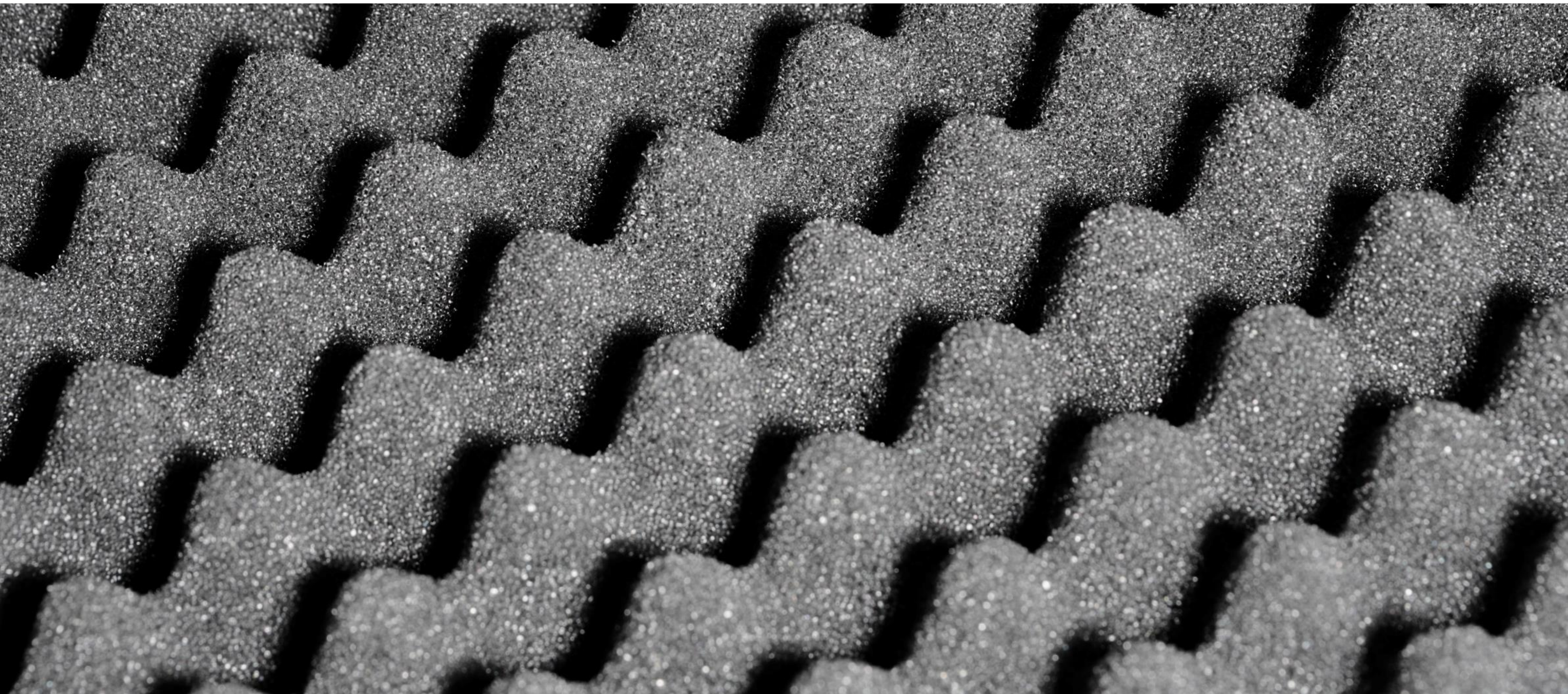
Zürcher Hochschule
für Angewandte Wissenschaften

Deep-learning-based speaker recognition
…and the problem of modeling supra-segmental temporal features

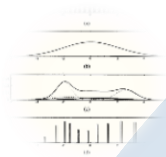
Lecture series on Speech and Text Technologies, University of Zurich Computational Linguistics, Nov 29, 2021

Thilo Stadelmann

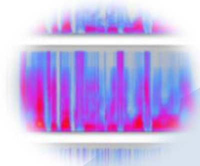


Agenda

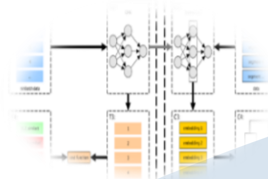
1. ZHAW CAI



2. Once upon a time

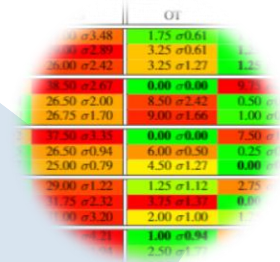


3. The problem



4. Enter deep learning

5. Problem solved?



6. Surprise, surprise?

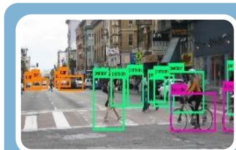
Zürcher Hochschule
für Angewandte Wissenschaften

The ZHAW Centre for Artificial Intelligence

Foundation: Machine Learning & Deep Learning
Cross-cutting concerns: Ethics, Generality

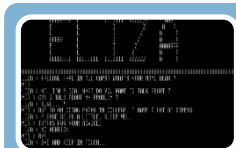
Autonomous Learning Systems

- *Reinforcement Learning*
- *Multi-Agent Systems*
- *Embodied AI*



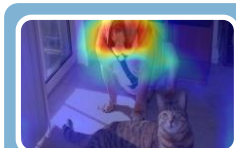
Computer Vision, Perception and Cognition

- *Pattern Recognition*
- *Machine Perception*
- *Neuromorphic Engineering*



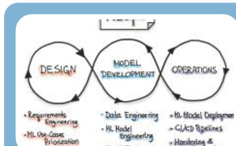
Natural Language Processing

- *Dialogue Systems*
- *Text Analytics*
- *Spoken Language Technologies*



Trustworthy AI

- *Explainable AI*
- *Robust Deep Learning*
- *AI & Society*



AI Engineering

- MLOps
- Data-Centric AI
- Continuous Learning

Areas of application & cooperation: medicine & health, IoT, robotics, AI ethics & regulation, predictive maintenance, automatic quality control, document analysis, chat bots, biometrics, earth observation, digital farming, meteorology, autonomous driving, further data science use cases in industries like manufacturing / finance / insurance / commerce / transportation / energy etc.

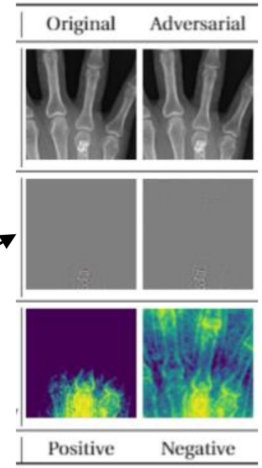
Computer Vision, Perception & Cognition Group

Machine learning-based Pattern Recognition

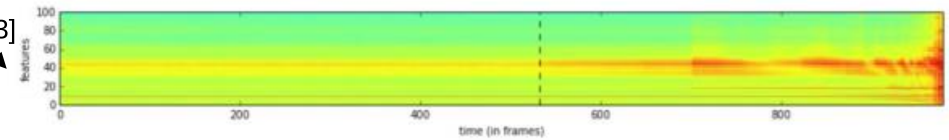
- Robust applications
- Biometrics
- Document Analysis
- Learning to act

[1]

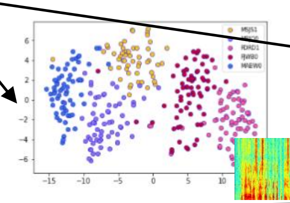
[2]



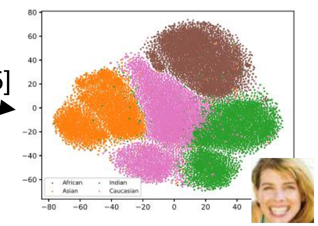
[3]



[4]



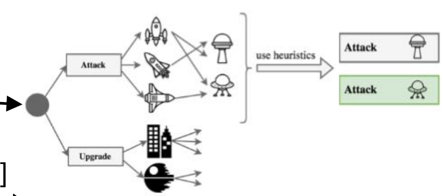
[5]



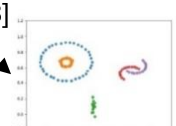
[7]

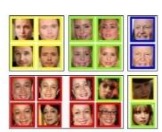
[6]

[1]

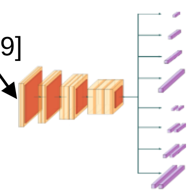


[8]

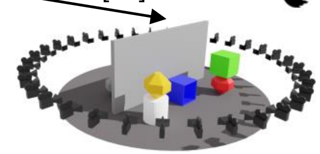




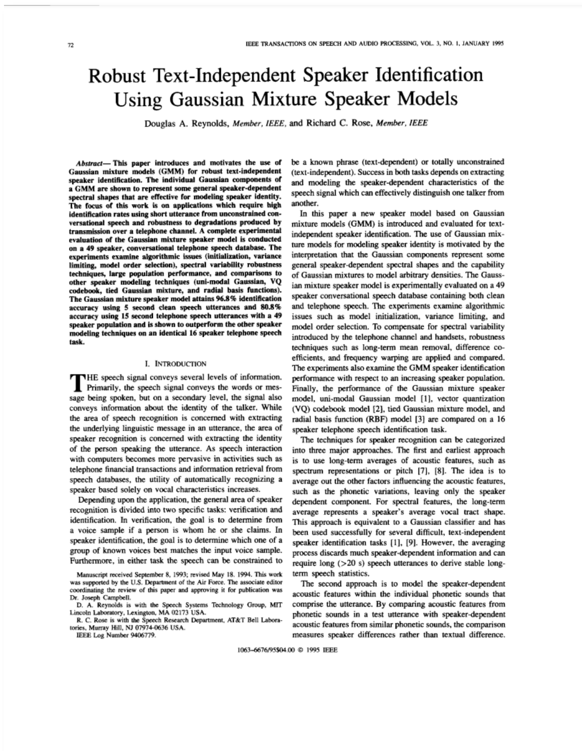
[9]



[10]



ONCE UPON A TIME



72 IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 3, NO. 1, JANUARY 1995

Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models

Douglas A. Reynolds, *Member, IEEE*, and Richard C. Rose, *Member, IEEE*

***Abstract*— This paper introduces and motivates the use of Gaussian mixture models (GMM) for robust text-independent speaker identification. The individual Gaussian components of a GMM are shown to represent some general speaker-dependent spectral shapes that are effective for modeling speaker identity. The focus of this work is on applications which require high identification rates using short utterance from unconstrained conversational speech and robustness to degradations produced by transmission over a telephone channel. A complete experimental evaluation of the Gaussian mixture speaker model is conducted on a 49 speaker, conversational telephone speech database. The experiments examine algorithmic issues (initialization, variance limiting, model order selection), spectral variability robustness techniques, large population performance, and comparisons to other speaker modeling techniques (uni-modal Gaussian, VQ codebook, tied Gaussian mixture, and radial basis functions). The Gaussian mixture speaker model attains 96.8% identification accuracy using 5 second clean speech utterances and 80.8% accuracy using 15 second telephone speech utterances with a 49 speaker population and is shown to outperform the other speaker modeling techniques on an identical 16 speaker telephone speech task.**

I. Introduction

THE speech signal conveys several levels of information. Primarily, the speech signal conveys the words or message being spoken, but on a secondary level, the signal also conveys information about the identity of the talker. While the area of speech recognition is concerned with extracting the underlying linguistic message in an utterance, the area of speaker recognition is concerned with extracting the identity of the person speaking the utterance. As speech interaction with computers becomes more pervasive in activities such as telephone financial transactions and information retrieval from speech databases, the utility of automatically recognizing a speaker based solely on vocal characteristics increases.

