet, alle übrigen Phoneme hingegen nicht geeignet sind. Der Bewertungskoeffizient berechnet sich dann als Anteil dieser Phoneme an einem Passwort.

	Orthographie	Phonemic	Eignungsbilanz
Cr1	*Rosemarie Maximilian*	ro:z@mari:maksImi:lIa:n	12/19 = 0,6316
Cr2	*Lieselotte Sebastian*	li:z@lOt@z@bastIa:n	6/17 = 0,3529
Cr3	*Veronika Ferdinand*	ve:ro:nIkafErdi:nant	10/17 = 0,5882
Cr4	*Evamaria Konstantin*	e:famari:akOnstanti:n	12/18 = 0,6667
Cr5	*Christiane Dagobert*	krIstIa:n@da:go:bErt	7/17 = 0,4118
GID	*GHI456*	ge:ha:i:fi:rfYnfsEks	7/16 = 0,4375
GName	*Karoline Mustermann*	karo:li:n@mUst@rman	8/17 = 0,4706
GPhrase	*Meine Stimme ist mein Passwort*	maIn@StIm@IstmaInpasvOrt	13/24 = 0,5417

Tabelle 2 - Passwörter, phonetische Transkription und deren Eignungsbilanz

Zur text-abhängigen Sprechererkennung wurde eine Ganzwort-Modelierung mittels eines DTW-HMM Klassifikators verwendet. Als Merkmalsgrößen wurden bei 30 ms Fenstergröße und 20 ms Überlapp je 12 Cepstral- und 12 Delta-Cepstral-Koeffizienten berechnet. Die Trainingsmuster wurden mit dem Dynamic Time Warp (DTW) Verfahren normalisiert, wobei die Äußerung mit der geringsten Länge als Referenz für die zeitliche Anpassung genutzt wurde. Dann wurde für jedes Passwort ein Hidden Markov Modell erstellt. Die Anzahl der Zustände richtete sich nach der Länge. Jeder Zustand enthielt zwei Gauß-Komponenten mit diagonalen Kovarianzmatrizen. Getestet wurden 100 Sprecher mit jeweils vier Trainingsäußerungen und einer Testäußerung für jedes Passwort.

Die nachfolgende Tabelle 3 zeigt hierfür eine Zusammenstellung der ermittelten Gleichfehlerraten der einzelnen Passwörter zusammen mit dem zugehörigen Wert des Bewertungskoeffizienten. Es zeigt sich, dass die Erkennungsleistung für Äußerungen mit hohem phonematischen Bewertungskoeffizienten tatsächlich ebenfalls hoch ist, wodurch die Brauchbarkeit des Verfahrens im Kontext der Registrierung bzw. Authentifizierung von Personen anhand ihres Stimmprofils belegt ist.

	EER in %:	Eignungsbilanz:
CR4	1,07	0,6667
CR1	1,66	0,6316
CR3	1,95	0,5882
GPhrase	2,00	0,5417
CR2	2,23	0,3529
GName	2,46	0,4706
CR5	2,55	0,4118
GID	4,83	0,4375

Tabelle 3 - Gleichfehlerraten der Passwörter, absteigend sortiert, und deren Eignungsbilanz

3 SUVI Datenbank

3.1 Aufbau der Datenbank

Die Datenbank enthält sowohl einzeln gesprochene Wörter als auch ganze Sätze. In Hinblick auf die Verwendung zum Evaluieren text-abhängiger Verfahren wurden stets dieselben Texte vorgegeben. Im Einzelnen sollten von jeder Sprecherin und jedem Sprecher folgende Aufnahmen gesprochen werden:

- 5 Listen mit jeweils 45 einzelnen Wörtern (Wortliste 1 bis 5)
- Ein kurzer Text mit phonetisch ausgewogenen Sätzen
- Die Buttergeschichte aus dem Kieler Corpus of Read Speech

Alle Wörter in den Listen haben eine phonetische Eignung von mindestens 0,6. Während der ersten Aufnahmen zeigte sich, dass einige Wörter schwierig vorzulesen waren. So kannten einige Probanden den Vornamen *Angelina* nicht und wussten die Aussprache nicht oder verwechselten ihn mit *Angelika*. Nach den Erfahrungen mit den ersten 7 Probanden wurden daher 5 Wörter ausgetauscht.

