Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie

Multimedia Analysis
Speaker Recognition

Material based on the lecture *„Video Retrieval"* by
Thilo Stadelmann, Ralph Ewerth, Bernd Freisleben
AG Verteilte Systeme, Fachbereich Mathematik & Informatik

Philipps Universität Marburg

Content

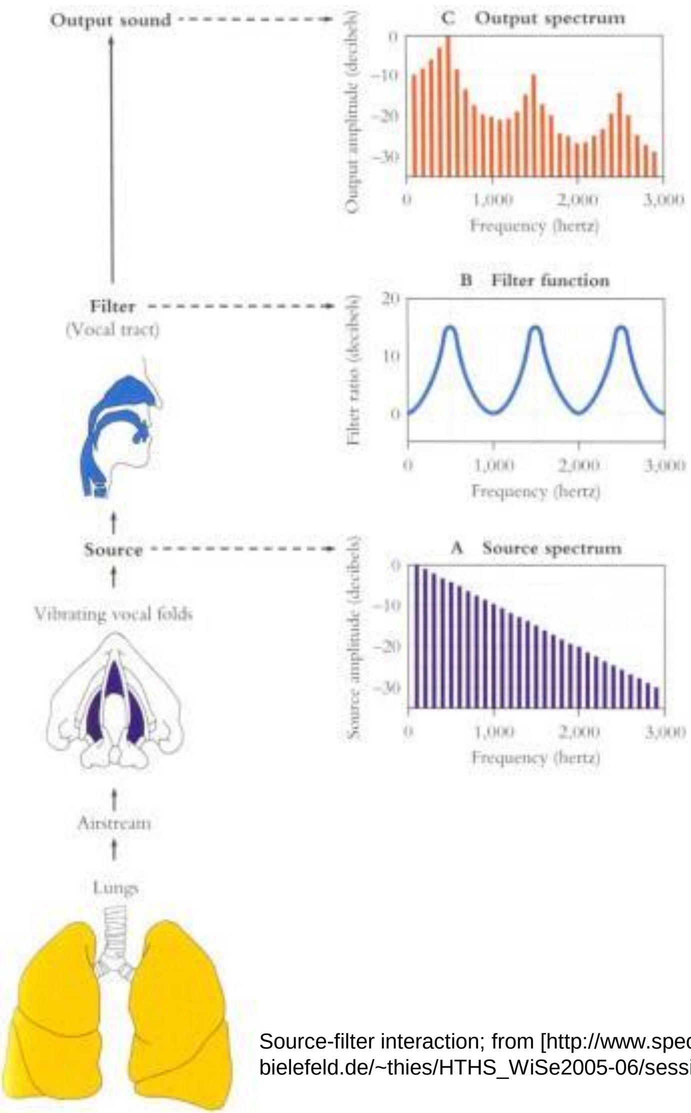
1. Introduction
   - What is speaker recognition
   - Speech production
   - Hints from other disciplines

2. The GMM approach to speaker modeling
   - The general idea
   - GMM in practice
   - An audio-visual outlook

Task and settings for speaker recognition

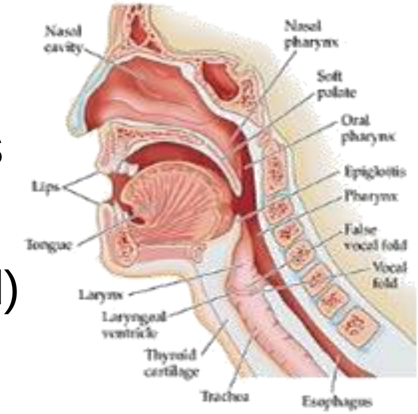
- Speaker recognition: **tell identity of** an utterances' **speaker**
- Typical: **score feature-sequence against** a speaker **model**
- Possible settings:
  - **verification**: verify that a given utterance fits a claimed identity (model) or not
  - **identification**: find the actual speaker among a list of prebuild models (or declare as unknown: open set identification)
  - **diarization**, tracking, clustering: segment an audio-stream by voice identity (who spoke when, no prior knowledge of any kind)

The source filter model



Source-filter interaction; from [http://www.spectrum.uni-bielefeld.de/~thies/HTHS_WiSe2005-06/session_05.html]

- **Source**: Air flows from the lungs through the vocal chords
  - noise-like (unvoiced) or
  - periodic (overtone-rich, voiced) excitation sound