Depending upon the application, the general area of speaker recognition is divided into two specific tasks: verification and identification. In verification, the goal is to determine from a voice sample if a person is whom he or she claims. In speaker identification, the goal is to determine which one of a group of known voices best matches the input voice sample. Furthermore, in either task the speech can be constrained to be a known phrase (text-dependent) or totally unconstrained (text-independent). Success in both tasks depends on extracting and modeling the speaker-dependent characteristics of the speech signal which can effectively distinguish one talker from another.

In this paper a new speaker model based on Gaussian mixture models (GMM) is introduced and evaluated for text-independent speaker identification. The use of Gaussian mixture models for modeling speaker identity is motivated by the interpretation that the Gaussian components represent some general speaker-dependent spectral shapes and the capability of Gaussian mixtures to model arbitrary densities. The Gaussian mixture speaker model is experimentally evaluated on a 49 speaker conversational speech database containing both clean and telephone speech. The experiments examine algorithmic issues such as model initialization, variance limiting, and model order selection. To compensate for spectral variability introduced by the telephone channel and handsets, robustness techniques such as long-term mean removal, difference coefficients, and frequency warping are applied and compared. The experiments also examine the GMM speaker identification performance with respect to an increasing speaker population. Finally, the performance of the Gaussian mixture speaker model, uni-modal Gaussian model [1], vector quantization (VQ) codebook model [2], tied Gaussian mixture model, and radial basis function (RBF) model [3] are compared on a 16 speaker telephone speech identification task.

The techniques for speaker recognition can be categorized into three major approaches. The first and earliest approach is to use long-term averages of acoustic features, such as spectrum representations or pitch [7], [8]. The idea is to average out the other factors influencing the acoustic features, such as the phonetic variations, leaving only the speaker dependent component. For spectral features, the long-term average represents a speaker's average vocal tract shape. This approach is equivalent to a Gaussian classifier and has been used successfully for several difficult, text-independent speaker identification tasks [1], [9]. However, the averaging process discards much speaker-dependent information and can require long (>20 s) speech utterances to derive stable long-term speech statistics.

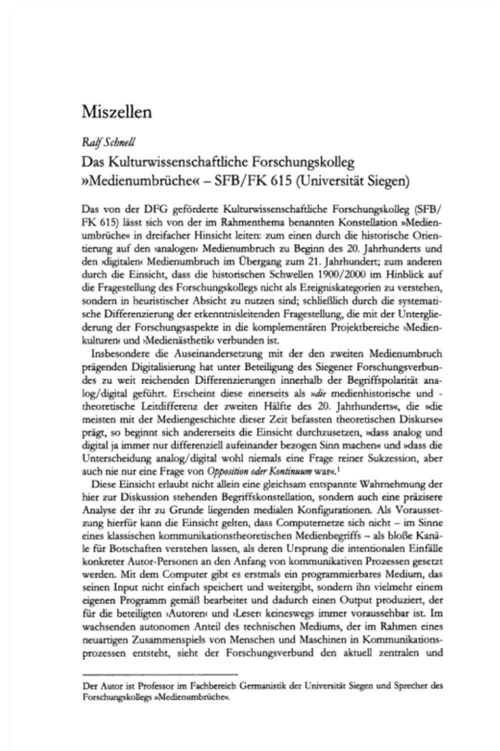
The second approach is to model the speaker-dependent acoustic features within the individual phonetic sounds that comprise the utterance. By comparing acoustic features from phonetic sounds in a test utterance with speaker-dependent acoustic features from similar phonetic sounds, the comparison measures speaker differences rather than textual difference.

Manuscript received September 8, 1993; revised May 18, 1994. This work was supported by the U.S. Department of the Air Force. The associate editor coordinating the review of this paper and approving it for publication was Dr. Joseph Campbell.

D. A. Reynolds is with the Speech Systems Technology Group, MIT Lincoln Laboratory, Lexington, MA 02173 USA.

R. C. Rose is with the Speech Research Department, AT&T Bell Laboratories, Murray Hill, NJ 07974-0636 USA.

IEEE Log Number 9406779.



Miszellen

Ralf Schnell

Das Kulturwissenschaftliche Forschungskolleg »Medienumbrüche« – SFB/FK 615 (Universität Siegen)

Das von der DFG geförderte Kulturwissenschaftliche Forschungskolleg (SFB/FK 615) lässt sich von der im Rahmenthema benannten Konstellation »Medienumbrüche« in dreifacher Hinsicht leiten: zum einen durch die historische Orientierung auf den ›analogen‹ Medienumbruch zu Beginn des 20. Jahrhunderts und den ›digitalen‹ Medienumbruch im Übergang zum 21. Jahrhundert; zum anderen durch die Einsicht, dass die historischen Schwellen 1900/2000 im Hinblick auf die Fragestellung des Forschungskollegs nicht als Ereigniskategorien zu verstehen, sondern in heuristischer Absicht zu nutzen sind; schließlich durch die systematische Differenzierung der erkenntnisleitenden Fragestellung, die mit der Untergliederung der Forschungsaspekte in die komplementären Projektbereiche ›Medienkulturen‹ und ›Medienästhetik‹ verbunden ist.

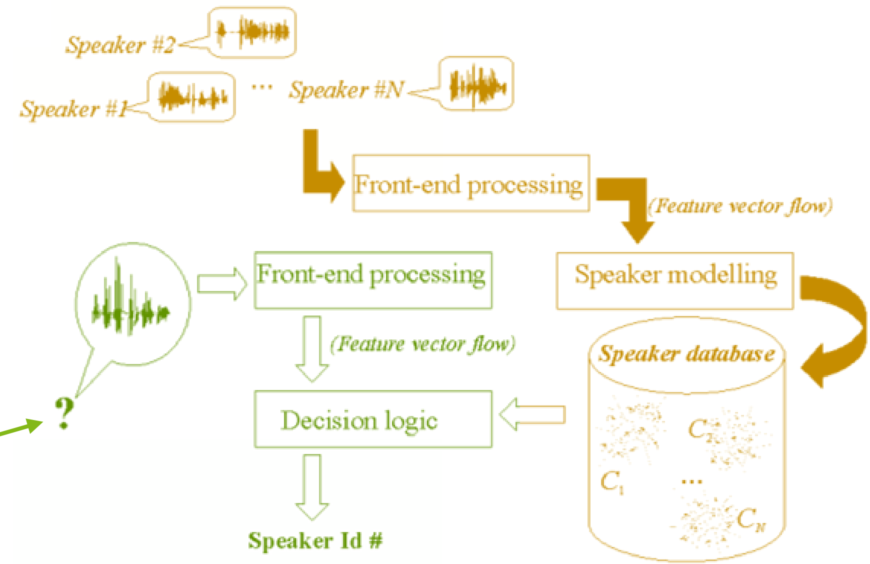
Insbesondere die Auseinandersetzung mit der den zweiten Medienumbruch prägenden Digitalisierung hat unter Beteiligung des Siegener Forschungsverbundes zu weit reichenden Differenzierungen innerhalb der Begriffspolarität analog/digital geführt. Erscheint diese einerseits als »*die* medienhistorische und -theoretische Leitdifferenz der zweiten Hälfte des 20. Jahrhunderts«, die »die meisten mit der Mediengeschichte dieser Zeit befassten theoretischen Diskurse« prägt, so beginnt sich andererseits die Einsicht durchzusetzen, »dass analog und digital ja immer nur differenziell aufeinander bezogen Sinn machen« und »dass die Unterscheidung analog/digital wohl niemals eine Frage reiner Sukzession, aber auch nie nur eine Frage von *Opposition oder Kontinuum* war«.[1]

Diese Einsicht erlaubt nicht allein eine gleichsam entspannte Wahrnehmung der hier zur Diskussion stehenden Begriffskonstellation, sondern auch eine präzisere Analyse der ihr zu Grunde liegenden medialen Konfigurationen. Als Voraussetzung hierfür kann die Einsicht gelten, dass Computernetze sich nicht – im Sinne eines klassischen kommunikationstheoretischen Medienbegriffs – als bloße Kanäle für Botschaften verstehen lassen, als deren Ursprung die intentionalen Einfälle konkreter Autor-Personen an den Anfang von kommunikativen Prozessen gesetzt werden. Mit dem Computer gibt es erstmals ein programmierbares Medium, das seinen Input nicht einfach speichert und weitergibt, sondern ihn vielmehr einem eigenen Programm gemäß bearbeitet und dadurch einen Output produziert, der für die beteiligten ›Autoren‹ und ›Leser‹ keineswegs immer voraussehbar ist. Im wachsenden autonomen Anteil des technischen Mediums, der im Rahmen eines neuartigen Zusammenspiels von Menschen und Maschinen in Kommunikationsprozessen entsteht, sieht der Forschungsverbund den aktuell zentralen und

Der Autor ist Professor im Fachbereich Germanistik der Universität Siegen und Sprecher des Forschungskollegs »Medienumbrüche«.

