Recent Advances in Speaker Diarization

Hagai Aronowitz

IBM Research – Haifa

Interspeech 2014 Tutorial

Outline

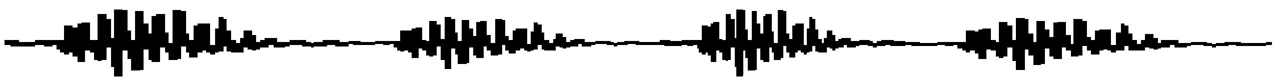
1. Introduction
2. Speech processing
3. Voice activity detection
4. Classic diarization methods
5. Speaker recognition
6. Advanced diarization methods
   - Geometry
   - Clustering
   - Techniques

Outline

1. **Introduction**
2. Speech processing
3. Voice activity detection
4. Classic diarization methods
5. Speaker recognition
6. Advanced diarization methods
   - Geometry
   - Clustering
   - Techniques

Definition

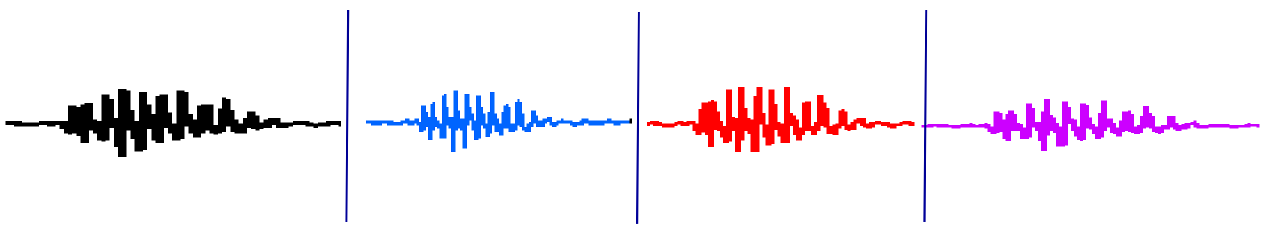
"Who spoke when"



- Speaker identities are unknown

Definition

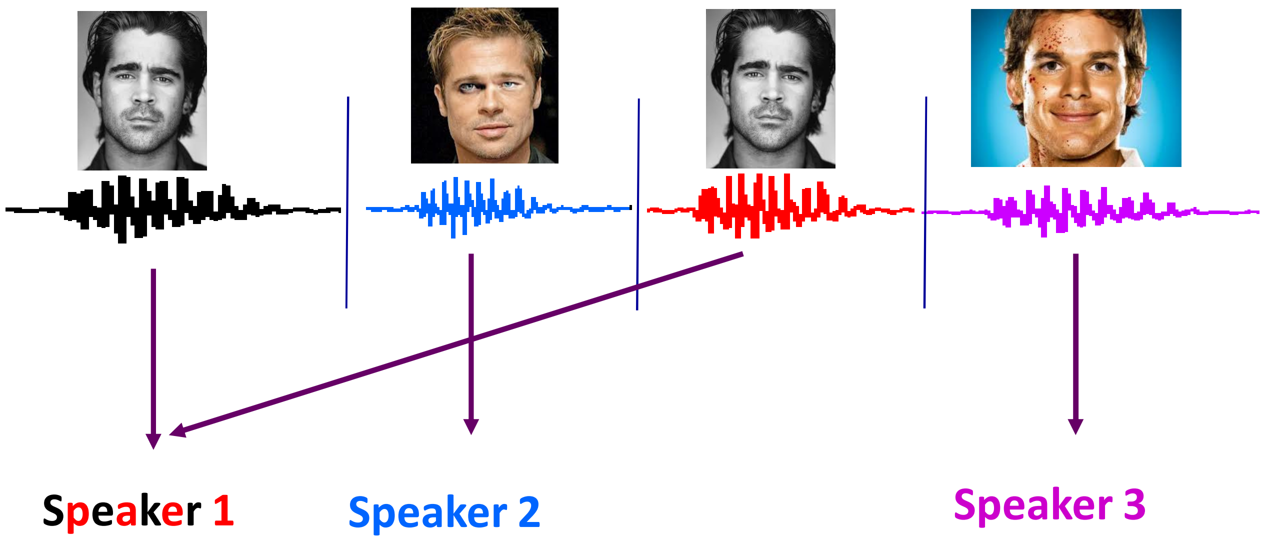
"Who spoke when"



- Speaker identities are unknown
- Speaker change detection (speaker segmentation)

Definition

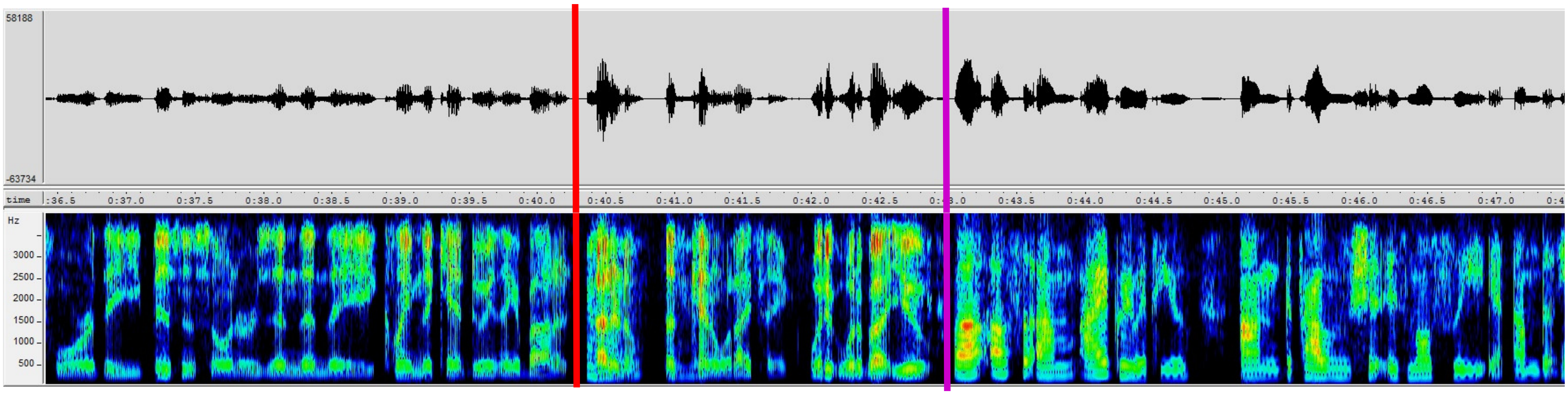
"Who spoke when"



- Speaker identities are unknown
- Speaker change detection (speaker segmentation)
- Speaker clustering

Example

What is the correct boundary – **red** or **purple**?



Meetings

- Number of participants is unknown
- Close talk / far-field microphones
  - Noisy data
  - Reverberation
- Spontaneous speech

Broadcast News (BN)

- Number of participants is unknown
- News: High quality, planned speech
- Conversations: Spontaneous
- Mixed sources: close-talk mic, telephone
- Background music, commercials

Telephone Conversations

- Two speakers
  - Realistic data may contain 3 speakers or more
- Spontaneous speech

Applications

- Enable speaker adaption in speech recognition systems
- Enable speaker recognition in multi-speaker data
  - Summed phone calls
- Pre-processing for speech analytics
  - Agent-customer segmentation in call-centers
- Pre-processing for speech transcription and speech translation
  - Accurate speaker turns are key to good performance
- Displaying speaker turns and speaker labels
  - Transcription services (meetings)

Basic Functionalities

- Voice activity detection
  - Detection of speech vs. non-speech (silence, noise, music)
  - Speech may be mixed with noise or music
- Speaker change detection
  - Find timestamps for which speaker before timestamp is different than speaker after timestamp
- Speaker clustering
  - Associate speaker turns according to speaker identity

Performance Measures

Measure	Description
False Acceptance (FA)	Probability of classifying non-speech as speech
Miss Detection (MD)	Probability of classifying speech as non-speech
Speaker Error Rate (SER)	Probability of speech to be assigned to the wrong speaker
Diarization Error Rate (DER)	FA + MD + SER

Forgiveness collar

- Regions around speaker boundaries are not scored
- In NIST Evaluations forgiveness collar is 0.25 sec

