Biometrics and Security
Speaker Recognition

Jonas Marquardt und Maik Riestock
Otto von Guericke University Magdeburg - Advanced Multimedia and Secruity Lab (AMSL)

speaker reco
orientiert am hyke(database)
Closed set speaker authentication
projektion auf 'Doddingtons Zoo'

Categories and Subject Descriptors:

Additional Key Words and Phrases: Speaker Recognition, Audio Feature Extraction, Doddingtons Zoo

1. MOTIVATION

Um Personen eindeutig zu identifizieren gibt es verschiedene Verfahren. Eines davon ist die Erkennung der Stimme. Als Aufnahmegert ist ein handelsbliches Mikrophone ausreichend. In dieser bung ging es darum, herauszufinden wie eine Stimmenerkennung umgesetzt wird und welche Eigenheiten dieses Verfahren mit sich bringt.

2. HYKE-SYSTEM

Das Hyke ist ein System welches durch die Stimme eines
es stamm von einer dezentralen Bildungseinrichtung aus Indien die es dazu verwendet um nachzuvollziehen onb die Lehrer ihren Unterricht
allgemein

—Bildungseinrichtung
—Rajasthan, nw india
—jede schule hat 1-3 lehrer
—hauptquartier in udaipur, 150 entfernt
—hauptquartier in udaipur, 150 entfernt

bisher

—visuelle kontrtolle der anwesenheit
—2 mal am tag bilder
—manuell jedes bild verifiziert wurde
—kosten fr angestellte
—jede schule bentigt eine digital kamera

neu

—neuer ansatz mit stimmen erkennung
—da 75 % der schulen eh schon ein telefon haben
—automatische verifikation

This report was created in the context of the course Biometrics and Security [BIOSEC] winter term 2014/15. This course was held by: Prof. Dr.-Ing. Jana Dittmann and Prof. Dr.-Ing. Claus Vielhauer; Research group Multimedia and Security, Otto-von-Guericke-University of Magdeburg, Germany. The course was supported by: Dr.-Ing. Christian Krtzer, M.Sc. Kun Qian

ergebnis

—95 % erkennungs rate, state of the art

3. UNSER ANSATZ

es geht um speaker recognition, wobei
Speaker recognition, sometimes referred to as speaker biometrics, includes identification, verification (authentication), classification, and by extension, segmentation, tracking and detection of speakers. It is a generic term used for any procedure which involves knowledge of the identity of a person based on his/her voice. [Beigi 2011]

In diesem rahmen sollten die folgenden aufgabenstellungen erfolgreich abgschlossen werden:

- Closed set speaker authentication on the Hyke speech database
- Compare the results achieved (in terms of authentication performance) to th results persented in [Azarias Reda 2011]
- A projection of the samples in your data set to the characters of 'Doddingtons Zoo'

Bei den Arbeitsschritte in unserem Projekt haben wir uns an das allgemeine Model fr das Authentifizieren von Benutzern anhand von Biometrischen Daten orientiert, welches in [**?**] vorgestellt wird. Dieses Model haben wir an unser Thema, der *Speaker Recognition*, und der Aufgabenstellung angepasst.

- Data Acquisition: Hyke Database 4
- Pre-processing: kein klassisches pre-pro..., einteilung in Sets 5
- Feature Extraction: Merkmalsentnahme der Audioaufnahmen mittels AAFE. 6
- Post-processing: eingefgt 7
- Comparsion and Classification: 8

die vorgestellten Ablauf eines Authentifizieungsprozesses
Bei unseren arbeitsschritten haben wir uns dabei an an der von ... vorgestellen pro chain gehalten
General model for biometric user authentication [**?**]
dieser prozesskette haben wir noch den schritt des post-processing hinzugefgt, weil das entfernen von noise aus feature mit weka angeboten hat.
fr diese aufgabe wurden uns die folgenden Programme bereitgestellt:

- **AAFE** (AMSL Audio Feature Extractor), ist ein Tool für die Extraktion von Merkmalen in Audio-Dateien und entstammt dem AMSL Audio Steganalysis Toolset (AAST).[**?**] Anwendung fand das AAFE-Tool in dem Kapitel:6.
- **WEKA**, ist eine Sammlung von Algorithmen des Maschinellen Lernens für Aufgaben im Bereich des Data-Mining.[**?**] Anwendung fand dieses Tool in den Kapieln:7 und 8.