Scientific media analysis

The task of speaker recognition



Speaker recognition

- **Tell identity** of an **utterances**' speaker
- Typical: score feature-sequence against a speaker model

Three tasks

- Identification: Given **one utterance** and a **set of speaker models**, **find** the actual **speaker**
  (or declare as unknown: open set identification)
- Clustering: Given a **set of utterances**, **sort them** into pure clusters **by voice** identity
  (if set originates from segmenting a longer recording: who spoke when; no prior knowledge of any kind)
- Verfication: Given **two** utterances, **decide if** both are **spoken by same speaker**
  (today's approach to the clustering problem)

Speaker recognition anno 2003: MFCC features and GMM models

Hybrid solution between non-parametric clusters (vector quantization) and compact smoothing (single Gaussian):

- **Smooth approximation** of arbitrary densities
- **Implicit clustering** into broad phonetic classes

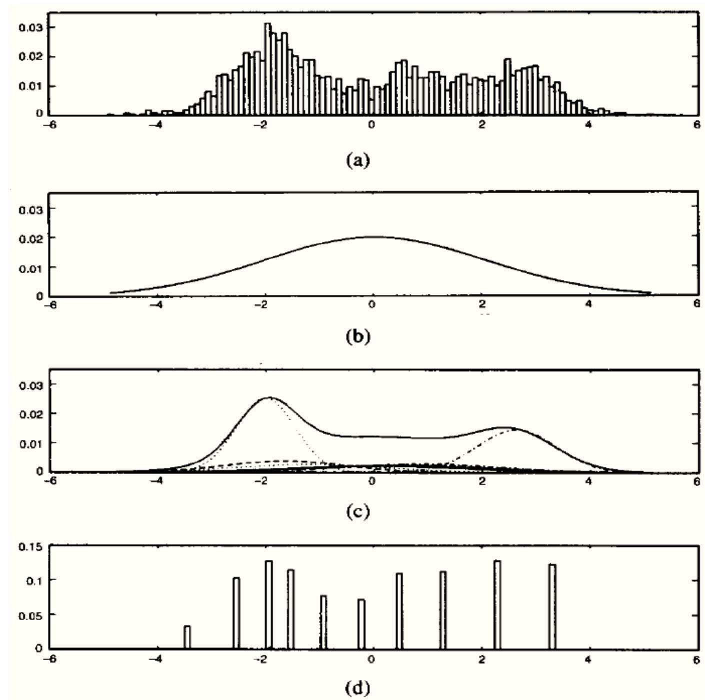
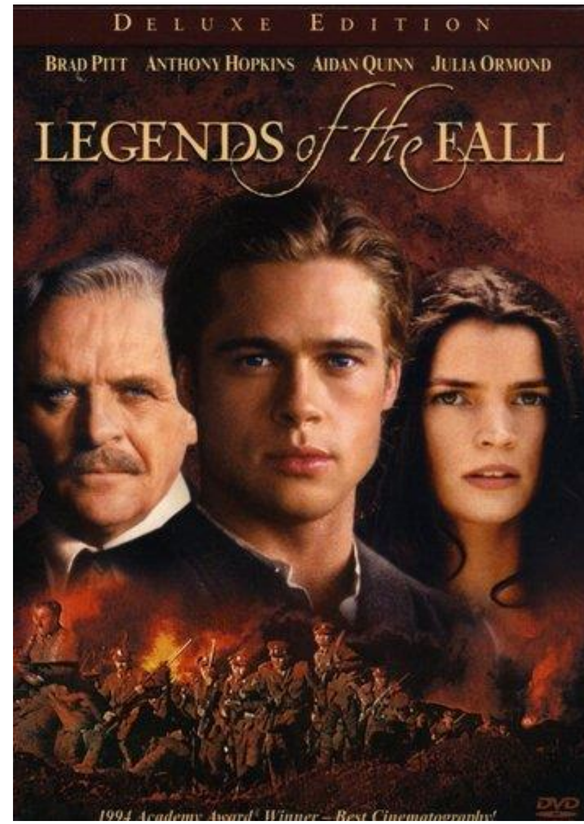


Fig. 3. Comparison of distribution modeling: (a) Histogram of a single cepstral coefficient from a 25 second utterance by a male speaker; (b) maximum likelihood unimodal Gaussian model; (c) GMM and its 10 underlying component densities; (d) histogram of the data assigned to the VQ centroid locations of a 10-element codebook.

GMM comparison with other techniques; from [Reynolds and Rose, 1995].

Results



ok

not ok

THE PROBLEM

Unfolding Speaker Clustering Potential: A Biomimetic Approach

Thilo Stadelmann Bernd Freisleben
Department of Mathematics & Computer Science, University of Marburg
Hans-Meerwein-Str. 3, D-35032 Marburg, Germany
{stadelmann, freisleb}@informatik.uni-marburg.de

ABSTRACT

Speaker clustering is the task of grouping a set of speech utterances into speaker-specific classes. The basic techniques for solving this task are similar to those used for speaker verification and identification. The hypothesis of this paper is that the techniques originally developed for speaker verification and identification are not sufficiently discriminative for speaker clustering. However, the processing chain for speaker clustering is quite large – there are many potential areas for improvement. The question is: *where* should improvements be made to improve the *final* result? To answer this question, this paper takes a biomimetic approach based on a study with human participants acting as an automatic speaker clustering system. Our findings are twofold: it is the stage of modeling that has the highest potential, and information with respect to the temporal succession of frames is crucially missing. Experimental results with our implementation of a speaker clustering system incorporating our findings and applying it on TIMIT data show the validity of our approach.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing; I.5.4 [**Pattern Recognition**]: Applications—*Signal processing, Waveform analysis*

General Terms

Algorithms, Design, Experimentation, Performance

Keywords

Speaker identification, Speaker clustering, Speaker diarization, GMM, MFCC, Temporal context, One-class SVM

1. INTRODUCTION

Recognizing voices automatically is useful for several applications. For example, it supports biometric authentication [64]. It helps making speech recognition robust [20]. It enables search engines to index spoken documents and thus improves retrieval performance [31]. These three examples refer to different subproblems of speaker recognition, namely: speaker verification [49], speaker identification [8] and speaker clustering [28] (or, when regarding the complete process including speech detection and segmentation: speaker diarization [46]).

Speaker verification is the most simple clustering task among these problems: the question is whether a given utterance can be assigned to a given model (identity) – a binary choice. Speaker identification is a $(1 : n+1)$ choice: the question is which (if any) of the given models can the given utterance be paired with? Finally, speaker clustering is a $(m : n)$ problem in which all utterances are equally important and each utterance may be grouped together with any other utterance – or stay alone. Both the number of clusters (speakers) and the actual cluster memberships must be determined automatically.

The speaker verification and identification tasks have been studied extensively in the literature. Using Mel Frequency Cepstral Coefficients (MFCCs) [12] as parametric speech features and Gaussian Mixture Models (GMMs) [49] (with more recent modifications [48]) as speaker models has become the quasi-standard, although other methods have been proposed [16]. This is due to quite satisfactory results with just moderate demands for the data: the utterances should be relatively noise-free (telephone speech works) and long enough (minimum 10 seconds, better more than 30 seconds per utterance) [62]. The canonical example is the experiment in Reynolds' classic paper on GMMs [47]: The 630 speakers of the TIMIT database [19] are split into a training set (8 sentences per speaker concatenated to one utterance) and a separate test set (2 sentences per speaker form one utterance). Each sentence is approximately 3 seconds long. The utterances are transformed to MFCC feature vectors. For the 630 training utterances, GMMs with 32 mixtures are built a priori, then an identification experiment is run for the 630 test utterances. It yields a satisfactory 0.5% closed set identification error.

Speaker clustering has also been studied extensively for more than a decade [24]. The basic techniques used for speaker clustering are largely along the lines of the previously discussed verification/identification techniques: MFCC features are modeled by GMMs [28][60]. Upon this, a step-by-step scheme using agglomerative hierarchical clustering is usually built using some metric (often the Generalized Likelihood Ratio (GLR)) and a termination criterion (frequently based on the Bayesian Information Criterion (BIC))

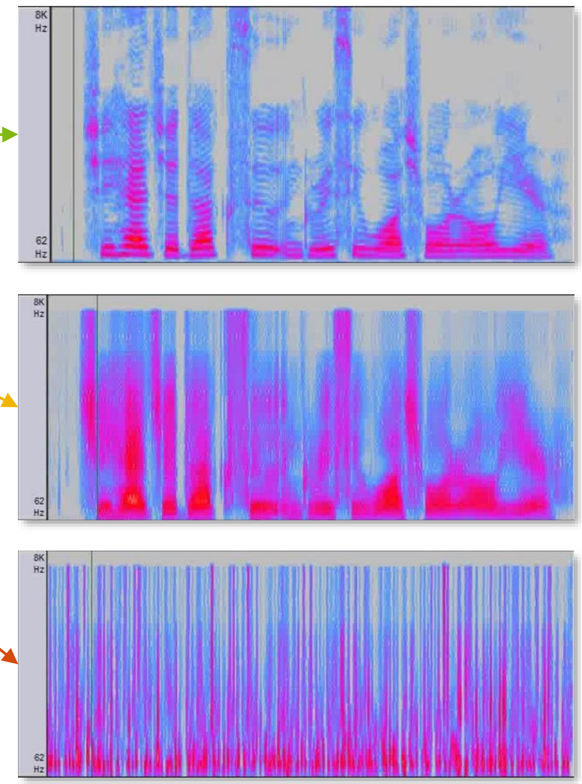
MM'09, October 19–24, 2009, Beijing, China.

185

What GMMs do not capture

Re-synthesizing speech from intermediate stages of the speaker modeling pipeline

- Original utterance
- Resynthesized feature vectors (MFCCs)
- Resynthesized MFCCs from GMM



Implication

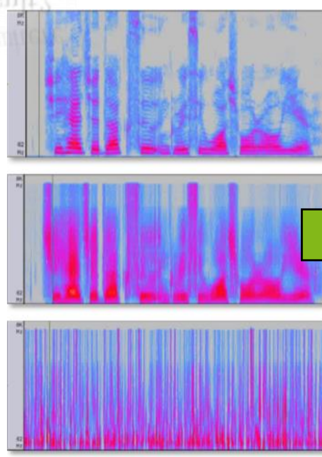
- **Temporal context isn't modeled** by GMMs

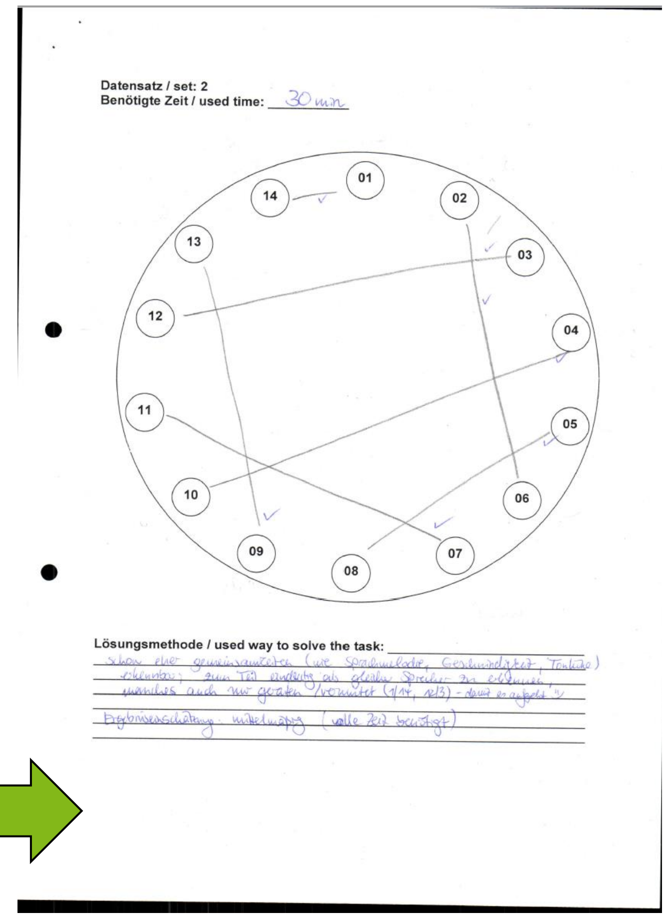
Searching for the bottleneck

For the 630 training utterances, GMMs with 32 mixtures are built a priori, then an identification experiment is run for the 630 test utterances. It yields a satisfactory 0.5% closed set identification error.