Outline

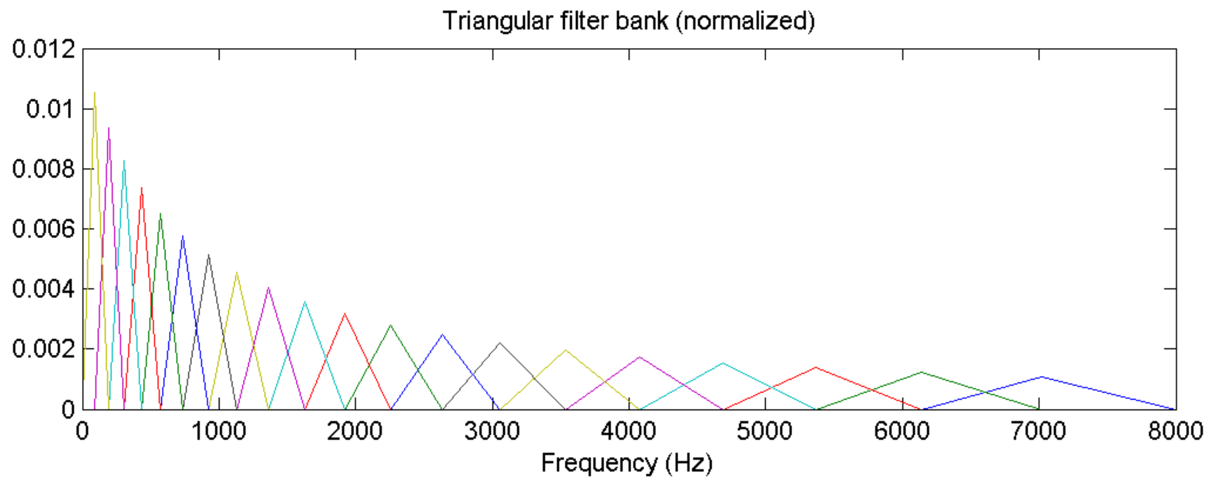
1. Introduction
2. **Speech processing**
3. Voice activity detection
4. Classic diarization methods
5. Speaker recognition
6. Advanced diarization methods
   - Geometry
   - Clustering
   - Techniques

Feature Extraction

- Speech is divided into frames (100/s)
- Signal in each frame is assumed to be stationary
- Frames are parameterized by vectors representing their spectral characteristics using MFCCs (Mel Frequency Cepstral Coefficients)
- Delta MFCCs (and delta-delta) are usually concatenated to the MFCCs

MFCCs

- Take the Fourier transform of a frame
- Map power spectrum into the Mel-scale, using triangular overlapping windows



- Take logs of the powers
- Take the discrete cosine transform of the log powers

Modeling Speech Segments

- Given a sequence of MFCC features $X = x_1, \ldots, x_t$
- Model sequence by a stochastic model
  - Gaussian Mixture Model (GMM)
  - Hidden Markov Model (HMM)
- Models can be used to
  - Learn a class of segments and classify a segment ***X***

$$p(C|X) = \frac{p(X|C)\,p(C)}{p(X)}$$

  - Map segments into a vector space

$$X \rightarrow \mathrm{pdf}(X)$$

Gaussian Mixture Model (GMM)

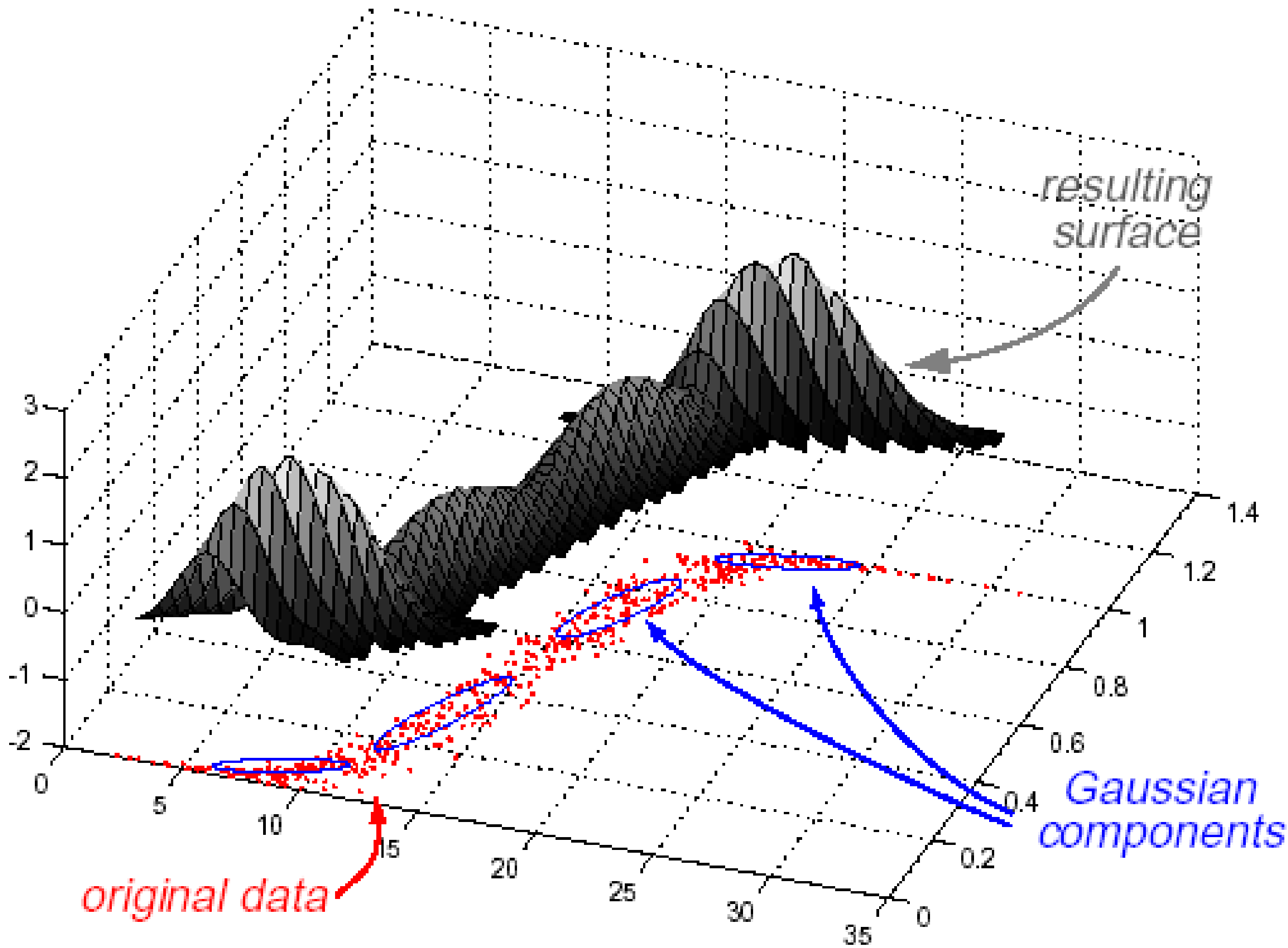
- A GMM is a convex combination of normal PDFs

$$p(x) = \sum_g w_g \, p(x \mid N(\mu_g, \Sigma_g))$$

$$\sum_g w_g = 1$$

- Most interesting PDFs may be reasonably approximated by GMMs
- Interpretation
    1. Gaussian index *g* is drawn from a generalized Bernoulli distribution $\{w_g\}$
    2. Observation *x* is drawn from normal distribution $N(\mu_g, \Sigma_g)$