The vocal tract; from [DUKE Magazine, Vol. 94, No. 3, 05/06 2008]

- **Filter**: vocal tract shapes the emitted spectrum

⇒ **Size** of the **glottis determines** fundamental frequency (**F_0**) **range**

⇒ **Shape** of the **vocal tract** and nasal cavity **determines** formant frequencies (**$F_{1\text{-}5}$**) and "**sound**"

Features from source-filter decomposition

- Represent **source** characteristics **via pitch & noise**
  - 1 `double` per frame

- Represent **filter** characteristics with filter coefficients $\boldsymbol{a_k}$ from **LPC analysis** (8-10 `double` per frame):
  - $s[n] = \sum_{k=1}^{p} a[k] \cdot s[n-k] + e[n]$
  - Btw.: this is the way **it is done in mobile phones**…

- LPC coefficients are also applied as (or further processed to be) speaker specific features, but typically, MFCCs are used

Speech properties

- Slowly time-varying
  - ⇒ **stationary over** sufficiently **short period** (5-100ms, **phoneme**)

- Speech range: 100 - 6800Hz (telephone: 300 - 3400Hz)
  - ⇒ 8kHz samplerate sufficient, 16kHz optimal

- Speech frames convey **multiple information**:
  1. Linguistic (phonemes, syllables, words, sentences, phrases, …)
  2. Identity
  3. Gender
  4. Dialect
  5. …

  ⇒ **fractal structure**

The human auditory system

- **High dynamic range** (120dB, $q_{dB} = 10 \cdot \log_{10}\left(\frac{q}{q_{ref}}\right)$ for some quantity *q*)

  $\Rightarrow$ work in the log domain (increase in 3dB => loudness doubled)

- Performs short-time spectral analysis (similar to wavelet-/Fourier-transform) with log-frequency resolution

  $\Rightarrow$ **Mel filterbank**

- **Masking** effects

  $\Rightarrow$ that's what makes mp3 successful in compressing audio

- Channel decomposition via "**auditory object recognition**"

  =>that's what a machine can not (yet)

$\Rightarrow$ lots of further interesting material, but no direct and simple applicability to ASR at the moment

- More on the auditory system: Moore, "An Introduction to the Psychology of Hearing", 2004

Forensic speaker identification (1)

- **Manual** or semi-automatic voice comparison done **by phoneticians**
  - "when it really matters“

- Useful insights:
  - **compare** only **matching** (i.e. hand-selected) **units** (i.e. phonemes; ca. 10 per second)
    - $\geq$ **30 realisations** per unit **needed** to get relatively sure
  - useful **features**: formants, fundametal frequency, energy, speaking rate, formant coupling, articulation, dialect, syllable grouping, breath pattern
  - **long term** ($\geq$ 60s) **F_0 statistics** (mean and range) are relevant (generally, the longer the better)

Forensic speaker identification (2)

- formants: F_1-F_3 resemble vowel quality, F_3 indicates vocal tract length, **F_4-F_5** are more **speaker specific** but **difficult to extract** automatically
- **vocal chord activity** (pitch, phonation) and nasals are relevant
- number and distribution of **speech pauses** is relevant
- **cepstral features don’t refer directly** to what is known about how speakers actually differ
- great use in **linguistic** rather than acoustic **parameters**
- understanding the language is relevant (=> **context information**)
- auditory analysis of **voice quality** is relevant

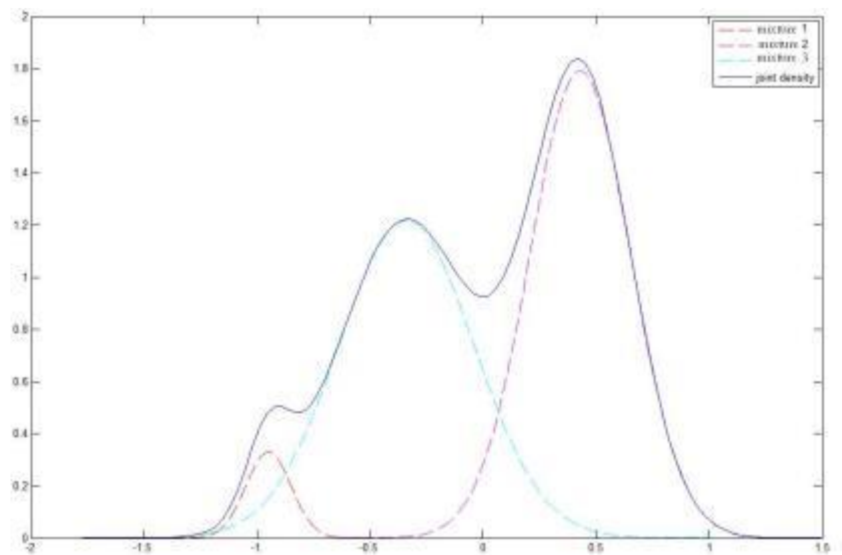
- More on forensic phonetics:
  - http://www.uni-marburg.de/fb09/igs/institut/abteil/phonetik
  - Rose, "Forensic Speaker Identification", 2002
  - Laver, "The Phonetic Description of Voice Quality", 1980