[34]. Evaluations typically concentrate on data sets built from broadcast news/shows and meeting recordings, where diarization error rates ranging from 8% to 24% are reported [28][34][45]. These results are confirmed by more recent

The hypothesis of this paper is: the techniques originally developed for speaker verification and identification are not suitable for speaker clustering, taking into account the escalated difficulty of the latter task. However, the processing chain for speaker clustering is quite large – there are many potential areas for improvement. The question is: *where should improvements be made to improve the final result?*

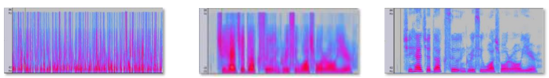




Stadelmann & Freisleben (2009). *«Unfolding Speaker Clustering Potential: A Biomimetic Approach»*. ACMMM'2009.

Zürcher Hochschule
für Angewandte Wissenschaften

Bottleneck: detected



feature	#dataset 1	#dataset 2	#dataset 3
rhythm/velocity	7	11	8
pitch	7	11	7
timbre/sound	3	6	14
perceived gender	0	2	13
perceived age	0	0	5
visual imagination	0	1	3
volume	2	1	0
nasalization	0	1	0
holistic judgment	0	0	1

The interpretation of our results has shown that it is the stage of modeling that bears the highest potential: the inclusion of temporal context information among feature vectors is what is crucially missing there. Furthermore, the inclusion

context vector. This corresponds to a syllable length of 130 ms and is found to best capture speaker specific sounds in informal listening experiments over a range of 32–496 ms (in intervals of 16 ms). Our context vector step is one orig-

Stadelmann & Freisleben (2009). *«Unfolding Speaker Clustering Potential: A Biomimetic Approach»*. ACMMM'2009.

Proof of concept

SVM-based “time model”

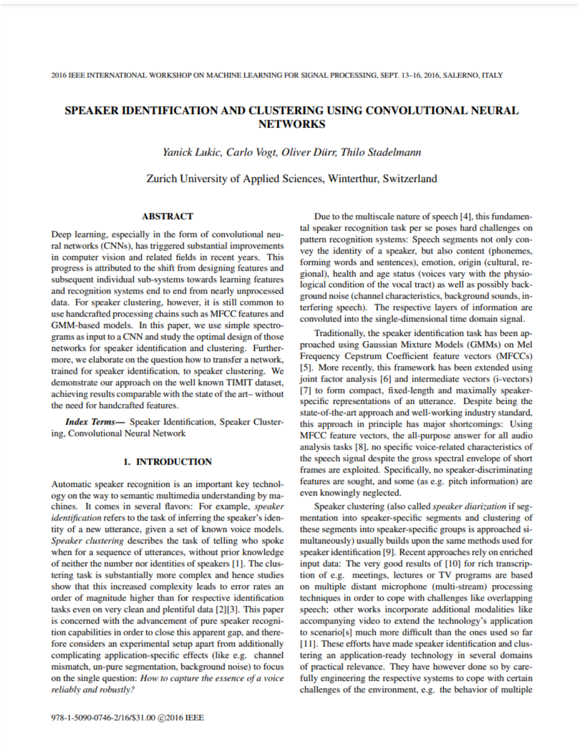
1. Speaking rate normalization (i.e., **removal** of **too similar subsequent frames**)
2. **Transformation** of basic features **to trajectories** (i.e., concatenation of feature vectors in a segment)
3. Estimation of the support of the trajectory’s distribution in time and frequency (using a **ν-SVM**)
4. Comparison of different trajectory models (by **scoring features** of one utterance **against model** of other)

approach	runtime [m]	MR	DER
baseline	2.70	0.125	0.04527
baseline+δ	4.95	0.65	0.5833
baseline+δ+$\delta\delta$	7.98	0.5	0.1731
baseline+F_0	2.15	0.2625	0.1551
baseline+δ+F_0	4.98	0.4875	0.4084
baseline+δ+$\delta\delta$+F_0	7.97	0.7125	0.6176
time model	523.13	0.0625	0.01962

-50% missclassification rate!

- Baseline: GMM per utterance on MFCCs
- Time model: One-class SVM per utterance on concatenated MFCCs of whole segments

ENTER DEEP LEARNING



2016 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 13–16, 2016, SALERNO, ITALY

SPEAKER IDENTIFICATION AND CLUSTERING USING CONVOLUTIONAL NEURAL NETWORKS

Yanick Lukic, Carlo Vogt, Oliver Dürr, Thilo Stadelmann

Zurich University of Applied Sciences, Winterthur, Switzerland

ABSTRACT

Deep learning, especially in the form of convolutional neural networks (CNNs), has triggered substantial improvements in computer vision and related fields in recent years. This progress is attributed to the shift from designing features and subsequent individual sub-systems towards learning features and recognition systems end to end from nearly unprocessed data. For speaker clustering, however, it is still common to use handcrafted processing chains such as MFCC features and GMM-based models. In this paper, we use simple spectrograms as input to a CNN and study the optimal design of those networks for speaker identification and clustering. Furthermore, we elaborate on the question how to transfer a network, trained for speaker identification, to speaker clustering. We demonstrate our approach on the well known TIMIT dataset, achieving results comparable with the state of the art– without the need for handcrafted features.

***Index Terms*—** Speaker Identification, Speaker Clustering, Convolutional Neural Network

1. INTRODUCTION

Automatic speaker recognition is an important key technology on the way to semantic multimedia understanding by machines. It comes in several flavors: For example, *speaker identification* refers to the task of inferring the speaker's identity of a new utterance, given a set of known voice models. *Speaker clustering* describes the task of telling who spoke when for a sequence of utterances, without prior knowledge of neither the number nor identities of speakers [1]. The clustering task is substantially more complex and hence studies show that this increased complexity leads to error rates an order of magnitude higher than for respective identification tasks even on very clean and plentiful data [2][3]. This paper is concerned with the advancement of pure speaker recognition capabilities in order to close this apparent gap, and therefore considers an experimental setup apart from additionally complicating application-specific effects (like e.g. channel mismatch, un-pure segmentation, background noise) to focus on the single question: *How to capture the essence of a voice reliably and robustly?*

Due to the multiscale nature of speech [4], this fundamental speaker recognition task per se poses hard challenges on pattern recognition systems: Speech segments not only convey the identity of a speaker, but also content (phonemes, forming words and sentences), emotion, origin (cultural, regional), health and age status (voices vary with the physiological condition of the vocal tract) as well as possibly background noise (channel characteristics, background sounds, interfering speech). The respective layers of information are convoluted into the single-dimensional time domain signal.

Traditionally, the speaker identification task has been approached using Gaussian Mixture Models (GMMs) on Mel Frequency Cepstrum Coefficient feature vectors (MFCCs) [5]. More recently, this framework has been extended using joint factor analysis [6] and intermediate vectors (i-vectors) [7] to form compact, fixed-length and maximally speaker-specific representations of an utterance. Despite being the state-of-the-art approach and well-working industry standard, this approach in principle has major shortcomings: Using MFCC feature vectors, the all-purpose answer for all audio analysis tasks [8], no specific voice-related characteristics of the speech signal despite the gross spectral envelope of short frames are exploited. Specifically, no speaker-discriminating features are sought, and some (as e.g. pitch information) are even knowingly neglected.

Speaker clustering (also called *speaker diarization* if segmentation into speaker-specific segments and clustering of these segments into speaker-specific groups is approached simultaneously) usually builds upon the same methods used for speaker identification [9]. Recent approaches rely on enriched input data: The very good results of [10] for rich transcription of e.g. meetings, lectures or TV programs are based on multiple distant microphone (multi-stream) processing speech; other works incorporate additional modalities like accompanying video to extend the technology's application to scenario[s] much more difficult than the ones used so far [11]. These efforts have made speaker identification and clustering an application-ready technology in several domains of practical relevance. They have however done so by carefully engineering the respective systems to cope with certain challenges of the environment, e.g. the behavior of multiple

978-1-5090-0746-2/16/$31.00 ©2016 IEEE

2017 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 25–28, 2017, TOKYO, JAPAN

LEARNING EMBEDDINGS FOR SPEAKER CLUSTERING BASED ON VOICE EQUALITY

Yanick X. Lukic, Carlo Vogt, Oliver Dürr, and Thilo Stadelmann

Zurich University of Applied Sciences, Winterthur, Switzerland

ABSTRACT

Recent work has shown that convolutional neural networks (CNNs) trained in a supervised fashion for speaker identification are able to extract features from spectrograms which can be used for speaker clustering. These features are represented by the activations of a certain hidden layer and are called embeddings. However, previous approaches require plenty of additional speaker data to learn the embedding, and although the clustering results are then on par with more traditional approaches using MFCC features etc., room for improvements stems from the fact that these embeddings are trained with a surrogate task that is rather far away from segregating unknown voices - namely, identifying few specific speakers.

We address both problems by training a CNN to extract embeddings that are similar for equal speakers (regardless of their specific identity) using weakly labeled data. We demonstrate our approach on the well-known TIMIT dataset that has often been used for speaker clustering experiments in the past. We exceed the clustering performance of all previous approaches, but require just 100 instead of 590 unrelated speakers to learn an embedding suited for clustering.