GMMs in 2D

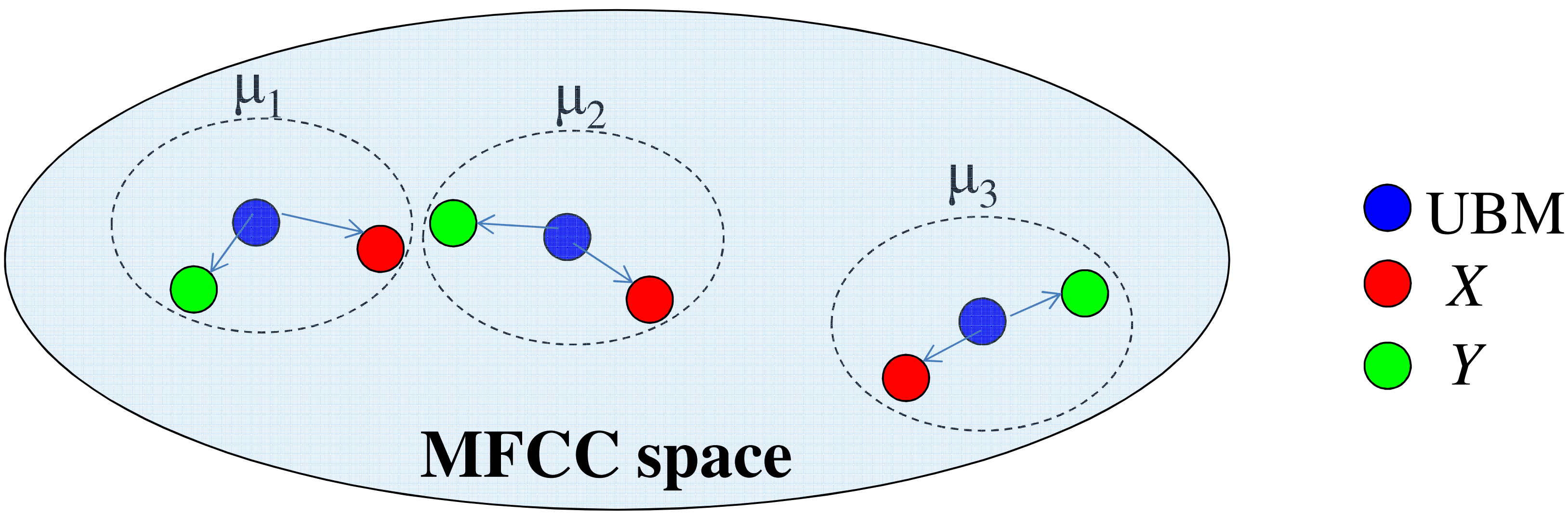

resulting surface
Gaussian components
original data
3
2
1
0
-1
-2
0
5
10
15
20
25
30
35
0
0.2
0.4
0.6
0.8
1
1.2
1.4

GMMs for Speech Modeling

- Frame independence assumption
  - Segment is a bag of frames
- Simplistic interpretation
  - Every phoneme is modeled by a single normal distribution
  - Mixture weights represent prior phoneme probabilities
- Realistic interpretation
  - Every phoneme is modeled by a low-order GMM

Universal Background Model (UBM)

1. Train a Universal GMM on data collected from many segments
2. Adapt a GMM for segment X by applying MAP (Maximum A-Posteriori) adaptation to the UBM
   → all GMMs are aligned



GMMs Supervectors

- Given segments X and Y, what is $p(Y|X)$?
- Classic approach
  - Estimate GMM G_X from X
  - Compute the log-likelihood of Y

$$\log p\left(Y\middle|X\right) = \log p\left(Y\middle|G_X\right) = \sum_t \log\left(\sum_g w_g\, p(y_t \mid N(\mu_g, \Sigma_g))\right)$$

- Disadvantages
  - Inefficient for scoring a session against multiple models
  - Inflexible to modifying modeling assumptions

GMMs Supervectors (2)

- Supervector approach
  - Estimate both G_X and G_Y
  - Average log-likelihood → Negative cross entropy

$$\frac{1}{|Y|}\log p\left(Y \middle| G_X\right) \xrightarrow[T\to\infty]{} \int_f p\left(f \middle| G_Y\right)\log p\left(f \middle| G_X\right)df = -H\left(G_Y, G_X\right)$$

  - Matching-based approximation

$$H\left(G_X, G_Y\right) \cong \frac{1}{2}\sum_{g=1} w_g\left(\mu_g^X - \mu_g^Y\right)^T \Sigma_g^{-1}\left(\mu_g^X - \mu_g^Y\right) + \text{const}$$

  - Supervector transform T

$$T : X \to x \quad ; \quad x_{g*D+d} = \sqrt{w_g}\,\frac{\mu_{g,d}^{G_X}}{\sigma_{g,d}}$$

$$\Rightarrow \frac{1}{|Y|}\log p\left(Y \middle| G_X\right) \cong x^T y + C_X + C_Y$$

GMMs Supervectors (3)

Advantages

- Efficient
- Enables (or simplifies) modifying GMM assumptions
  - Channel compensation: Nuisance Attribute Projection
  - Inter-segment variability modeling
  - SVMs for classification
  - Factor analysis
  - Joint Factor analysis

Hidden Markov Model (HMM)

- Define a set of states $\{s_1, s_2, \ldots, s_N\}$
- Process moves from one state to another generating a sequence of states : $s_{i_1}, \ldots, s_{i_t}, \ldots, s_{i_T}$
- Markov assumption: probability of each subsequent state depends only on identity of previous state:

$$p(s_{i_t} \mid s_{i_1}, s_{i_2}, \ldots, s_{i_{t-1}}) = p(s_{i_t} \mid s_{i_{t-1}})$$

- States are hidden, we only see the emitted observations $O = o_1, \ldots, o_t, \ldots, o_T$

Hidden Markov Model (2)

- To define an HMM, the following distributions have to be specified: $\lambda = (\pi, A, B)$
  - Initial probabilities vector $\boldsymbol{\pi}$

$$\pi_i = p(s_i)$$

  - Transition probabilities matrix $\boldsymbol{A}$

$$a_{ij} = p(s_i \mid s_j)$$

  - Emission PDFs $\boldsymbol{b_i}$

$$b_i(o_t) = p(o_t \mid s_i)$$

Three Basic Problems of HMM

1. Evaluation

   Given the observation sequence O and a model λ, how do we efficiently compute the likelihood of the observations $p(O|\lambda)$?

2. Decoding

   Given the observation sequence O, and the model λ, how do we find the optimal state sequence?

3. Training

   How do we learn the model parameters λ to maximize $p(O|\lambda)$?

Three Basic Problems of HMM

1. Evaluation

   Efficiently compute the likelihood $p(O|\lambda)$?

   Solution

   Dynamic programming

   Use

   Model selection

$$p(\lambda|O) = \frac{p(O|\lambda)p(\lambda)}{p(O)}$$

Three Basic Problems of HMM

2. Decoding

Find the optimal state sequence

Solution

The **Viterbi** algorithm finds the optimal sequence using dynamic programming

$$\arg\max_{q_1, q_2, \ldots, q_T} p\left(s_1, s_2, \ldots, s_T = q_1, q_2, \ldots, q_T \middle| O, \lambda\right)$$

Alternatively, posterior probabilities may be computed

$$p\left(s_t = q_t \middle| O, \lambda\right)$$

Three Basic Problems of HMM

3. Training

Learning the model parameters λ

Solution

i. Initialize either state sequence $s_1, s_2, \ldots, s_T$ or model λ

ii. Iterate (EM algorithm)

- Given posteriors, estimate model
- Given model, estimate posteriors

HMM - Examples

- Speech recognition
  - Each context-dependent phoneme is represented by 3-5 HMM states
  - Labeled speech is used to train the HMM
- Speaker diarization
  - Each speaker is represented by an HMM state
  - When number of speakers is unknown, how do we set the number of states?
  - How do we train the HMM?

Outline

1. Introduction
2. Speech processing
3. **Voice activity detection**
4. Classic diarization methods
5. Speaker recognition
6. Advanced diarization methods
   - Geometry
   - Clustering
   - Techniques

Voice Activity Detection (VAD)

Goal

- Segment speech into speech and non-speech (silence, noise, tones, music)

Uses

- Non-speech removal (SID, LID, diarization, ASR)
- Sentence segmentation (diarization, ASR)
- Speech detection (enhancement, transmission)

Performance measures

- False acceptance / miss-detection

Features for VAD: **Energy**

- Log-energies can be computed per sub-band
- Log-energy should be normalized
  - Find 90% and 10% percentiles and set threshold to a weighted average
  - Estimate one Gaussian for speech, and one for silence
- Issues
  - Low SNRs
  - Music and tones
  - Mix of loud and weak speakers

Features for VAD: **Pitch**

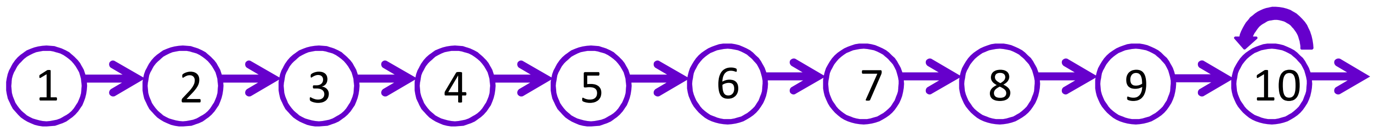
- Existence of pitch is a cue for speech
- Unvoiced speech should be covered by some sort of smoothing over time
- Issues
  - Reliable and robust pitch detection
  - Music and tones

Features for VAD: **MFCC**

- MFCCs contain the required information
- Advanced modeling techniques are required
  - ASR
  - Phonetic decoding
- Issues
  - Availability of training data
  - Language dependency
  - Robustness
- Variants
  - PLP (Perceptually Linear Predictive)
  - FDLP (Frequency domain linear prediction)