A multimodal, multivariate Gaussian model

- Reynolds, Rose, *"Robust Text-Independant Speaker Identification Using Gaussian Mixture Speaker Models"*, 1995

- Idea: **Take** the estimated probability density function (**pdf**) $p(\vec{x} \mid \lambda)$ **of** a **speaker's** (D-dim.) training **vectors** $\vec{x}$ **as** a **model** of his voice

$\Rightarrow$ **Model** the **pdf via** a **weighted** sum (linear **combination**) **of** *M* *D*-dimensional **gaussians** g_i



GMM with 3 mixtures in 1 dimension; from [Y. Wang, Diplomarbeit, 2009]

Rationale

- Hybrid solution between non-parametric clusters (VQ) and compact smoothing (Gaussian):
  - Smooth aproximation of **arbitrary densities**
  - Implicit clustering into **broad phonetic classes**

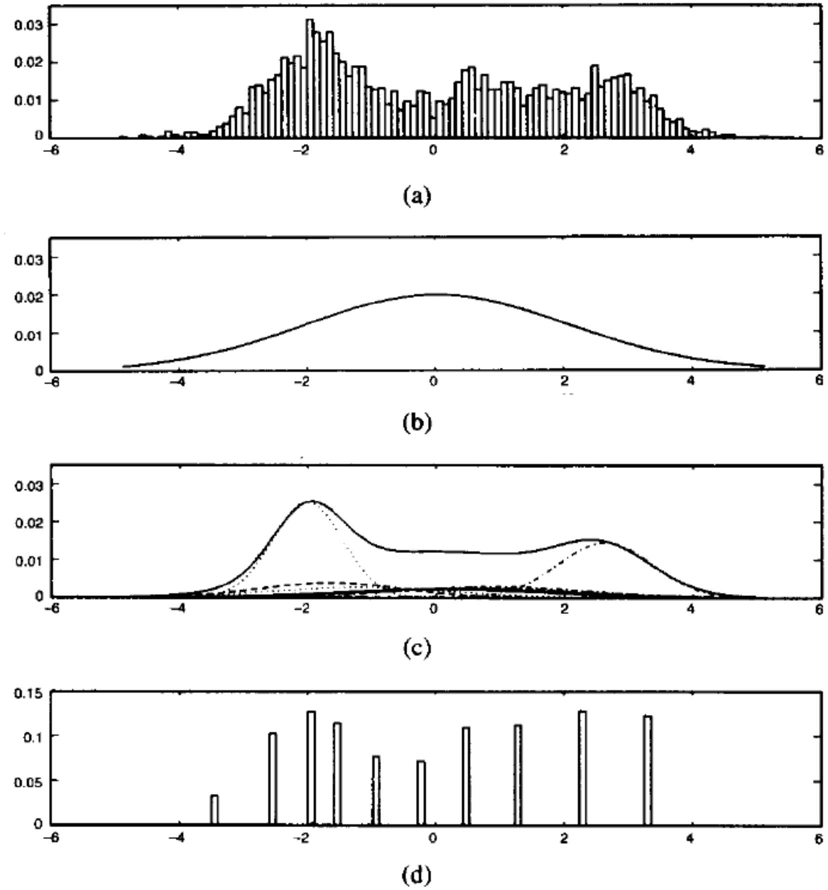


Fig. 3. Comparison of distribution modeling: (a) Histogram of a single cepstral coefficient from a 25 second utterance by a male speaker; (b) maximum likelihood unimodal Gaussian model; (c) GMM and its 10 underlying component densities; (d) histogram of the data assigned to the VQ centroid locations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