***Index Terms*—** Speaker Clustering, Speaker Recognition, Convolutional Neural Network, Speaker Embedding

1. INTRODUCTION

Speaker clustering handles the *"who spoke when"* challenge in a given audio recording without knowing how many and which speakers are present in the audio signal. It is called speaker diarization when the task of segmenting the audio stream into speaker-specific segments is handled simultaneously [1]. The problem of speaker clustering is eminent in digitizing audio archives like e.g. recordings of lectures, conferences or debates [2]. For their quantitative indexing, automatic extraction of key figures like number of speakers or talk time per person is important. This further facilitates automatic transcripts using existing speech recognition procedures, based on the accurate automatic assignment of speech utterances to groups that each represent a (previously unknown) speaker.

The lack of knowledge of the number and identity of speakers leads to a much more complex problem compared to the related tasks of speaker verification and speaker identification, and in turn to less accurate results. One reason is that well-known speech features and models, originally fitted to the latter tasks, might not be adequate for the more complex clustering task [3]. The use of deep learning methods offers a solution [4]: in contrast to classical approaches (e.g., based on MFCC features and GMM models [5]), where general features and models are designed manually and independently for a wide variety of tasks, deep models learn hierarchies of suitable representations for the specific task at hand [6]. Especially convolutional neural networks (CNNs) have proven to be very useful for pattern recognition tasks mainly on images [7], but also on sounds [8]. Previous work [9] has shown that CNNs are able to learn a voice-specific vector representation (embedding) suitable for clustering when trained for the surrogate task of speaker identification. The authors report state of the art results for speaker clustering using an embedding learned from 590 different speakers.

In this paper, we investigate a novel training approach for CNNs for speaker clustering that learns embeddings more directly based on pairwise voice equality information of speech snippets (i.e., the binary information if the two snippets come from the same speaker or not). This weak labeling is neither a fitting to particular individuals, nor depending on hard to obtain voice similarity measures (i.e., real-valued distances amongst snippets). For evaluation, we focus on the pure speaker clustering performance, given its role as a performance bottleneck in the complete diarization process [3]. Section 2 reviews related work and introduces our approach. Section 3 reports on our results that not only reach state of the art consistently with one approach and improve the clustering quality in certain scenarios, but also reduce the necessary amount of pre-training data to 17%. We also report on a number of experiments in order to give insight on which part of our system is responsible for the improved results. We conclude the paper with an outlook in section 4.

2. LEARNING SPEAKER DISSIMILARITY

2.1. Related work

The design of CNNs makes it possible to recognize patterns in minimally preprocessed digital images or other data with

978-1-5090-6341-3/17/$31.00 ©2017 IEEE

Capturing Suprasegmental Features of a Voice with RNNs for Improved Speaker Clustering

Thilo Stadelmann[1], Sebastian Glinski-Haefeli[1], Patrick Gerber[1], and Oliver Dürr[1,2]

[1] ZHAW Datalab, Zurich University of Applied Sciences, Winterthur, Switzerland
[2] Institute for Optical Systems, Konstanz University of Applied Sciences, Germany
stdm@zhaw.ch, sebastian.glinski@gmail.com, gerber.pat@gmail.com, oliver.duerr@gmail.com

Abstract. Deep neural networks have become a veritable alternative to classic speaker recognition and clustering methods in recent years. However, while the speech signal clearly is a time series, and despite the body of literature on the benefits of prosodic (suprasegmental) features, identifying voices has usually not been approached with sequence learning methods. Only recently has a recurrent neural network (RNN) been successfully applied to this task, while the use of convolutional neural networks (CNNs) (that are not able to capture arbitrary time dependencies, unlike RNNs) still prevails. In this paper, we show the effectiveness of RNNs for speaker recognition by improving state of the art speaker clustering performance and robustness on the classic TIMIT benchmark. We provide arguments why RNNs are superior by experimentally showing a "sweet spot" of the segment length for successfully capturing prosodic information that has been theoretically predicted in previous work.

Keywords: speaker clustering · speaker recognition · recurrent neural network

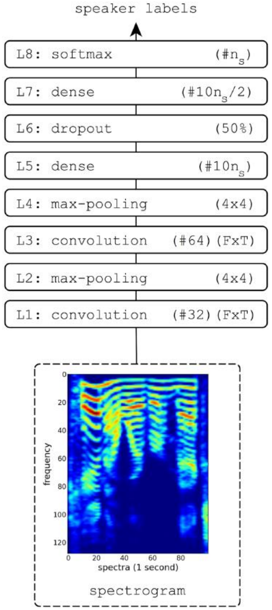
1 Introduction

Automatic speaker recognition comes in many flavors, of which speaker clustering is the most unconstrained and hence the most difficult one [3,41]. It can be defined as the task of judging if two short utterances come from the same (previously unknown) speaker, and thus forms a suitable benchmark for the general ability of a system to capture what makes up a voice: speaker clustering can only be solved satisfactory by regarding all available cues in the utterances themselves. This distinguishes speaker clustering from a more complex experimental setup like e.g. speaker diarization, where engineering a complex system of many components has a not negligible influence on the final result besides the pure voice modeling [2]; and for example from speaker identification, where more available data enables the creation of models that work well just because of the sheer amount of collected training statistics [34]. Previous work [41] hence suggests that the bottleneck for speaker clustering performance lies in exploiting the supra-frame

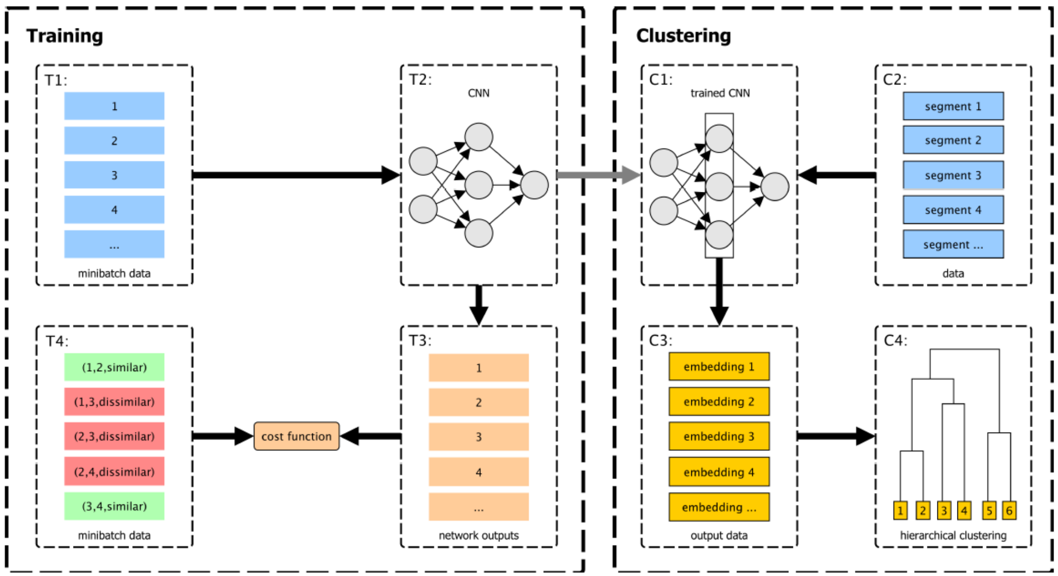
Zürcher Hochschule
für Angewandte Wissenschaften

Exploiting time information with deep learning

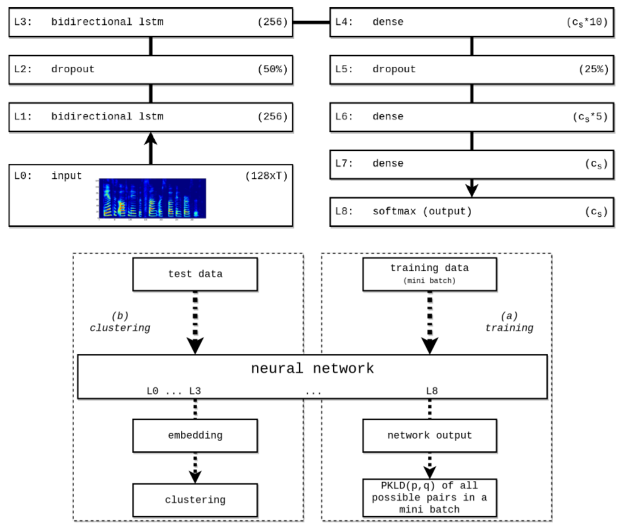
CNN (MLSP'16)



CNN & clustering-loss (MLSP'17)



RNN & clustering-loss (ANNPR'18)



Method	MR	MR (legacy)
RNN /w PKLD	**2.19%** $\left(\frac{1.25\%+2.5\%+1.25\%+3.75\%}{4}\right)$	**4.38%** (average of 4 runs)
CNN /w PKLD [24]	-	**5%**
CNN /w cross entropy [23]	-	**5%**
ν-SVM [40]	6.25%	-
GMM/MFCC [40]	12.5%	-

Lukic, Vogt, Dürr & Stadelmann (2016). «Speaker Identification and Clustering using Convolutional Neural Networks». MLSP'2016.
Lukic, Vogt, Dürr & Stadelmann (2017). «Learning Embeddings for Speaker Clustering based on Voice Equality». MLSP'2017.
Stadelmann, Glinski-Haefeli, Gerber & Dürr (2018). «Capturing Suprasegmental Features of a Voice with RNNs for Improved Speaker Clustering». ANNPR'2018.

PROBLEM SOLVED?

Speaker Clustering Using Dominant Sets

Feliks Hibraj*, Ca' Foscari University, Venice, Italy, feliks.hibraj@gmail.com
Sebastiano Vascon*, Ca' Foscari University, Venice, Italy, sebastiano.vascon@unive.it
Thilo Stadelmann, ZHAW Datalab, Winterthur, Switzerland, stdm@zhaw.ch
Marcello Pelillo, Ca' Foscari University, Venice, Italy, pelillo@unive.it

***Abstract*—Speaker clustering is the task of forming speaker-specific groups based on a set of utterances. In this paper, we address this task by using Dominant Sets (DS). DS is a graph-based clustering algorithm with interesting properties that fits well to our problem and has never been applied before to speaker clustering. We report on a comprehensive set of experiments on the TIMIT dataset against standard clustering techniques and specific speaker clustering methods. Moreover, we compare performances under different features by using ones learned via deep neural network directly on TIMIT and other ones extracted from a pre-trained VGGVox net. To asses the stability, we perform a sensitivity analysis on the free parameters of our method, showing that performance is stable under parameter changes. The extensive experimentation carried out confirms the validity of the proposed method, reporting state-of-the-art results under three different standard metrics. We also report reference baseline results for speaker clustering on the entire TIMIT dataset for the first time.**

I. Introduction

Speaker clustering (SC) is the task of identifying the unique speakers in a set of audio recordings (each belonging to exactly one speaker) without knowing who and how many speakers are present altogether [1]. Other tasks related to speaker recognition and SC are the following:

- *Speaker verification* (SV): A binary decision task in which the goal is to decide if a recording belongs to a certain person or not.
- *Speaker identification* (SI): A multiclass classification task in which to decide to whom out of n speakers a certain recording belongs.