HMM for VAD

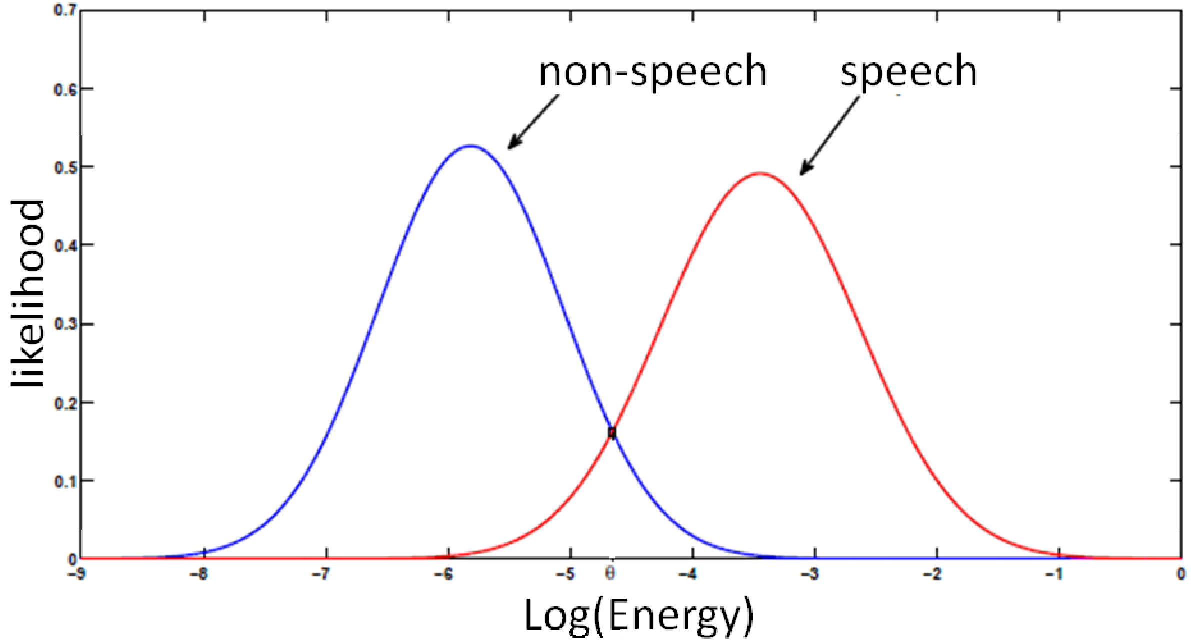
- Each class (speech, silence, noise, music, tone) is represented by a sequence of states



- Minimum duration is enforced by number of states
- Transition probabilities
  - either hand tuned or trained
- Emission PDFs
  - either hand tuned or trained

HMM for VAD: Example #1

- Features: log-energies
- Emission PDFs are assumed to be Gaussian



- Gaussians may be estimated using EM

HMM for VAD: Example #2

- Features: MFCCs
- Emission PDFs are pre-trained on labeled data
  - GMMs
  - SVMs
- HMM is primarily used for Viterbi decoding
- Emission PDFs may be retrained using segmentation

Phoneme recognizer for VAD

- Phonetic decoding is used to transcribe the signal into
  - Phonemes
  - Silence
  - Noise
- The Hungarian phoneme recognizer tool is widely used for telephone speech [Schwarz 09’]
- Issues
  - Language dependency
  - Robustness
  - Time complexity

Segmental Modeling for VAD

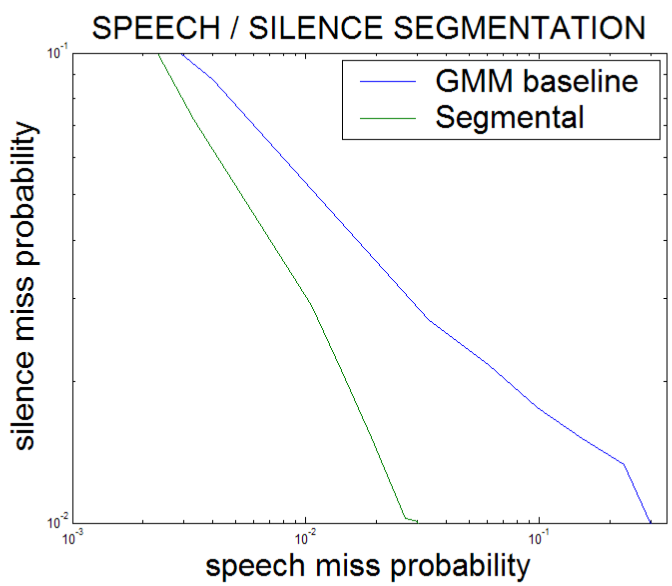
- High level features are extracted on the segment level
  - Audio is segmented into evenly spaced overlapping segments (length is 300ms)
  - Every segment is represented by a GMM supervector
  - Data (train & test) is represented by sequences of supervectors
- Audio classes modeled in supervector space using GMMs
- HMM is used to integrate local scores

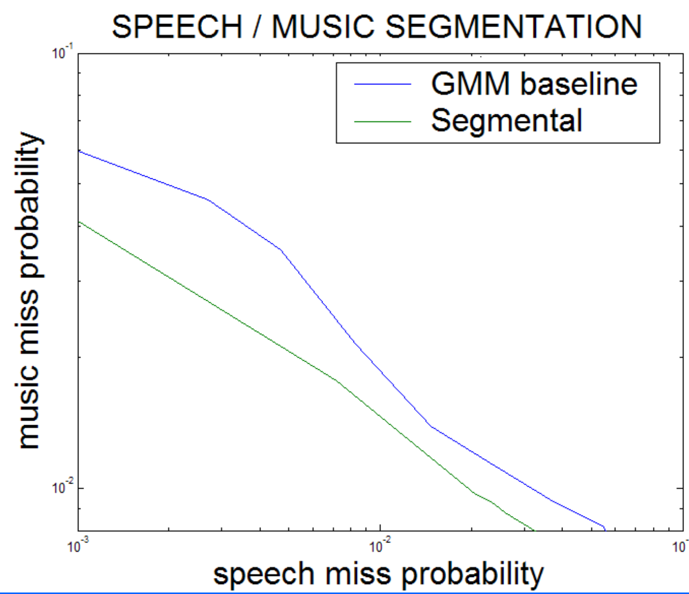
H.A. "Segmental modeling for speech segmentation", in Proc. *ICASSP* 2007

Segmental Modeling: Experiments

Data: BN in Arabic (GALE)

System	speech vs. silence EER in (%)	Speech vs. music EER in (%)
Baseline (GMM)	2.9	1.4
Segmental	1.7	1.3





Deep Neural Networks (DNN) for VAD

Setup

- DARPA RATS: audio transmitted over extremely noisy and/or highly distorted channels
- VADs are channel dependent
- HMM is used on top of GMM/DNN

Baseline

- GMMs with 1024 Gaussians/class

G. Saon et al., “The IBM Speech Activity Detection System for the DARPA RATS Program”, in Proc. *Interspeech* ,2013.

DNNs: Features

- PLP Cepstra + probability of voicing
  - Mean and variance normalization (for Cepstra only)
  - ARMA filtering per dimension with a temporal window of 20 frames (for Cepstra only)
  - Splicing of 17 consecutive frames
  - LDA projection to 40 dimensions
  - Augment with 1^{st},2^{nd} and 3^{rd} deltas (40 → 4x40)
- FDLP (13)
  - Same processing as for PLP Cepstra

DNNs: Architecture

- 3 hidden layers: 1024 neurons each
- Output layer has 3 neurons:
  Speech, noisy-silence and no-transmission

Training

1. Fully train n hidden layers
2. Use trained network to initialize a network with n+1 hidden layers
3. Iterate to 1

DNNs for VAD - Results

Method	Features	Amount of training data (in hrs)	DEV1 (EER in %)	DEV2 (EER in %)
GMM	PLP+v	200	2.0	3.3
GMM	PLP+v	2000	2.0	3.3
NN 1 hidden layer	PLP+v	200	1.8	2.6
NN 2 hidden layers	PLP+v	200	1.7	2.6
NN 2 hidden layers	PLP+v+FDLP	2000	1.5	2.3
NN 3 hidden layers	PLP+v+FDLP	2000	1.2	2.1

Outline

1. Introduction
2. Speech processing
3. Voice activity detection
4. **Classic diarization methods**
5. Speaker recognition
6. Advanced diarization methods
   - Geometry
   - Clustering
   - Techniques