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

Mathematics

- Reminder: λ model (GMM), w weight, μ mean, Σ covariance, p pdf, $\vec{x}$ feature vector, g_i i^th^ (out of M) Gaussian mixture

$$\lambda = \{w_i, \mu_i, \Sigma_i\}, i = 1..M$$

$$g_i(\vec{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} \cdot \left|\Sigma_i\right|^{\frac{1}{2}}} \cdot e^{-\frac{1}{2}\cdot(\vec{x}-\vec{\mu}_i)'\Sigma_i^{-1}(\vec{x}-\vec{\mu}_i)}$$

$$\sum_{i=1}^{M} w_i = 1$$

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} w_i \cdot g_i(\vec{x})$$

Model training and testing

- A GMM is trained via the **Expectation Maximization (EM)** Algorithm
  - **Maximum likelihood** (ML) training, **initialized by k-Means**
  - Maximum *a posteriori* (MAP) adaptation (i.e. uses *a priori* knowledge)

- Finding the speaker *s* of a new utterance (represented by its feature vector sequence $X = \{\vec{x}_1..\vec{x}_T\}$) from a given a set of speakers (represented by their models $\{\lambda_1..\lambda_S\}$):
  - $$s = \arg\max_s p(X \mid \lambda_s) = \arg\max_s \Pi_{t=1}^{T} p(\vec{x}_t \mid \lambda_s)$$

- More on EM and current GMM trends:
  - Mitchel, „Machine Learning“, chapter 6.2 „The EM Algorithm“, 1997
  - Reynolds, Quatieri, Dunn, „Speaker Verification Using Adapted Gaussian Mixture Models“, 2000

Best practices

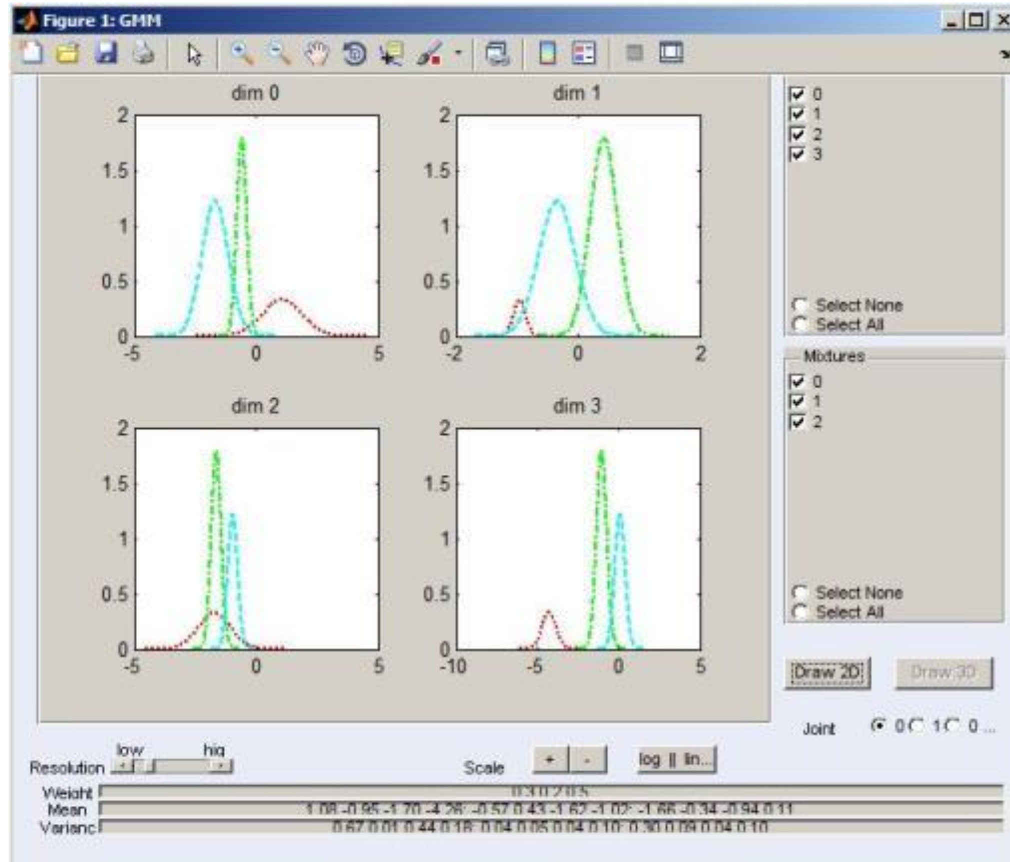
- Use **diagonal covariances**
  $\Rightarrow$ Simpler/faster training, same/better results (with more mixtures)
- Use a **variance limit** and beware of **curse of dimensionality**
  $\Rightarrow$ Prohibit artifacts through underestimation of components
- Use **16-32 mixtures** and a **minimum of 30s of speech** (ML)
- Adapt only means from 512-1024 mixtures per gender (MAP)
  - Score only with top-scoring mixtures
- **Find optimal number of Mixtures** for data via brute force and BIC
- **Compare** models via
  - Generalized Likelihood Ratio (GLR) or (**score-wise**)
  - Earth Mover‘s Distance (EMD) or Beigi/Maes/Sorensen Distance (**parameter-wise**)