SC is also referred to as *speaker diarization* when a single (usually long) recording involves multiple speakers and thus needs to be automatically segmented prior to clustering. Since SC is a completely unsupervised problem (the number of speakers and segments per speaker is unknown), it is straightforward to note that it is considered of higher complexity with respect to both SV and SI. The complexity of SC is comparable to the problem of image segmentation in computer vision, in which the number of regions to be found is typically unknown.

The SC problem is of importance in the domain of audio analysis due to many possible applications, for example in lecture/meeting recording summarization [2], as a pre-processing step in automatic speech recognition, or as part of an information retrieval system for audio archives [3]. Furthermore, SC represents a building block for speaker diarization [4].

The SC problem has been widely studied [5], [6]. A typical pipeline is based on three main steps: *i.a)* acoustic feature extraction from audio samples, *i.b)* voice feature aggregation from the lower-level acoustic features by means of a speaker modeling stage, and *ii)* a clustering technique on top of this feature-based representation.

The voice features after phase *i)* have been traditionally created based on Mel Frequency Cepstral Coefficient (MFCC) acoustic features modeled by a Gaussian Mixture Model (GMM) [7], or i-vectors [8], [9]. More recently, with the rise of deep learning, the community is moving towards learned features instead of hand-crafted ones, as surveyed by Richardson et al. [10]. Recent examples of deep-feature representations for SI, SV, and SC problems come for example from Lukic et al. [11], after Convolutional neural networks (CNN) have been introduced in the speech processing field by LeCun et al. already in the nineties [12]. McLaren et al. used a CNN for speaker recognition in order to improve robustness to noisy speech [13]. Chen et al. used a novel deep neural architecture to learn speaker specific characteristics directly from MFCC features [14]. Yella et al. exploited the capabilities of an artificial neural network of 3 layers to extract features directly from a hidden layer, which are used for speaker clustering [15].

However advanced phase *i)* has become during the last years, the clustering phase *ii)* still relies on traditional methodologies. For example, Khoury et al. demonstrated good results for speaker clustering using a hierarchical clustering algorithm [16], while Kenny et al. report hierarchical clustering to be unsuitable for the speaker clustering stage in a speaker diarization system [17]. In [18] they performed clustering with K-means on dimensionality-reduced i-vectors which showed to work better than spectral clustering as noted in [4].

In this paper, we therefore improve the results of the speaker clustering task by first using state-of-art learned features and then, a different and more robust clustering algorithm, dominant sets (DS) [19]. The motivation driving the choice of dominant sets is the following: *a)* no need for an a-priori number of clusters; *b)* having a notion of compactness to be able to automatically detect clusters composed of noise; *c)* for each cluster the centrality of each element is quantified (centroids emerge naturally in this context); and *d)* extensive

* = Equal contribution

Learning Neural Models for End-to-End Clustering

Benjamin Bruno Meier[1,2], Ismail Elezi[1,3], Mohammadreza Amirian[1,4], Oliver Dürr[1,5], and Thilo Stadelmann[1]

[1] ZHAW Datalab & School of Engineering, Winterthur, Switzerland
[2] ARGUS DATA INSIGHTS Schweiz AG, Zurich, Switzerland
[3] Ca' Foscari University of Venice, Venice, Italy
[4] Institute of Neural Information Processing, Ulm University, Germany
[5] Institute for Optical Systems, HTWG Konstanz, Germany

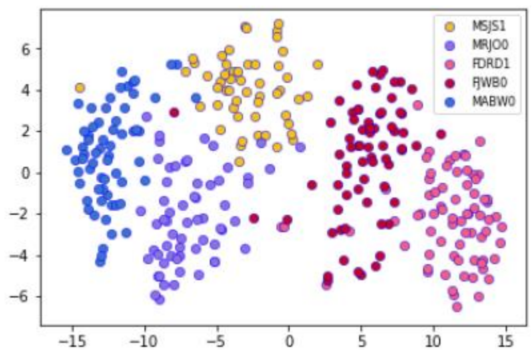
Abstract. We propose a novel end-to-end neural network architecture that, once trained, directly outputs a probabilistic clustering of a batch of input examples in one pass. It estimates a distribution over the number of clusters k, and for each $1 \leq k \leq k_{\max}$, a distribution over the individual cluster assignment for each data point. The network is trained in advance in a supervised fashion on separate data to learn grouping by any perceptual similarity criterion based on pairwise labels (same/different group). It can then be applied to different data containing different groups. We demonstrate promising performance on high-dimensional data like images (COIL-100) and speech (TIMIT). We call this "learning to cluster" and show its conceptual difference to deep metric learning, semi-supervise clustering and other related approaches while having the advantage of performing learnable clustering fully end-to-end.

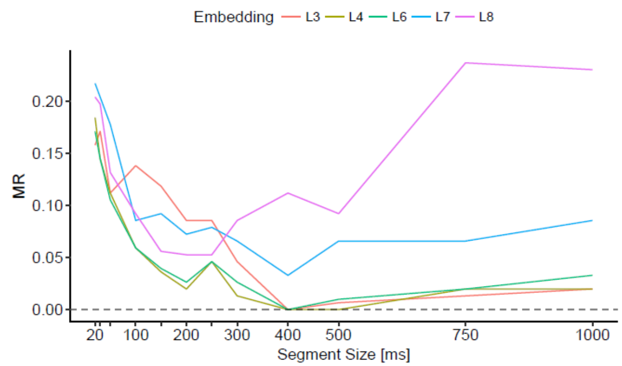
Keywords: perceptual grouping · learning to cluster · speech & image clustering

1 Introduction

Consider the illustrative task of grouping images of cats and dogs by *perceived* similarity: depending on the intention of the user behind the task, the similarity could be defined by animal type (foreground object), environmental nativeness (background landscape, cp. Fig. 1) etc. This is characteristic of clustering perceptual, high-dimensional data like images [15] or sound [24]: a user typically has some similarity criterion in mind when thinking about naturally arising groups (e.g., pictures by holiday destination, or persons appearing; songs by mood, or use of solo instrument). As defining such a similarity for every case is difficult, it is desirable to learn it. At the same time, the learned model will in many cases not be a classifier—the task will not be solved by classification—since the number and specific type of groups present at application time are not known in advance (e.g., speakers in TV recordings; persons in front of a surveillance camera; object types in the picture gallery of a large web shop).

Results of best speaker recognition model





FULL TIMIT	CNN-T Features MR ↓	CNN-T Features ARI ↑	CNN-T Features ACP ↑	CNN-V Features MR ↓	CNN-V Features ARI ↑	CNN-V Features ACP ↑
HC ◇	0.0770	0.8341	0.9283	0.0571	0.8809	0.9484
SP ◇	0.2294	0.0432	0.8355	0.0675	0.5721	0.9488
KM ◇	0.1071	0.7752	0.9071	0.1286	0.6982	0.8730
HC k	0.0762	0.8343	0.9280	0.0706	0.8502	0.9295
SP k	0.2341	0.0421	0.8332	0.0635	0.4386	0.9427
KM k	0.1079	0.7682	0.9007	0.1429	0.6646	0.8485
HC #	0.9921	0.0050	0.0079	0.9984	0.0000	0.0016
SP #	0.9921	0.0003	0.0075	0.9984	0.0000	0.0016
KM #	0.9921	0.0052	0.0076	0.9984	0.0000	0.0016
AP	0.0753	0.8330	0.9030	0.1396	0.7127	0.8222
HDBS	0.1825	0.6214	0.7825	0.3000	0.4112	0.6527
SCDS	**0.0048**	**0.9897**	**0.9947**	**0.0349**	**0.9167**	**0.9578**
SCDS⋆	**0.0048**	**0.9897**	**0.9947**	**0.0349**	**0.9167**	**0.9578**
SCDSbest	**0.0032**	**0.9929**	**0.9966**	**0.0024**	**0.9944**	**0.9974**

«Pure» voice modeling seem largely solved

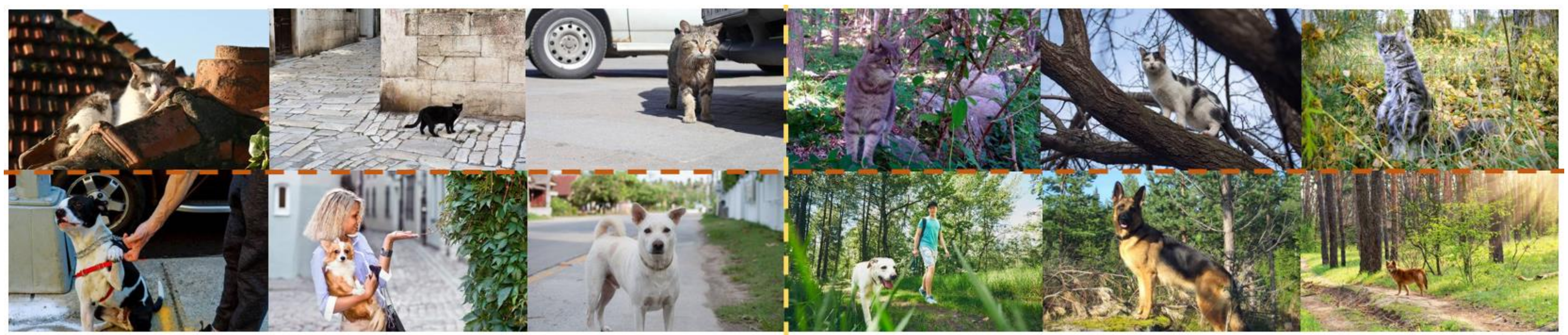
- RNN model robustly exhibits *the predicted* **«sweet spot» for** the used **time information**
- Speaker clustering on clean & reasonably long input works **an order of magnitude better** (*as predicted*)
- Additionally, using a smarter clustering algorithm on top of embeddings makes **clustering on TIMIT as good as identification** (see ICPR'18 paper on dominant sets)