Classic Diarization Methods

Components

- VAD
- Segmentation
- Clustering
- Viterbi resegmentation

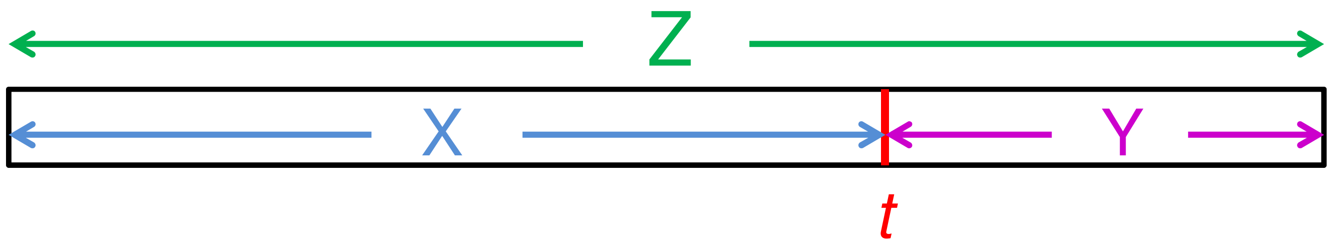
Architectures

- Bottom-up vs. top-down

Number of speakers

- Two vs. unknown

Speaker Change Detection



- H_0: no speaker change at time *t*
- H_1: speaker change point at time *t*

$$L_0 = \log p(Z \mid M_Z)$$
$$L_1 = \log p(X \mid M_X) + \log p(Y \mid M_Y)$$

M: statistical model (usually a Gaussian)

- GLR criterion [Gish 91']: $d_{GLR} = L_1 - L_0$
- BIC criterion [Chen 98']: $d_{BIC} = d_{GLR} - (P_1 - P_o)$
- KL2-dist [Seigler 97']: $d_{KL2} = \mathrm{KL}(M_X, M_Y) + \mathrm{KL}(M_Y, M_X)$
- CLLR [Reynolds 98']: $d_{CLLR} = \log \dfrac{p(X|M_Y)}{p(X|UBM)} + \log \dfrac{p(Y|M_X)}{p(Y|UBM)}$

Bayesian Information Criterion

- A criterion for model selection among a set of models $\{M_i\}$

$$BIC(M_i) = \log p(X \mid M_i) - \underbrace{\tfrac{1}{2}\lambda(\#M_i)\log|X|}_{\text{penalty}}$$

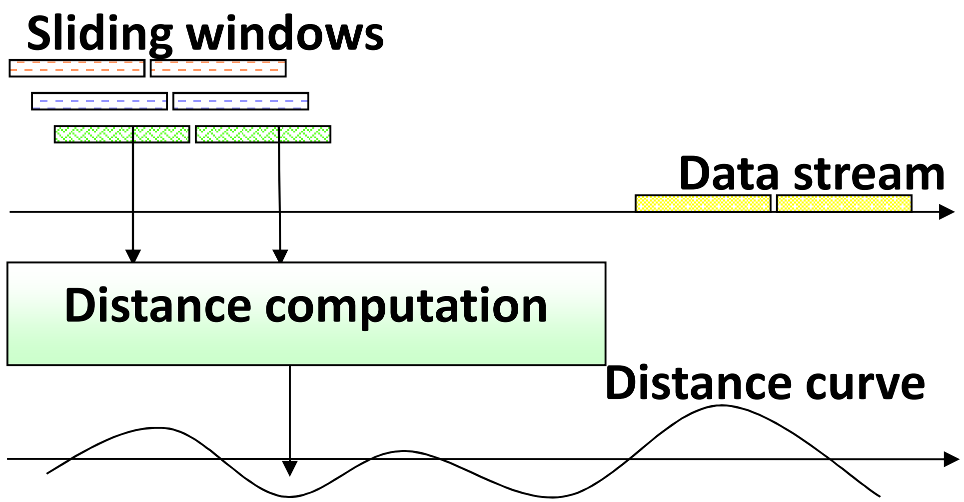
- In theory λ=1 but is tuned in practice
- $\#M_i$: the number of parameters in model M_i
- $|X|$: length of X (number of frames)
- For Gaussian models in R^d:

$$\Delta BIC = |Z|\log|\hat{\Sigma}_Z| - |X|\log|\hat{\Sigma}_X| - |Y|\log|\hat{\Sigma}_Y| - \tfrac{\lambda}{2}\left(d + \tfrac{d(d+1)}{2}\right)\left(\log|Z|\right)$$

Segmentation

Fixed-size analysis running window

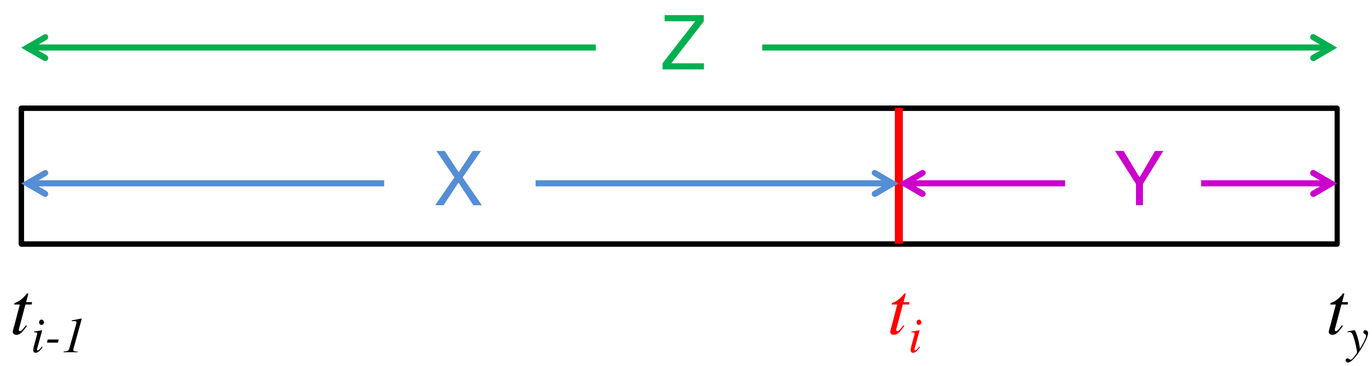
- ΔBIC is used for distance computation



- Issues
  - Short window: ΔBIC is too noisy
  - Long window: short speaker-turns may be missed

Segmentation

Variable-size analysis running window



- t_{i-1} is latest detected speaker change
- $[t_{i-1}\ t_y]$ is variable-size analysis window increasing until speaker change is detected
- t_i is hypothesized speaker change
- Issues: past errors may be propagated

Segmentation

Other BIC-based approaches

- Top-down speaker change detection [Wu 06‘]
  - Find single best scoring speaker change analyzing whole session (if score is too low, terminate)
  - Partition utterance using detected speaker change
  - Recursively apply method on both parts of the partitioned utterance
- Dynamic programming [Cettolo 05’]
  - Find optimal segmentation according to BIC criterion

Speaker Clustering

Hierarchical agglomerative clustering (HAC)