4. DATENBASIS

Die Datenbasis wurde dem Hyke-Projekt entnommen. Sie kann unter folgender URL heruntergeladen werden: Sie umfasst Aufnahmen von 83 verschiedenen Sprechern, davon 48 mnnlich und 35 weiblich. Von jeder Person gibt es fnf Aufnahmen in denen Abfolgen verschiedener Ziffern gesprochen werden. Die Sprache dabei ist Englisch. Die Lnge der Aufnahmen liegt zwischen 5 und 35 Sekunden. Es gibt auch Aufnahmen, die keine Stimme enthalten. Die Stimmen wurden ber das Telefon aufgenommen und bieten daher eine geringere Bandbreite als die menschliche Stimme hat. Bei den Sprechern handelt es sich um Inder mit verschiedenen Hintergrnden. Die Aufnahmen enthalten teilweise Hintergrundgerusche, vom leisem Rauschen bis zu Gesprchen und Musik.

5. VORVERARBEITUNG

In diesem Kapitel geht es um die Daten unserer Datenbank auf die folgenden schritte vorzubereiten.

Da wir in unseren Ergebnissen am ende einen möglichen Unterschied zwischen den Ergebnissen der Authentifizierung beider Geschlechtern beobachten zu knnen, wurden die Datenbank in sechs Sets unterteilt. Hierfür wurde das Set mit Sprechern von beiden Geschlechtern, *mixed set*, aufgeteilt in zwei Sets mit ausschlielich Stimmen von weiblichen Sprechern, *female set*, und mit ausschlielich mnnlichen Sprechern, *male set*.

Zusätzlich benötigen die Klassifikatoren zwei verschiedene Sets von Daten. mit dem einen Set wird das Model trainiert, hier *train set*, und mit dem andren Set evaluiert, hier *test set*.

Die Datenbasis aufgeteilt in folgende Sets:

- mixed train set
- mixed test set
- female train set
- female test set
- male train set
- male test set

6. FEATURE EXTRACTION

Um aus den Aufnahmen die Features zu extrahieren wurde der *AMSL Feature Extractor* verwendet. Dieser zerlegt eine Audiodatei in sehr kurze Samples und berechnet aus diesen verschiedene Features. Die Lnge der Samples kann man frei whlen. Wir haben einen Lnge von 1024 gewhlt und ein berlappung von Null. Die Hemmingenfenster Funktion wurde aktiviert. Der Feature Extractor berechnet aus jedem Sample 593 verschiedene Features berechnet.

7. NACHTVERARBEITUNG

Die extrahierten Daten wurden mit Hilfe von Weka aufbereitet. Dadurch sollten bessere Ergebnisse bei der Klassifikation erzeugt werden. Dazu wurde die Features *lbs flipping ratio* (in allen Instanzen 922337203685477.6000) und *lbs flipping rate* (in allen Instanzen 0) entfernt. Weil sie in allen Fllen gleich sind lassen sich an ihnen keine Unterschiede in den Aufnahmen feststellen.

In den Aufnahmen gibt es Bereiche die keine Stimme enthalten. Diese konzentrieren sich auf Anfang und Ende der Datei. Es gibt auch Pausen zwischen den gesprochenen Ziffern. Die “stillen“Bereiche enthalten keine Information ber die Stimme und somit den Sprecher. Dadurch wird die sptere Klassifikation erschwert. Um die “Stille“herauszufiltern wurden alle Samples mit einer geringen Amplitude gelscht. Dazu wurde das Feature *rms amplitude* genutzt und alle Samples mit einem Wert unter 10 gefiltert. Es wurde der *RemoveWithValues* Filter von Weka mit den Parametern -S 10.0 -C 5 -L first-last verwendet. Dadurch wurden von 50.424 Samples 32.026 entfernt. Das heit es wurden rund 64 Prozent der Datenbasis entfernt.

8. KLASSIFIKATION

in diesem kapitel geht es um die klassifikation unserer daten.