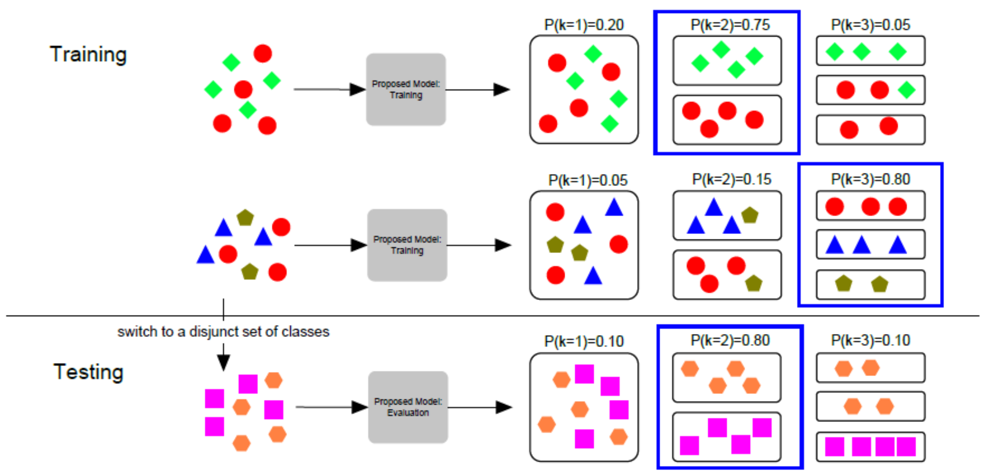
Future work (as seen 2018)

- Make models robust on **real-worldish data** (noise and more speakers/segments)
- Exploit findings for robust reliable **speaker diarization**
- **Learn** embeddings and the clustering algorithm **end to end**

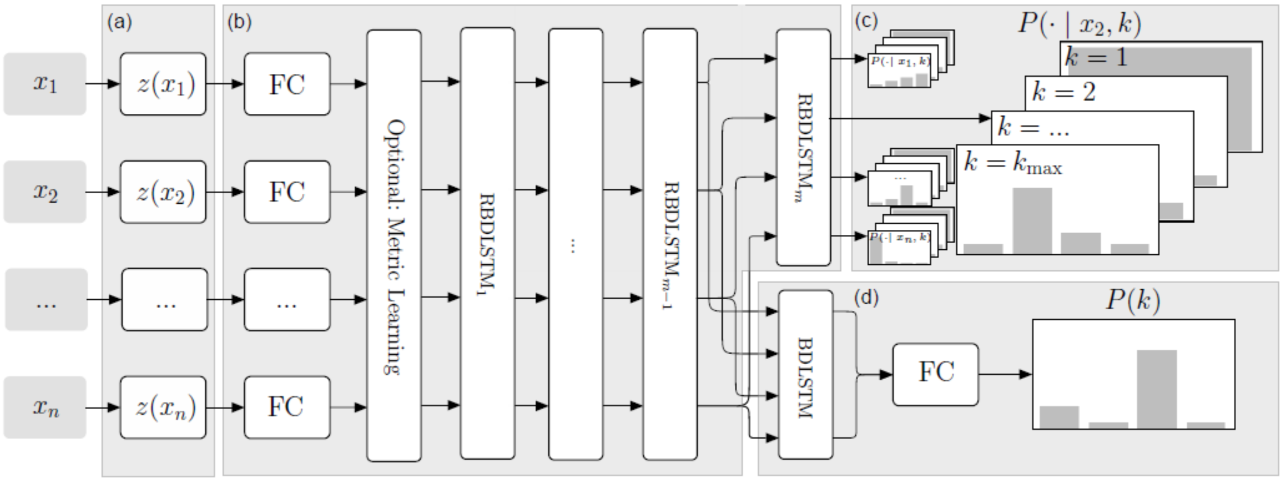
Hibraj, Vascon, Stadelmann & Pelillo (2018). «Speaker Clustering Using Dominant Sets». ICPR'2018.
Meier, Elezi, Amirian, Dürr & Stadelmann (2018). «Learning Neural Models for End-to-End Clustering». ANNPR'2018.

Learning to cluster

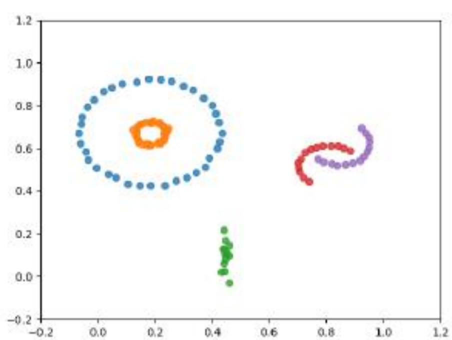




Learning to cluster – architecture & examples

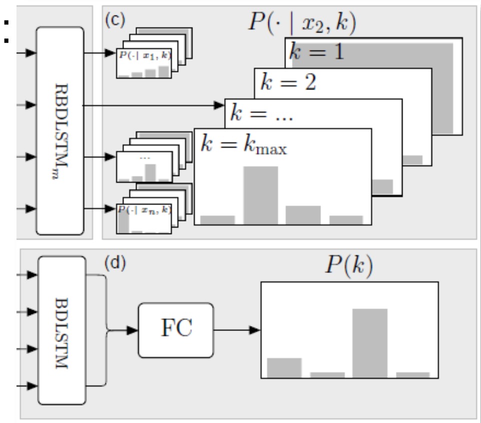


a) **Embedding network:** examples x_i are processed by (data-type specific) embedding network $z(x)$
b) **Clustering network:** embeddings are processed by $m = 14$ bi-directional LSTM layers w/ residual con.
c) **Cluster-assignment network:** for each x_i and cluster count k, output a distribution over the cluster idx
d) **Cluster count estimation network:** output a distribution over the cluster count $1 \leq k \leq k_{max}$



Meier, Elezi, Amirian, Dürr & Stadelmann (2018). «*Learning Neural Models for End-to-End Clustering*». ANNPR'2018.

Learning to cluster – loss



Probability of two instances ***i*, *j*** being **in the same cluster** ℓ (of k clusters):

$$P_{ij}(k) = \sum_{\ell=1}^{k} P(\ell \mid x_i, k) P(\ell \mid x_j, k).$$

Probability of two instances i, j being in the same cluster ℓ **in general**:

$$P_{ij} = \sum_{k=1}^{k_{\max}} P(k) \sum_{\ell=1}^{k} P(\ell \mid x_i, k) P(\ell \mid x_j, k).$$

Cluster assignment loss (with $y_{ij} = 1\ iif$ the two instances are from the same cluster, 0 otherwise):
Weighted binary cross entropy (weights account for imbalance due to more dissimilar pairs)

$$L_{\mathrm{ca}} = \frac{-2}{n(n-1)} \sum_{i<j} \left(\varphi_1 y_{ij} \log(P_{ij}) + \varphi_2 (1 - y_{ij}) \log(1 - P_{ij}) \right)$$

Number of cluster loss:
Categorical cross entropy

$$L_{\mathrm{cc}} = -\log(P(k))$$

Total loss:

$$L_{\mathrm{tot}} = L_{\mathrm{cc}} + \lambda L_{\mathrm{ca}}$$

Zürcher Hochschule
für Angewandte Wissenschaften

SURPRISE, SURPRISE?

Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

InIT Institut für angewandte
Informationstechnologie

Masterthesis (MSE)

Exploiting the Full Information of Varying-Length Utterances for DNN-Based Speaker Verification

Autoren	Daniel Neururer
Hauptbetreuung	Thilo Stadelmann
Datum	31.08.2020

Zürcher Fachhochschule www.engineering.zhaw.ch Studium

Quantifying to which extent DNNs use supra-segmental temporal information

Assumption

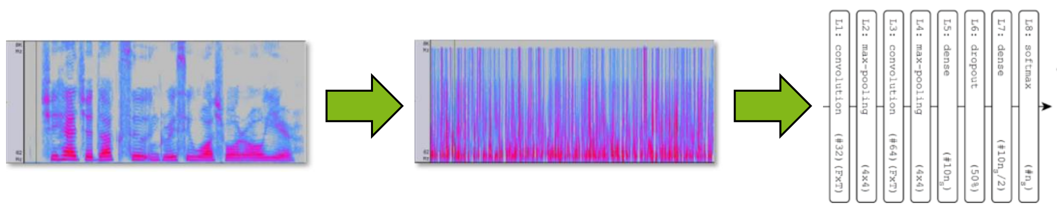
- DNNs are superior voice models ***because*** they model supra-segmental temporal (**SST**) aspects

Evidence

- The **ability is there in principle**: CNNs can use filters along the temporal axis of spectrograms; RNNs have in-built sequence modelling capabilities
- The achieved **results resemble closely the predicted improvements** when modeling temporal aspects: increase in recognition rate, optimal length of temporal context

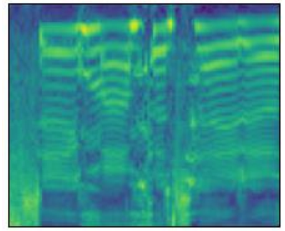
Test

- What happens if we **scramble the time axis** of a spectrogram as a preprocessing to DNN input?

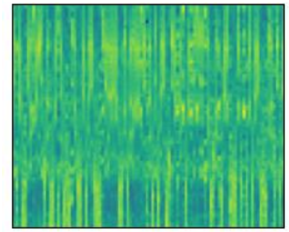


- Rationale: if the sequence of frames is random, the **only usable information are frame-based** acoustic cues (**FBA**) => the **recognition should become worse**, confirming proper exploitation of SSTs

OT



RS



Setup

METHODOLOGY

3 DNNs: **LUVO** (Lukic, Vogt et al., 2016/17), **LSTM** (Stadelmann et al., 2018) and **ResNet34s** (Xie et al., 2019)

Training details

- **CosFace loss** (Wang et al, 2018) instead of PKLD for computational efficiency and larger margins
- **Per epoch** (64x): draw 1s segment from random starting point from each utterance; batch size 100

Evaluation

- **Evaluate speaker clustering** with Misclassification rate (***MR***) and **speaker verification** with ***EER***
- **Utterance representation**: 1s segments w/ 50% overlap → average over resulting embeddings

EXPERIMENTS

TIMIT dataset

- 630 speakers, studio conditions, 10 sentences/speaker
- Training set: 462 speakers (8 sentences train, 2 val)
- Test set: 168 speakers (10 sentences)

Setup

- As **similar** as possible **to prior work** (2009-2018)
- **Train** each DNN with **original** (**OT**) or **randomized** (**RS**) time axis
- **Evaluate** each trained model **with OT** and **RS** segments
- **Clustering**: hierarchical clustering of 2 utterances (8 or 2 concatenated sentences) per speaker (40 speakers)
- **Verification**: for all test speakers & each sentence: selected 2 matched & 2 unmatched random sentences

Stadelmann & Freisleben (2009). «*Unfolding Speaker Clustering Potential: A Biomimetic Approach*». ACMMM'2009.
Lukic, Vogt, Dürr & Stadelmann (2016). «Speaker Identification and Clustering using Convolutional Neural Networks». MLSP'2016.
Lukic, Vogt, Dürr & Stadelmann (2017). «Learning Embeddings for Speaker Clustering based on Voice Equality». MLSP'2017.
Stadelmann, Glinski-Haefeli, Gerber & Dürr (2018). «Capturing Suprasegmental Features of a Voice with RNNs for Improved Speaker Clustering». ANNPR'2018.
Xie, Nagrani, Chung & Zisserman: *"Utterance-level Aggregation for Speaker Recognition in the Wild"*. ICASSP 2019.
Wang, Wang, Zhou, Ji, Gong, Zhou, ... & Liu: *"Cosface: Large margin cosine loss for deep face recognition."* CVPR 2018.