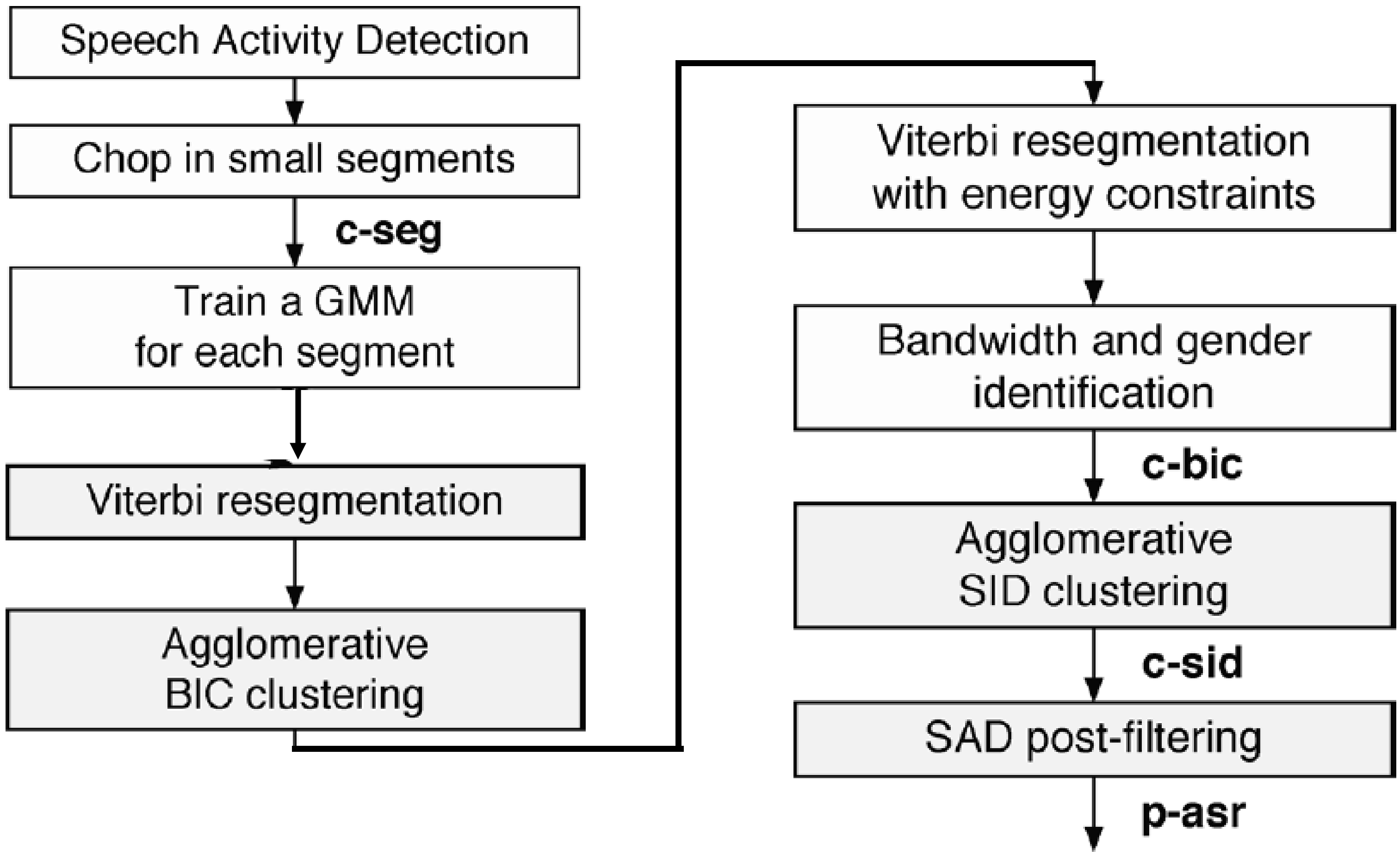
[Chen 98'; Barras 06']

- Initiate each segment as a distinct cluster
- Iterate:
  - Find closets pair of clusters
    - Distance measure: ΔBIC / CLLR
  - Merge closets pair
  - Stopping criteria
    - Local BIC
    - Global BIC

Viterbi Resegmentation

- GMMs/HMMs are trained per cluster
- Viterbi is used for obtaining a new segmentation
- Energy local minima [Barras 06'] may be used to refine boundaries
  - Boundaries are shifted to the nearest point of low energy within an interval of 1s
  - Purpose is to avoid cutting words
- GMMs/HMMs are re-estimated
- Process may be reiterated

Multistage Diarization for BN (LIMSI)



C. Barras, et al., "Multi-Stage Speaker Diarization of Broadcast News," IEEE TASLP, vol. 14, no. 5, 2006 (and image credit)

Agglomerative SID Clustering

Purposes

- Once an initial clustering is early stopped, more complex models can be used with more data per cluster
- Robust modeling is necessary to regroup multiple clusters obtained from a single speaker
(due to noise, music, etc.)

Agglomerative SID Clustering

Outline

- Features: 15 MFCC+15 ΔMFCC+ ΔEnregy
- Feature warping [Pelecanos 01']
- Each cluster is used to train a speaker recognition model
- Agglomerative clustering is performed separately for each gender and bandwidth condition using CLLR

$$d_{CLLR} = \frac{1}{|X|}\log\frac{p(X|M_Y)}{p(X|UBM)} + \frac{1}{|Y|}\log\frac{p(Y|M_X)}{p(Y|UBM)}$$

- Stopping criterion: comparing d_{CLLR} to a threshold δ

Multistage Diarization for BN: Results

Datasets

- NIST RT-04F
  (Fall 2004 Rich Transcription evaluation)
- The French data from the ESTER BN evaluation

System	RT-04F test DER (in %)	ESTER test DER (in %)
No agglomerative SID clustering	17	13.8
With agglomerative SID clustering	9.1	11.5

Outline

1. Introduction
2. Speech processing
3. Voice activity detection
4. Classic diarization methods
5. **Speaker recognition**
6. Advanced diarization methods
   - Geometry
   - Clustering
   - Techniques

Speaker Recognition

Definition

Given enrollment session X and test session Y, are they the same speaker?

Progress of the state-of-the-art

Algorithm	Year	EER (in %)
GMM	1995	>10
GMM-UBM + score-norm	2000	6.2
GMM-supervectors	2004	6.2
NAP / WCCN / Eigen-channels	2005	3.6
JFA	2006	1.4
i-vectors + PLDA	2011	1.0

Setup: NIST 10' SRE, telephone data

Speaker Recognition: **GMMs**

Every speaker is treated as a bag-of-frames

Enrollment

- Speaker is modeled by a GMM using ML estimation (EM training)

Test

- Likelihood of every test-frame is computed given enrolled GMM
- Score is average log-likelihood of test frames

$$S(Y|X|) = \frac{1}{|Y|}\log p(Y|X) = \frac{1}{|Y|}\log \prod_{t=1}^{|Y|} p(y_t|X) = \langle \log p(y_t|G_X) \rangle$$

Speaker Recognition: **GMM-UBM**

- A UBM is estimated from a large dev-set
- UBM is used as a prior for speaker models

Enrollment

- Speaker is modeled by a GMM using MAP adaptation from the UBM (which serves as a prior)

Speaker Recognition: **Score Norm**

$$p(S|X) = \frac{p(X|S)\,p(S)}{p(X)}$$

1. UBM based normalization: $p(X) \cong p(X|UBM)$
2. In practice, score normalization methods improve accuracy
3. Normalization works because it cancels out some effects of channel, noise and modeling inaccuracies
4. Is beneficial for normalizing scores in diarization

Znorm

- Znorm [Auckenthaler 00'] standardizes the distribution of scores for speaker S given a calibration set of imposter test sessions
- Let $\varphi(S,Y)$ be the score of session Y for speaker S
- Estimate mean and variance of $\varphi(S,\cdot)$

$$\mu_Z(S,\cdot) = E_Y \varphi(S,Y)$$

$$\sigma_Z(S,\cdot) = \sqrt{V_Y \varphi(S,Y)}$$

- Standardize $\varphi(S,Y)$

$$\varphi_{Znorm}(S,Y) = \frac{\varphi(S,Y) - \mu_Z(S,\cdot)}{\sigma_Z(S,\cdot)}$$

Tnorm

- Tnorm [Auckenthaler, 00'] standardizes the distribution of scores for test session Y given a calibration set of imposter speakers
- Estimate mean and variance of $\varphi(\cdot, Y)$

$$\mu_T(\cdot, Y) = E_S \varphi(S, Y)$$
$$\sigma_T(\cdot, Y) = \sqrt{V_S \varphi(S, Y)}$$

- Standardize $\varphi(S,Y)$

$$\varphi_{Tnorm}(S, Y) = \frac{\varphi(S, Y) - \mu_T(\cdot, Y)}{\sigma_T(\cdot, Y)}$$

Combining Znorm and Tnorm

ZTnorm [H.A. 05']

- Apply Znorm followed by Tnorm
- Calibration scores for Tnorm must be Znorm-ed

$$\mu_{ZT}(\cdot,Y) = E_S \varphi_{Znorm}(S,Y)$$

$$\sigma_{ZT}(\cdot,Y) = \sqrt{V_S \varphi_{Znorm}(S,Y)}$$

- Standardize $\varphi(S,Y)$

$$\varphi_{ZTnorm}(S,Y) = \frac{\varphi_{Znorm}(S,Y) - \mu_{ZT}(\cdot,Y)}{\sigma_{ZT}(\cdot,Y)}$$

Snorm [Shum 10']

- Sum Znorm and Tnorm scores

$$\varphi_{Snorm}(S,Y) = \varphi_{Znorm}(S,Y) + \varphi_{Tnorm}(S,Y)$$

Speaker Recognition: **GMM Supervectors**

- Similar modeling as for GMM-UBM

Paradigm shift [H.A. 04']