A tool for visual experimentation and debugging

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering
InIT Institut für angewandte Informationstechnologie

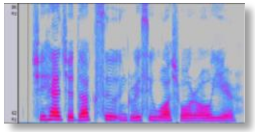
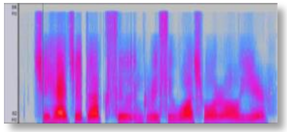
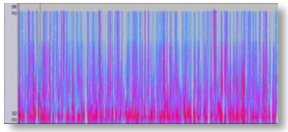
- **Matlab tool** for GMM visualization

- Developed by Y. Wang, diploma-thesis 2009 at Marburg University

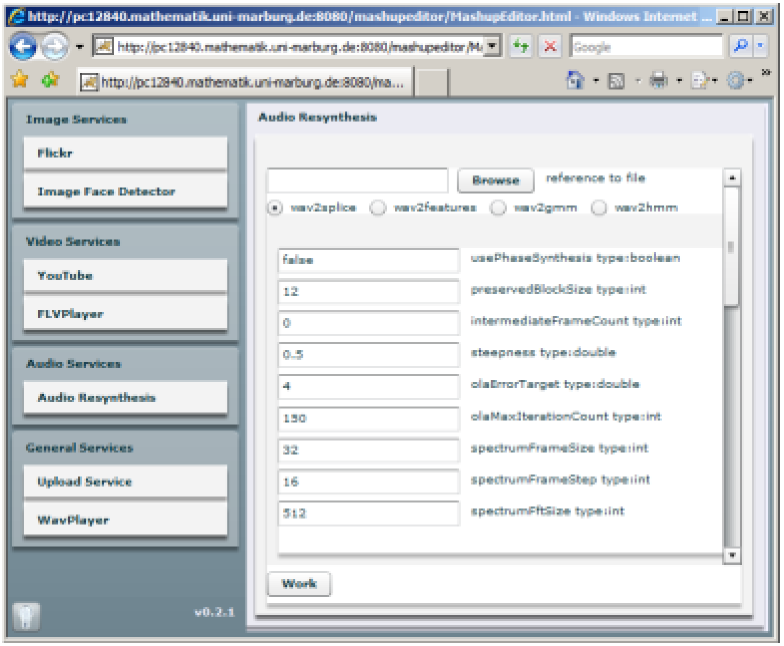


- Available at `http://www.mathematik.uni-marburg.de/~stadelmann/`

What GMMs might fail to capture…

- **Re-synthesizing** what a **speech** processing result conveys:
  - Tool at http://mage.uni-marburg.de/audio/audio.html
  - Original/spliced signal: `examples/SA1_spliced.wav`
  - Resynthesized MFCCs: `examples/SA1_features.wav`
  - Resynthesized MFCCs from GMM: `examples/SA1_gmm.wav`



- Implications?
  - **Model temporal context!**

- More on temporal context:
  - Friedland, Vinyals, Huang, Müller, „Prosodic and other Long-Term Features for Speaker Diarization“, 2009
  - Stadelmann, Freisleben, „Unfolding Speaker Clustering Potential – A Biomimetic Approach“, 2009

Review

- **Speaker recognition** comes in the flavours of **verification**, **identification** or **diarization**
- Lots of useful **insight** for automated systems comes **from other disciplines**: psychoacoustics, signal processing and (forensic) phonetics

- The classic (still **quasi-standard**) approach is **MFCC** features and **GMM** models
- There’s lots of engineering to find optimal parameters, but **best practices** help a lot