3.2 Aufnahmeverfahren

Die Aufnahmen erfolgten jeweils in einer Sitzung mit einer fest vorgegebenen Reihenfolge. Verwendet wurde ein Fostex FR-2 Field Memory Recorder und ein Sennheiser Kondensatormikrofon ME64. Die Sprachdaten wurden mit 44 kHz aufgezeichnet. Das Mirkophon stand in einem Abstand von etwa 50 cm vor den Probanden. Die Aufnahmen fanden in unterschiedlichen Räumen sowohl innerhalb der Fachhochschule als auch in Wohnungen statt. Dabei wurde darauf geachtet, dass die Räume geringen Nachhall haben und keine Störgeräusche auftreten [6].

Nach einer kurzen Einweisung wurden die Probanden aufgefordert, die vorgegebenen Wörter und Texte vorzulesen. Dabei wurde immer die gleiche Reihenfolge

1. Wortliste 1
2. Wortliste 2
3. Phonetisch ausgewogene Sätze
4. Wortliste 3
5. Wortliste 4
6. Buttergeschichte
7. Wortliste 5

eingehalten. In der Regel erfolgten die Aufnahmen mit Zweiergruppen, so dass die benötigte Zeit zum Abholen und Instruieren der Probanden etwa halbiert werden konnte. Insgesamt lag der Zeitaufwand für die Aufnahme zweier Probanden bei etwa 20 Minuten. Die Aufnahmen begannen im Sommer 2006 und waren im Januar 2007 beendet.

Alle Aufnahmen wurden vor der Übernahme in die Datenbank überprüft. Lange Pausen vor und nach den Sprachdaten wurden dabei entfernt. Soweit möglich wurden grobe Fehler wie falsche Anfänge oder Korrekturen entfernt. Kleinere Fehler wurden in den Aufnahmen belassen. Einige wenige Aufnahmen mit schlechter Qualität konnten nicht weiter verwendet werden.

3.3 Statistische Auswertung

Insgesamt enthält die Datenbank je 130 Sprecher und Sprecherinnen im Alter zwischen 17 und 66 Jahren. Alle Aufnahmen erfolgten in der Region um Friedberg, wobei nahezu die Hälfte der Sprecherinnen und Sprecher in dieser Region aufgewachsen waren. Für insgesamt 8 Sprecherinnen und Sprecher war Deutsch nicht Muttersprache. Angesichts der geringen Stichprobengröße wurde lediglich auf eine Ausgewogenheit bezüglich des Geschlechts geachtet. Im Alter und in der Dialektfärbung ergaben sich durch einen hohen Anteil von Studierenden aus der Region deutliche Ungleichverteilungen. So lag das Durchschnittsalter bei nur 30,5 Jahren.

Bei der Mehrzahl der Probanden kam es zu Fehlern beim Vorlesen. Bei den Einzelwörtern resultierten die allermeisten Abweichungen von den Vorgaben aus Lesefehlern. Am häufigsten traten dabei die folgenden Verwechslungen auf:

Vorgegebenes Wort	Gelesenes Wort
Autor	*Auto*
Karma	*Kamera*
Komma	*Koma*
Mahnmal	*manchmal*
wahr	*warm*

Insgesamt stellten wir in etwa 17% aller Aufnahmen der Wortlisten Fehler fest. Erwartungsgemäß erwiesen sich die phonetisch ausgewogenen Sätze und die Kurzgeschichte als fehleranfälliger. Insbesondere die Sätze fielen vielen Probanden schwer. Sie enthalten einige wenig gebräuchliche Wörter und schwierigere syntaktische Konstruktionen. Außerdem fehlt ein durchgängiger inhaltlicher roter Faden. Die Fehlerquote betrug bei den Sätzen 72%. Demgegenüber lag sie bei der Kurzgeschichte um 63%. Einige Fehler entstanden durch Ersetzungen mit dem gleichen Sinngehalt. Zwei Beispiele dazu sind:

Vorgegeben	Gelesen
Brötchenkorb	*Brotkorb*
vorlassen	*vorbei lassen*

Um einen Überblick über den Umfang der aufgezeichneten Sprachdaten zu gewinnen, wurden die Dauern aller Sprachaufnahmen ausgewertet. In Tabelle 4 sind die Werte der einzelnen Aufnahmen zusammengestellt. Für die fünf Wortlisten liegen die mittleren Werte relativ konstant um 53 Sekunden. Die jeweils kürzesten und längsten Aufnahmen unterscheiden sich um einen Faktor von 2,5 bis 3. Im Mittel haben die Sprachdaten eine Gesamtlänge von 7:20 Minuten pro Sprecher/in.

	Minimum	Durchschnitt	Maximum
Wortliste 1	32,92	53,62	92,01
Wortliste 2	34,01	53,51	83,24
Wortliste 3	32,45	52,66	83,31
Wortliste 4	33,01	53,99	81,17
Wortliste 5	32,43	52,88	81,90
Text 1	58,08	82,34	131,71
Text 2	64,24	90,75	163,40

Tabelle 4 - Länge der einzelnen Aufnahmen in Sekunden

4 Sprechererkennung

Als erste Untersuchung wurde in Simulationsmessungen ein System zur Sprecheridentifikation mittels Vektorquantisierung erprobt. Dabei wurde die Teilmenge der ersten 150 Sprecher/innen verwendet. Von jedem Sprecher bzw. jeder Sprecherin wurde zunächst jeweils aus der ersten Aufnahme mit Wörtern ein Codebuch generiert. Alle anderen Aufnahmen dienten dann als Testmuster. Jedes Testmuster wurde mit allen Codebücher verglichen und der Sprecher oder die Sprecherin mit dem geringsten Abstand zu dem Codebuch galt als erkannt. Im Einzelnen wurden folgende Schritte durchgeführt:

1. Umsetzung der Abtastrate auf 16 kHz mit Hilfe des Programms SoX[1]
2. Berechnung der MFCC-Merkmalsvektoren (Dimension 12)
3. Erstellung eines Codebuchs mit Mustervektoren aus einer Aufnahme jeder Sprecherin und jedes Sprechers (standardmäßig Wortliste 1)
4. Alle verbleibenden Aufnahmen wurden mit allen Codebüchern verglichen. Das Codebuch mit dem minimalen Abstand wurde bestimmt, die zugehörige Sprecherin beziehungsweise der zugehörige Sprecher galt als erkannt.

Im Test wird jede Aufnahme einzeln betrachtet. Von jedem Probanten gilt eine Äußerung als Referenz und die sechs anderen dienen als unabhängige Testmuster. In diesem Experiment verwendeten wir die Programme und Dateiformate des Hidden Markov Toolkits (HTK) der Universität Cambridge [7]. Da allerdings das HTK-Programm zur Vektorquantisierung keine Abstandsinformation ausgibt, wurde für diesen Verarbeitungsschritt ein eigenes Programm eingesetzt. Das Programm speichert die Abstandswerte und Erkennungsergebnisse in einer XML-Datei für die weitere Verarbeitung. Über eine Verknüpfung mit einem XSLT-Layout werden die Ergebnisse für die Darstellung in Browsern übersichtlich formatiert. Beispielsweise werden nicht korrekt zugeordnete Aufnahmen durch farbliche Markierung hervorgehoben.