- Embed both enrollment and test sessions into a supervector space using GMM parameters
- Compute score in the embedded space

$$\varphi(X,Y)=\sum_g \left(\sqrt{w_g}\,\Sigma_g^{-\frac{1}{2}}\mu_g^X\right)^T\left(\sqrt{w_g}\,\Sigma_g^{-\frac{1}{2}}\mu_g^Y\right)$$

Normalized GMM-supervector of X Normalized GMM-supervector of Y

- Improved flexibility and speed

Modeling Session Variability

PDF of frame is dependent on both speaker and session

- Inter-session intra-speaker variability (ISISV)
- channel

Methods

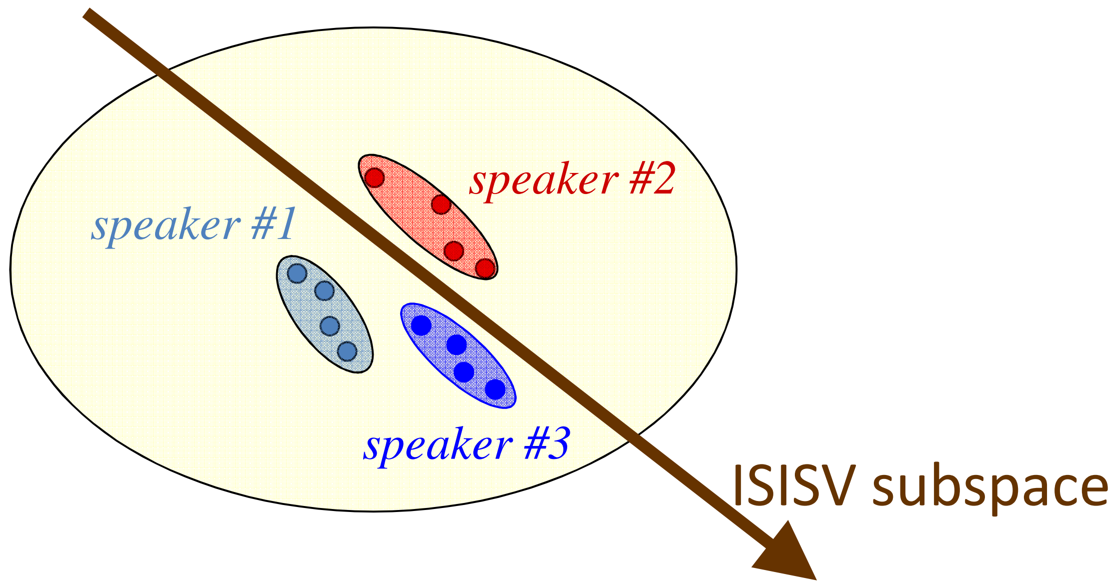
- ISIS [H.A. 05']
- WCCN [Hatch 06']
- NAP [Campbell 06']

Framework

- Assume ISISV is shared among all speakers
- Estimate variability from dev data
- Use ISISV modeling for supervector cleanup/better scoring

Nuisance Attribute Projection (NAP)

- Sessions are mapped into supervectors
- ISISV is assumed to be restricted to a low dimensional subspace
- ISISV subspace is estimated from a dev set



NAP (2)

ISISV estimation

1. Compute the within-speaker covariance matrix

$$\hat{\mathrm{W}} = \tfrac{1}{S}\sum_{s=1}^{S}\tfrac{1}{n_s}\sum_{i=1}^{n_s}\left(z_i^s - \overline{z}^s\right)\left(z_i^s - \overline{z}^s\right)^T$$

2. Find top eigenvectors using PCA → V
3. Projection $\mathrm{P}=(\mathrm{I}-\mathrm{V}\mathrm{V}^{\mathrm{T}})$ removes the estimated ISISV subspace

Enrollment

- Apply P on enrollment supervector: $x \rightarrow \mathrm{P}x$

Scoring

- No need of applying P on test supervector y

$$\varphi_{NAP}(x, y) = (\mathrm{P}x)^{\mathrm{T}}\mathrm{P}y = x^{\mathrm{T}}\mathrm{P}^{\mathrm{T}}\mathrm{P}y = (\mathrm{P}x)^{\mathrm{T}} y$$

Within Class Covariance Normalization (WCCN)

Motivation

- ISISV and Inter-speaker subspaces are not disjoint
- Instead of removing the ISISV subspace, deemphasize it

Estimation

- W is estimated similarly to NAP and inverted
- W may be smoothed

$$\dot{\mathrm{W}} = (1-\alpha)\hat{\mathrm{W}} + \alpha \mathrm{I}$$

or

$$\ddot{\mathrm{W}} = (1-\alpha)\hat{\mathrm{W}} + \alpha \cdot \mathrm{diag}(\hat{\mathrm{W}})$$

Scoring

$$\varphi_{WCCN}(x, y) = x^{\mathrm{T}} \mathrm{W}^{-1} y$$

Joint Modeling of Session & Speaker Variabilities

Innovation

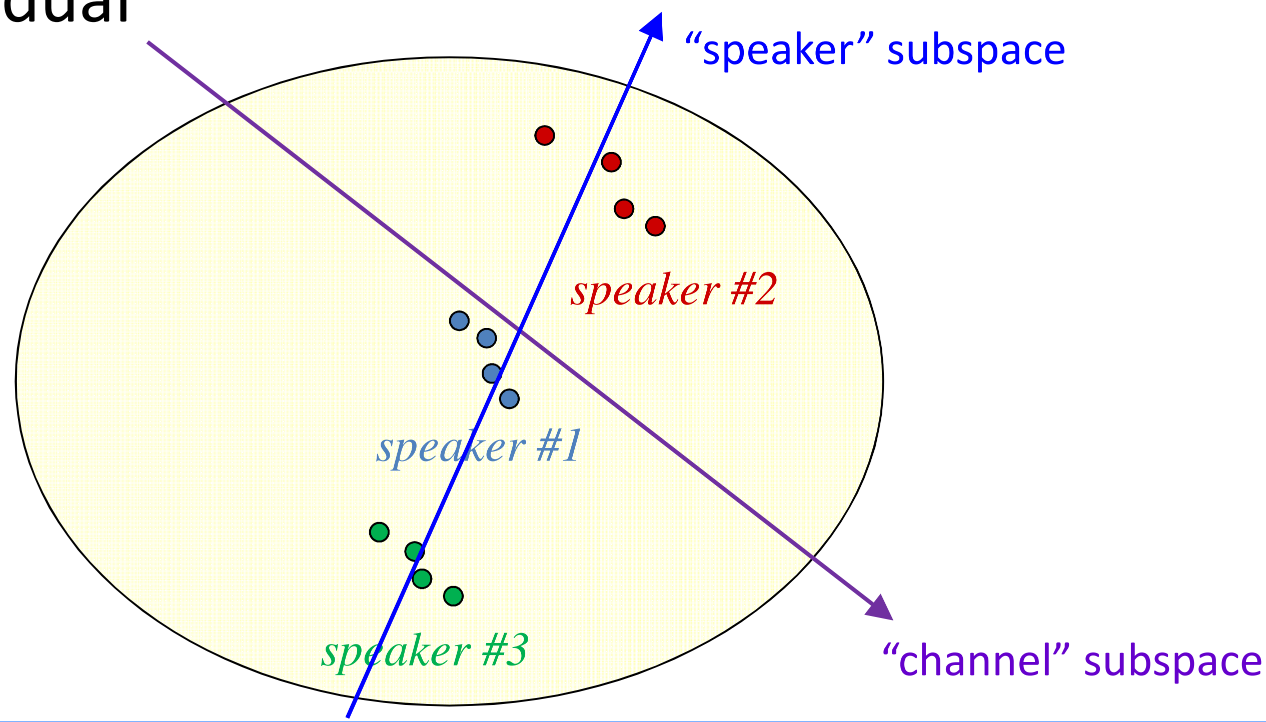
- Inter-speaker variability (ISV) is modeled jointly with ISISV
- Supervectors are estimated using MAP adaptation with strong priors
(contrary to the NAP/WCCN framework)

Methods

- JFA [Kenny 08']
- i-vector PLDA [Dehak 11'; Kenny 10'; Romero 11']

Joint Factor Analysis (JFA)

- Sessions are mapped into supervectors
- ISISV is restricted to a low-dimensional subspace
- ISV is decomposed into a low-dimensional subspace and a residual



JFA Model

$$M = m + \underbrace{Vy + Dz}_{s} + \underbrace{Ux}_{c}$$

M: supervector for a given session

m : overall mean (UBM supervector)

s: speaker dependent supervector

V : rectangular matrix of low-rank (eigenvoices)