Klassifizierung ist def durch

Fr diese Aufgabe haben wir das tool WEKA verwendet, welches in Kapitel:3 vorgestellt wurde.

Usnere ausgangsdaten fr die klassifikatoren sind die aufbearbeiteten features, welche in Kapitel:7 vorgestellt wurden, die wir in den folgenden sets unterscheiden:

- mixed train set as feature matrix
- mixed test set as feature matrix
- female train set as feature matrix
- female test set as feature matrix
- male train set as feature matrix
- male test as feature matrix

Zur bestimmung der bessten klassifizieren haben wir die methode *try and error* verwendet. Das heit wir haben alle anwendbaren klassifikatoren in der Standarteinstellung auf unsere datenbasis angewandt und danach die ergebisse verglichen.

ein gutes ergebnis bestand darin dass mglichst viele samples eines sprechers dem richtigen sprecher zugeordnet wurden. Also der kalssifikator unter verwendung des *female/male/mixed test set* eine gute treffergenauigkeit aufwies.

dabei hat sich ein klassifikator als besonders gut erwiesen, der ibk. dieser klassifikator erzielte ein ergebnis von 54.94 % Treffergenauigkeit bei dem *mixed test set*. als vergleich haben wir den klassifikator mit dem zweit besten ergebnis mit aufgefhrt, der RandomForest. Die Tabelle:I zeigt das Ergebnis beider klassifikatioren mit den dazugehrigen konfiguration des klassifikators.

Table I. Ergebnisse der Klassifikation des IBK und RandomForest

Datenset	IBK	RandomForest
female test set	54.94%	39.14%
male test set	58.50%	41.7679%
mixed test set	53.16%	33.86%
Konfiguration	-K 1 -W 0 -A	-I 10 -K 0 -S 1

Aus den ergebnissen ist zu entnehmen dass es keinen signifikanten unterschied zwischen der den ergebnisse der sets *female teest set* und *male test set* exisitert. der bestehtnde unterschied lsst sich aus der geringen gre des datensets erklren.

Auerdem ist zu beobachten dass sich die Ergebnisse beioder klassifikatoren verschlechter hat bei erhhung der anzahl von sprechern. Dies ist jedoch ein zu erwartendes ergebniss da der klassifikator nun das sample eines sprechers mit 82 anderen samples statt mit 47 bzw. 34 anderen samples verglcihen muss.

8.1 Authentifizierung

in diesem kaptitel geht es nun um die Aufgabe der *Closed set speaker authentication*. Wobei die akustische Aufnahme eines Sprechers mit der aller anderen mglichen Sprechern verglichen wird und die beste bereinstimmung als Ergebnis ausgegeben wird.[Beigi 2011] Zu Beachten ist, dass hier im Gegensatz zu der *Open set speaker authentication* es in jedem Fall zu einem Ergebnis kommt.

die aufgabe besteht nun darin das ergebnis der klassifikation der samples zu interpretieren. dafr betrachteten fr jeden sprecher die verteilung seiner samples. Hierbei wurde ein sprecher richtig erkannt wenn bei ihm die grte menge samples zugeordnet wurden. dies bedeutet dass wir auch mit einemgeringen anzahl richtig klassifizierter samples einen sprecher authentifizieren konnten solang die brigen samples gleichmig verteilt waren.

dieses verfahren wurde bei allen sprechern angewandt und das entstandende ergebniss ist zu sahen in Tabelle:II.

Table II. Ergebnisse der Authentifizierung

Datenset	Gesamt	Richtig	Falsch	Anteil
female test set	35	33	2	94.29%
male test set	48	46	2	95.83%
mixed test set	83	79	4	95.18%

Das Ergebnis von 95.18% richtig erkannten sprechern ist gut und entspricht damit dem *state-of-the-art*.[Beigi 2011] Im vergleich dazu wurde im hyke-projekt ein ergebnis von 95% erreicht, welches mit unseren nahezu identisch ist.[Azarias Reda 2011]

9. DODDINGTONS ZOO

Doddingtons Zoo geht es darum dass Sprecher ein unterschiedliches Verhalten bezglich den Erfolg ihrer Authentifizierung aufzeigen. [George Doddington 1998] Dadurch lassen sich Sprecher in vier Kategorien unterscheiden die jeweils von einem Tier repräsentiert werden.

