

Biometrics and Security Speaker Recognition

Jonas Marquardt und Maik Riestock

Otto von Guericke University Magdeburg - Advanced Multimedia and Security 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 Aufnahmegerät ist ein handelsübliches Mikrophon ausreichend. In dieser Übung 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 Leh-
res stamm von einer dezentralen Bildungseinrichtung aus Indien
die es dazu verwendet um nachzuvollziehen ob 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 kontrolle 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 % erkenntungs 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 abgeschlossen werden:

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

Bei den Arbeitsschritten in unserem Projekt haben wir uns an den vorgestellten Ablauf eines Authentifizierungsprozesses aus der Vorlesung von Prof. Dr. Jana Dittmann orientiert.

Bei unseren arbeitsschritten haben wir uns dabei an den von ... vorgestellten prozess gehalten [?]

dieser prozesskette haben wir noch den schritt des post-processing hinzugefügt, weil das entfernen von noise aus feature mit weka angeboten hat.

für 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:??.
- **WEKA**, ist eine Sammlung von Algorithmen des Maschinellen Lernens für Aufgaben im Bereich des Data-Mining.[?] Anwendung fand dieses Tool in den Kapiteln:?? und 6.

4. DATABASE

Die Datenbasis wurde dem Hyke-Projekt entnommen. Sie kann unter folgender URL heruntergeladen werden: Sie umfasst Aufnahmen von 83 verschiedenen Sprechern, davon 48 männlich und 35 weiblich. Von jeder Person gibt es fünf Aufnahmen in denen Abfolgen verschiedener Ziffern gesprochen werden. Die Sprache dabei ist Englisch. Die Länge der Aufnahmen liegt zwischen 5 und 35 Sekunden. Es gibt auch Aufnahmen, die keine Stimme enthalten. Die Stimmen wurden über das Telefon aufgenommen und bieten daher eine geringere Bandbreite als die menschliche Stimme hat. Bei den Sprechern handelt es sich um Indier mit verschiedenen Hintergründen. Die Aufnahmen enthalten teilweise Hintergrundgeräusche, vom leisen Rauschen bis zu Gesprächen 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 können, wurden die Datenbank in sechs Sets unterteilt. Hierfür wurde das Set mit Sprechern von beiden Geschlechtern, *mixed set*, aufgeteilt in zwei Sets mit ausschließlich Stimmen von weiblichen Sprechern, *female set*, und mit ausschließlich männlichen Sprechern, *male set*.

Zusätzlich benötigen die Klassifikatoren zwei verschiedene Sets von Daten. Mit dem einen Set wird das Modell trainiert, hier *train set*, und mit dem anderen 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. KLASSIFIKATION

In diesem Kapitel geht es um die Klassifikation unserer Daten.

Klassifizierung ist definiert durch

Für diese Aufgabe haben wir das Tool WEKA verwendet, welches in Kapitel 3 vorgestellt wurde.

Unsere Ausgangsdaten für die Klassifikatoren sind die aufbereiteten Features, welche in Kapitel ?? 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 besten Klassifizierung haben wir die Methode *try and error* verwendet. Das heißt wir haben alle anwendbaren Klassifikatoren in der Standardeinstellung auf unsere Datenbasis angewandt und danach die Ergebnisse verglichen.

Ein gutes Ergebnis bestand darin, dass möglichst viele Samples eines Sprechers dem richtigen Sprecher zugeordnet wurden. Also der Klassifikator 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 zweitbesten Ergebnis mit aufgeführt, der *RandomForest*. Die Tabelle I zeigt das Ergebnis beider Klassifikatoren mit den dazugehörigen Konfigurationen des Klassifikators.

Aus den Ergebnissen ist zu entnehmen, dass es keinen signifikanten Unterschied zwischen den Ergebnissen der Sets *female test set* und *male test set* existiert. Der bestehende Unterschied lässt sich aus der geringen Größe des Datensets erklären.

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

Außerdem ist zu beobachten, dass sich die Ergebnisse beider Klassifikatoren verschlechtert haben bei Erhöhung der Anzahl von Sprechern. Dies ist jedoch ein zu erwartendes Ergebnis, da der Klassifikator nun das Sample eines Sprechers mit 82 anderen Samples statt mit 47 bzw. 34 anderen Samples vergleichen muss.

6.1 Authentifizierung

In diesem Kapitel geht es nun um die Aufgabe der *Closed set speaker authentication*. Wobei die akustische Aufnahme eines Sprechers mit der aller anderen möglichen Sprechern verglichen wird und die beste Übereinstimmung 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. Dafür betrachten wir für jeden Sprecher die Verteilung seiner Samples. Hierbei wurde ein Sprecher richtig erkannt, wenn bei ihm die größte Menge Samples zugeordnet wurden. Dies bedeutet, dass wir auch mit einer geringen Anzahl richtig klassifizierter Samples einen Sprecher authentifizieren konnten, solange die übrigen Samples gleichmäßig verteilt waren.

Dieses Verfahren wurde bei allen Sprechern angewandt und das entstandene Ergebnis ist zu sehen 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]

7. DODDINGTONS ZOO

Doddingtons Zoo geht es darum, dass Sprecher ein unterschiedliches Verhalten bezüglich 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 Kategorien:

- **Sheeps:** außerordentlich *leicht* von dem System erkannt, die Mehrheit der Sprecher gehört dieser Kategorie an
- **Goats:** außerordentlich *schwer* von dem System erkannt
- **Lambs:** außerordentlich *verwundbar* gegenüber Nachahmung
- **Wolves:** außerordentlich *erfolgreich* bei der Nachahmung anderer 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.

8. CONCLUSION

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.

9. FUTURE WORK

—future stuff

REFERENCES

- Edward Cutrell Azarias Reda, Saurabh Panjwani. 2011. Hyke: A Low-cost Remote Attendance 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 Przyboc 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 Security Lab (AMSL).

List of Tables

I	Ergebnisse der Klassifikation des IBK und RandomForest	2
II	Ergebnisse der Authentifizierung	2
III	Ergebnisse der Kategorisierung nach Doddingtons Zoo	3

APPENDIX

A. TASK DESCRIPTION

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 the results presented in
- A projection of the samples in your data set to the characters of 'Doddingtons Zoo'