In einem ersten Simulationsexperiment wurde die Größe der Codebücher von 8 bis 100 Vektoren variiert. Die Erkennungsergebnisse sind in Tabelle 5 aufgeführt. In diesen Messungen erwies sich eine Codebuchgröße von 32 als ausreichend. Die Fehlerquote lag dann bei etwa 0,8%. Die Vergrößerung der Codebücher führt zu keinen signifikanten Verbesserungen. Alle 7 Fehler traten bei den beiden Texten auf. Bei zwei Sprechern wurden sogar beide Texte jeweils demselben falschen Sprecher zugeordnet. Die Unterschiede

[1] sox.sourceforge.net

zwischen einzeln gesprochenen Wörtern und ganzen Sätzen verhindern hier eine bessere Identifikation. In einem Gegenversuch wurden die phonetisch ausgewogenen Sätze als Referenz verwenden. Dann wurden alle „Buttergeschichten“ korrekt zugeordnet, aber in 17 Fällen Wortlisten falsch erkannt.

Codebuchgröße	Fehler	Fehlerquote
8	70	7,78 %
16	11	1,22 %
32	7	0,78 %
48	5	0,56 %
64	6	0,67 %
100	7	0,78 %

Tabelle 5 - Erkennungsfehler bei verschiedenen Codebuchgrößen

In einem weiteren Experiment wurde die Abhängigkeit der Erkennungsgenauigkeit von der Dauer der Testproben untersucht. Dazu wurde jeweils nur ein vorgegebener Anteil jedes Testmusters in der Auswertung berücksichtig. Von den Merkmalsvektoren wurde ein entsprechender Prozentsatz – immer gezählt ab dem Beginn – ausgewertet. In Tabelle 6 sind die entsprechenden Ergebnisse gezeigt. Ab 25% erhält man eine akzeptable Erkennungsquote und 50 % scheinen bereits ausreichend.

Anteil jedes Testmusters	Erkennungsquote
5 %	7,11 %
10 %	2,44 %
25 %	1,33 %
50 %	0,78 %
75 %	0,67 %
100 %	0,78 %

Tabelle 6 - Erkennungsquote bei Berücksichtigung unterschiedliche Anteile der Testmuster

5 Zusammenfassung

Die in diesem Beitrag vorgestellte Datenbank SUVI umfasst Aufnahmen von 260 verschiedenen Sprecherinnen und Sprecher. Von jedem Probanden wurden 7 Aufnahmen mit einer mittleren Gesamtlänge von 7:20 Minuten erstellt. Der Schwerpunkt liegt auf dem Vergleich verschiedener Wörter und Sätze hinsichtlich ihrer Eignung zur Sprecher-Verifikation oder –Identifikation. Die Daten erfassen keine zeitlichen Veränderungen der Sprechcharakteristika. Wir haben daher damit begonnen, von einigen Sprechern und Sprecherinnen mehrere Proben der gleichen Wörter und Texte über einen längeren Zeitraum aufzuzeichnen.

Literatur

[1] M. Falcone und A. Gallo, “The SIVA speech database for speaker verification: description and evaluation,” in *ICSLP96*, Philadelphia, 1996.

[2] U. Türk und F. Schiel, “Speaker verification based on the German VeriDat database,” in *EUROSPEECH 2003*, Geneva, 2003.
[3] M. Savic und S. Gupta, “Variable parameter speaker verification system based on hidden Markov modeling,” in *ICASSP-90*, Albuquerque, 1990, pp. 281–284.
[4] S. Euler, R. Langlitz, und J. Zinke, “Comparison of whole word and subword modeling techniques for speaker verification with limited training data,” in *ICASSP–97*, München, 1997.
[5] M. A. Fattah, F. Ren, und S. Kuroiwa, “Phoneme based speaker modeling to improve speaker identification,” in *International Conference on Artificial Intelligence and Machine Learning (AIML-06)*, Sharm El Sheikh, Egypt, June 2006.
[6] M. Thieme, “Entwicklung und Auswertung eines Simulationssystems zur Sprecher-Identifikation“. Diplomarbeit FH Gießen-Friedberg, April 2007
[7] S. J. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, und P. Woodland, *The HTK Book*, Cambridge University Engineering Department, 2002.