Beschreibung dieser Kategorieren:

- **Sheeps:** auerordentlich *leicht* von dem System erkannt, die Mehrheit der Sprecher gehrt dieser Kategorie an
- **Goats:** auerordentlich *schwer* von dem System erkannt
- **Lambs:** auerordentlich *verwundbar* gegenber Nachahmung
- **Wolves:** auerordentlich *erfolgreich* bei der Nachahmung anderen Sprecher

[Prof. Dr. Jana Dittmann 2014]

Nun sollte eine Projektion dieser Kategorien auf die Ergebnisse unserer Klassifizierung vollzogen werden. Als Entscheidungsgrundlage dienten nun nicht nur die erfolgreich klassifizierten Samples sondern auch die Verteilung der falsch klassifizierten Samples. Um die Projektion umzusetzen haben wir folgendes Schema erarbeitet und auf unsere Datenbasis angewandt.

verwendetes Schema zur Kategorisierung der Sprecher:

- **Sheeps:** viele richtig klassifizierte Samples
- **Goats:** wenig richtig klassifizierte Samples
- **Lambs:** viele Sample von anderen Sprechern wurden diesem Sprecher zugeordnet
- **Wolves:** viele Samples bei wenigen anderen Sprechern zugeordnet

Dieser Ansatz wurde auf das Ergebnis der Klassifikation mit dem Klassifikator IBK auf das *mixed test set* angewandt. Das Ergebnis dieser Projektion ist zu sehen in Tabelle:III.

Table III. Ergebnisse der Kategorisierung nach Doddingtons Zoo

Animal	female	male	mixed	Anteil
Sheep	32	43	75	90.36%
Goat	2	2	4	4.82%
Lamb	1	1	1	2.41%
Wolf	0	2	2	2.41%

Das Ergebnis der kategorisierung zeigt dass der groteil unsere Sprecher **Sheeps** sind. Wie zu erwarten wurden unsere falsch authentifizierten Sprecher zu der Kategorie der **Goat** zugeordnet.

10. ZUSAMMENFASSUNG

Es ist mglich einen Menschen anhand seiner Stimme zu identifizieren. Dies erffnet Anwendungsbereiche, die mit anderen biometrischen Verfahren nicht mglich sind. Ein Beispiel ist die Identifizierung einer Person ber das Telefon. Wir konnten in unseren Experimenten, mit geringem Aufwand, 96,68 Prozent der Personen eindeutig Identifizieren. Bei den 3,32 Prozent der nicht identifizierten lag eine schlechte Datenbasis vor. Das heit der Erfolg bei der Identifizierung hngt signifikant von der Datenbank ab.

11. FUTURE WORK

—future stuff

REFERENCES

Edward Cutrell Azarias Reda, Saurabh Panjwani. 2011. Hyke: A Low-cost Remote Attendence Tracking System for Developing Regions. *Networked System for Developing Regions* (2011).

Homayoon Beigi. 2011. *Fundamentals of Speaker Recognition*. Springer Science+Business Media.

Alvin Martin-Mark Przybocki Douglas Reynolds George Doddington, Walter Liggett. 1998. SHEEP, GOATS, LAMBS and WOLVES - A Statistical Analysis of Speaker Performance in the NIST 1998 Skeaper Recognition Evaluation. *National Institute of Standards and Technology* (1998).

Prof. Dr.-Ing. Claus Vielhauer Prof. Dr. Jana Dittmann. 2014. *Biometrics and Security - Lecture*. Faculty of Computer Science, Institute of Technical and Business Information Systems, Advanced Multimedia and Secruity Lab (AMSL).

List of Tables

APPENDIX

A. TASK DESCRIBTION

Run your prototype on the collected data and perform a performance evaluation with your prototype. The evaluation must include:
The evaluation must include:

- Closed set speaker authentication on the Hyke speech database
- Compare the results achieved (in terms of authentication performance) to th results persented in
- A projection of the samples in your data set to the characters of 'Doddingtons Zoo'