Zürcher Hochschule
für Angewandte Wissenschaften

Results

Speaker clustering on TIMIT

(MR, averaged over 5 runs)

		H50 OT	H50 RF	H50 RS
LUVO	OT	**0.00 σ0.00**	9.75 σ0.94	9.00 σ2.15
	RF	8.50 σ2.42	0.50 σ0.61	1.75 σ0.61
	RS	9.00 σ1.66	1.00 σ0.50	1.25 σ0.00
LSTM	OT	1.25 σ1.12	2.75 σ0.94	2.75 σ0.50
	RF	3.75 σ1.37	**0.00 σ0.00**	2.50 σ1.58
	RS	2.00 σ1.00	1.25 σ0.79	0.25 σ0.50
RESNET34S	OT	**1.00 σ0.94**	8.25 σ4.78	11.50 σ4.29
	RF	2.50 σ1.77	**1.00 σ0.50**	3.00 σ1.27
	RS	2.75 σ0.94	1.25 σ1.12	**1.00 σ0.94**

Speaker verification on TIMIT

(EER, averaged over 5 runs)

		H50 OT	H50 RF	H50 RS
LUVO	OT	6.38 σ0.12	12.02 σ0.51	11.90 σ0.46
	RF	8.55 σ0.49	5.55 σ0.06	6.12 σ0.12
	RS	8.16 σ0.42	**5.33 σ0.18**	5.78 σ0.16
LSTM	OT	**3.53 σ0.07**	4.19 σ0.09	3.90 σ0.12
	RF	3.99 σ0.16	3.78 σ0.10	3.66 σ0.13
	RS	4.00 σ0.07	3.89 σ0.06	3.54 σ0.05
RESNET34S	OT	**4.96 σ0.19**	10.34 σ1.56	9.21 σ1.15
	RF	6.59 σ0.25	6.25 σ0.23	6.37 σ0.35
	RS	5.89 σ0.25	6.11 σ0.31	5.80 σ0.11

- **RF**: fill a segment by picking frames at random from *full utterance* (i.e., more phonetic variability)

➔ **DNNs** seem to **ignore SST information** and still almost exclusively rely on FBA features

Follow-up question

- Can we **force DNNs to use SST** features by „scrambling“ FBA information?

Testing if DNNs can be forced to not rely on frame-based acoustic information alone

1. Make the problem acoustically harder by decreasing the SNR

Speaker verification on VoxCeleb (speech „in the wild“, 5994 speakers, 1+ mio. utterances)

		H50 OT	H50 RF	H50 RS
LUVO	OT	6.38 σ0.12	12.02 σ0.51	11.90 σ0.46
	RF	8.55 σ0.49	5.55 σ0.06	6.12 σ0.12
	RS	8.16 σ0.42	**5.33 σ0.18**	5.78 σ0.16
LSTM	OT	**3.53 σ0.07**	4.19 σ0.09	3.90 σ0.12
	RF	3.99 σ0.16	3.78 σ0.10	3.66 σ0.13
	RS	4.00 σ0.07	3.89 σ0.06	3.54 σ0.05
RESNET34S	OT	**4.96 σ0.19**	10.34 σ1.56	9.21 σ1.15
	RF	6.59 σ0.25	6.25 σ0.23	6.37 σ0.35
	RS	5.89 σ0.25	6.11 σ0.31	5.80 σ0.11

➔

		H50 OT	H50 RF	H50 RS
LUVO	OT	**25.75 σ0.13**	37.23 σ0.74	36.96 σ0.78
	RF	32.70 σ0.34	27.04 σ0.34	27.99 σ0.30
	RS	33.26 σ0.29	27.91 σ0.32	28.50 σ0.28
LSTM	OT	**20.67 σ0.23**	30.67 σ0.36	30.00 σ0.32
	RF	26.20 σ0.18	22.02 σ0.10	23.57 σ0.09
	RS	28.28 σ1.30	26.30 σ0.59	26.58 σ0.84
RESNET34S	OT	**12.49 σ0.15**	34.11 σ0.54	32.19 σ0.39
	RF	22.05 σ0.43	19.08 σ0.26	20.02 σ0.16
	RS	20.74 σ0.46	21.02 σ0.34	20.36 σ0.23

(EER, averaged over 5 runs)

➔ Being able to exploit **SST** information **helps in the presence of more noise**

Testing if DNNs can be forced to not rely on frame-based acoustic information alone

2. Remove discriminative power of FBAs by equalizing timbre of speakers

Speaker verification on TIMIT-NV (noise-vocoded w/ original amplitude contours in 4 bands)

		H50 OT	H50 RF	H50 RS
LUVO	OT	6.38 σ0.12	12.02 σ0.51	11.90 σ0.46
	RF	8.55 σ0.49	5.55 σ0.06	6.12 σ0.12
	RS	8.16 σ0.42	**5.33 σ0.18**	5.78 σ0.16
LSTM	OT	**3.53 σ0.07**	4.19 σ0.09	3.90 σ0.12
	RF	3.99 σ0.16	3.78 σ0.10	3.66 σ0.13
	RS	4.00 σ0.07	3.89 σ0.06	3.54 σ0.05
RESNET34S	OT	**4.96 σ0.19**	10.34 σ1.56	9.21 σ1.15
	RF	6.59 σ0.25	6.25 σ0.23	6.37 σ0.35
	RS	5.89 σ0.25	6.11 σ0.31	5.80 σ0.11

		H50 OT	H50 RF	H50 RS
LUVO	OT	32.56 σ0.62	35.32 σ0.46	35.41 σ0.55
	RF	35.16 σ0.52	**30.39 σ0.30**	30.91 σ0.47
	RS	35.25 σ0.69	30.63 σ0.38	31.23 σ0.27
LSTM	OT	**19.34 σ0.16**	27.20 σ0.42	26.12 σ0.44
	RF	22.95 σ0.24	21.48 σ0.40	21.15 σ0.25
	RS	22.82 σ0.40	21.89 σ0.25	21.04 σ0.12
RESNET34S	OT	**21.12 σ0.43**	37.83 σ1.17	36.57 σ1.45
	RF	27.03 σ0.63	23.38 σ0.41	24.02 σ0.25
	RS	27.25 σ1.37	23.57 σ0.46	23.32 σ0.58

(EER, averaged over 5 runs)

➔ Being able to exploit **SST** information **helps with less speaker-discriminating FBAs**
➔ Disclaimer: not evident for speaker clustering using MR

Testing if DNNs can be forced to not rely on frame-based acoustic information alone

2. Remove discriminative power of FBAs by equalizing timbre of speakers

Speaker verification on TIMIT-Syn (re-synthesized w/ original, normalized pitch tracks and phone-level timing information from annotations [Slowsoft synthesizer, similar for MBROLA])

		H50 OT	H50 RF	H50 RS
LUVO	OT	6.38 σ0.12	12.02 σ0.51	11.90 σ0.46
	RF	8.55 σ0.49	5.55 σ0.06	6.12 σ0.12
	RS	8.16 σ0.42	**5.33 σ0.18**	5.78 σ0.16
LSTM	OT	**3.53 σ0.07**	4.19 σ0.09	3.90 σ0.12
	RF	3.99 σ0.16	3.78 σ0.10	3.66 σ0.13
	RS	4.00 σ0.07	3.89 σ0.06	3.54 σ0.05
RESNET34S	OT	**4.96 σ0.19**	10.34 σ1.56	9.21 σ1.15
	RF	6.59 σ0.25	6.25 σ0.23	6.37 σ0.35
	RS	5.89 σ0.25	6.11 σ0.31	5.80 σ0.11

		H50 OT	H50 RF	H50 RS
LUVO	OT	46.24 σ0.18	48.94 σ0.15	48.97 σ0.23
	RF	47.26 σ0.15	45.98 σ0.34	46.16 σ0.27
	RS	47.14 σ0.22	45.88 σ0.12	**45.66 σ0.12**
LSTM	OT	**40.39 σ0.07**	44.29 σ0.65	42.43 σ1.40
	RF	43.63 σ0.35	41.93 σ0.26	41.64 σ0.25
	RS	43.62 σ0.21	42.55 σ0.34	41.53 σ0.23
RESNET34S	OT	**40.33 σ1.32**	47.28 σ2.06	46.60 σ2.02
	RF	43.44 σ0.86	42.97 σ0.51	42.65 σ0.59
	RS	42.48 σ0.45	43.07 σ0.72	41.59 σ0.36

(EER, averaged over 5 runs)

➔ Being able to exploit **SST** information **helps without any speaker-discriminating FBAs**
➔ Disclaimer: less evident for speaker clustering using MR

Zürcher Hochschule
für Angewandte Wissenschaften

Discussion

- **DNNs** are **lazy in picking up higher-level features** like SSTs
  → there is still the **potential for improvement, possibly still one order of magnitude**
- Recent results are still preliminary and open many areas for future work
  → **who helps** to uncover their depth?
- Happy to collaborate interdisciplinary & internationally

About me

- Prof. Dr. Thilo Stadelmann
- Director Centre for AI, head Computer Vision, Perception & Cognition Group
- Email: stdm@zhaw.ch
- Phone: +41 58 934 72 08
- Social media: @thilo_on_data, in/thilo-stadelmann/

Further contacts:

- info.cai@zhaw.ch, datalab@zhaw.ch, info.office@data-innovation.org, office-switzerland@claire-ai.org

Datalab
zhaw.ch/datalab

data innovation alliance

SwissNLP
Swiss Association for
Natural Language Processing

CLAIRE