D : diagonal matrix

y,z: random vectors with standard normal prior (speaker factors)

c: channel dependent supervector

U : rectangular matrix of low rank *(*eigenchannels*)*

x: random vector with standard normal prior (channel factors)

JFA: Speaker Supervector Estimation

1. Estimate Baum-Welch (BW) statistics for enrollment session
   - Counts: $N_g = \sum_t p(g \mid X_t)$
   - Sums: $F_g = \sum_t p(g \mid X_t) X_t$

   using $p(g|X_t) = \frac{w_g p(X_t \mid N(\mu_g, \Sigma_g))}{\sum_{\tilde{g}} w_{\tilde{g}} p(X_t \mid N(\mu_g, \Sigma_g))}$
2. Estimate MAP values of speaker factors y and z given BW statistics
3. Estimated speaker supervector is $\hat{s} = V\hat{y} + D\hat{z}$

JFA Scoring

Fast scoring [Glembeck 09']

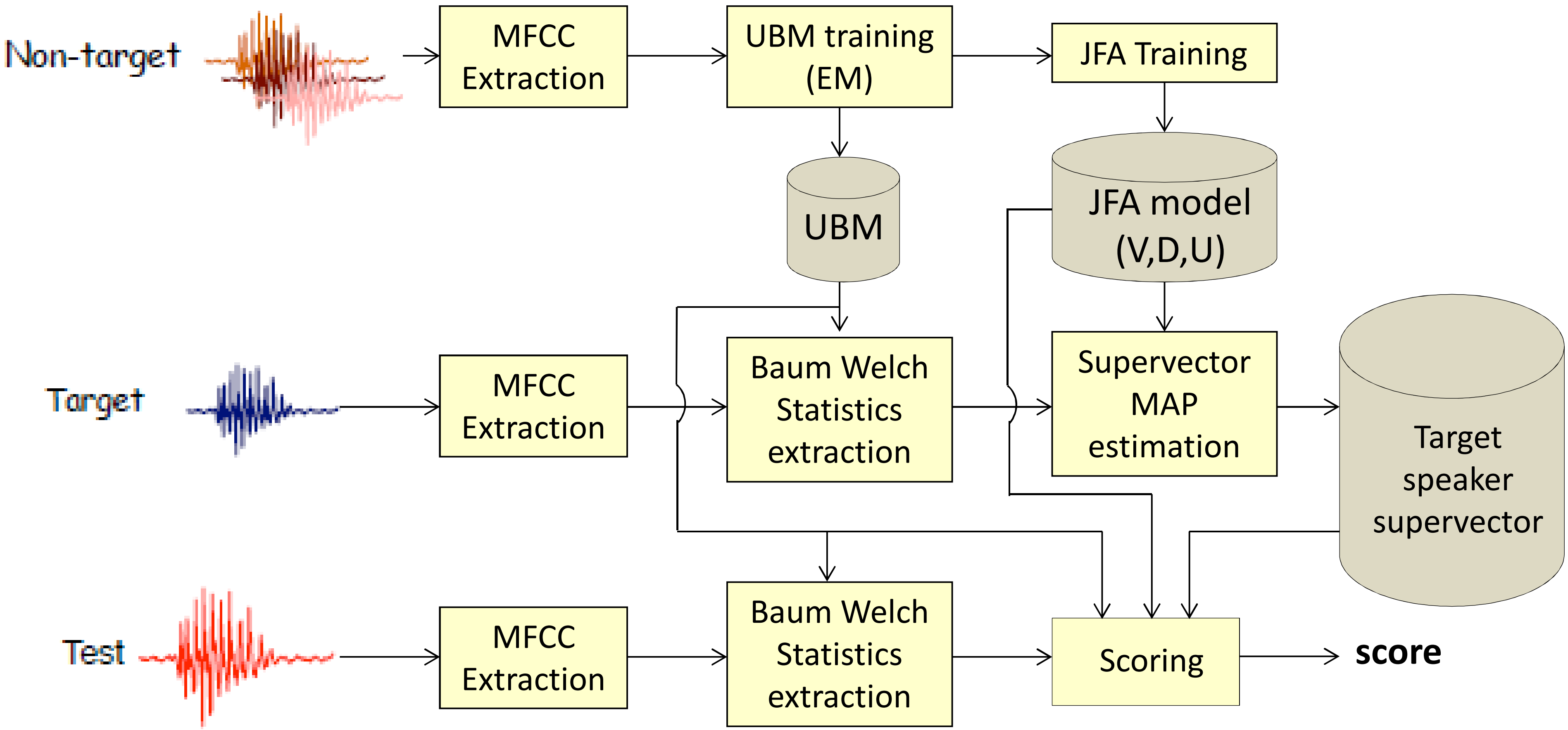
1. Estimate BW statistics for test session Y
2. Estimate MAP values of channel factors x given BW statistics
3. Remove channel and UBM effect from BW statistics

$$\tilde{F} = F - NUx - Nm$$

4. Linear Scoring $\dfrac{\hat{s}^t \Sigma^{-1} \tilde{F}}{|Y|}$

Σ is a stacking of the UBM covariance matrices

JFA Outline


Non-target
MFCC Extraction
UBM training (EM)
JFA Training
UBM
JFA model (V,D,U)
Target
MFCC Extraction
Baum Welch Statistics extraction
Supervector MAP estimation
Target speaker supervector
Test
MFCC Extraction
Baum Welch Statistics extraction
Scoring
score

i-vectors

Factor analysis as a high-level feature extractor

$$M = m + T\phi$$

M: supervector for a given session

m : overall mean (UBM supervector)

T : rectangular matrix of low-rank (total variability matrix)

ϕ: standard normal random vector (**i-vector**)

- i-vectors capture both speaker and channel variabilities
- i-vectors are extracted symmetrically for both enrollment and test sessions

i-vector Extraction

1. Estimate Baum-Welch (BW) statistics
   - Counts: $N_g = \sum_t p(g \mid X_t)$
   - Sums: $F_g = \sum_t p(g \mid X_t) X_t$

   using $p(g|X_t) = \dfrac{w_g p(X_t \mid N(\mu_g, \Sigma_g))}{\sum_{\tilde{g}} w_{\tilde{g}} p(X_t \mid N(\mu_g, \Sigma_g))}$

2. i-vector MAP estimate is $\hat{\phi} = L^{-1} T^T \Sigma^{-1} F$

   with $L = \mathrm{I} + T^T \Sigma^{-1} N T$

   and Σ is a stacking of the UBM covariance matrices

Probabilistic Linear Discriminant Analysis (PLDA)

- The PLDA framework assumes that i-vectors distribute according to:

$$\phi = \mu + s + c$$

ϕ - i-vector

μ - global mean

s - speaker component

c - channel / ISISV component

- s and c are assumed to distribute normally:

$$s \sim N(0, B)\ ,\ c \sim N(0, W)$$

- The PLDA model is parameterized by $\{\mu, B, W\}$

PLDA: Details

PLDA training

- Given a dev set, hyper-parameters W and B are trained using EM

PLDA scoring

- Given i-vectors x,y:

$$score = \frac{p(x,y|H_=)}{p(x,y|H_{\neq})} = \frac{p(x,y|H_=)}{p(x)p(y)}$$

with

$$p(x,y|H_=) = \int_s p(x|s)\,p(y|s)\,p(s)\,ds = N\left(\begin{bmatrix} x \\ y \end{bmatrix}; \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \Sigma_{tot} & B \\ B & \Sigma_{tot} \end{bmatrix}\right)$$

$$p(x) = \int_s p(x|s)\,p(s)\,ds = N(x; \mu, \Sigma_{tot})$$

$$\Sigma_{tot} = B + W$$

PLDA – Details (cont.)

PLDA scoring (cont.)

for μ=0:

$$score = x^T Qx + y^T Qy + 2x^T Py + const$$

with

$$Q = \Sigma_{tot}^{-1} - \left(\Sigma_{tot} - B\Sigma_{tot}^{-1}B\right)^{-1}$$

$$P = \Sigma_{tot}^{-1}B\left(\Sigma_{tot} - B\Sigma_{tot}^{-1}B\right)^{-1}$$

Outline

1. Introduction
2. Speech processing
3. Voice activity detection
4. Classic diarization methods
5. Speaker recognition
6. **Advanced diarization methods**
   - **Geometry**
   - Clustering
   - Techniques

Advanced Diarization Methods: **Geometry**

1. High level features
2. Intra-speaker variability modeling
3. PLDA
4. PCA
5. Spectral clustering
6. Score normalization

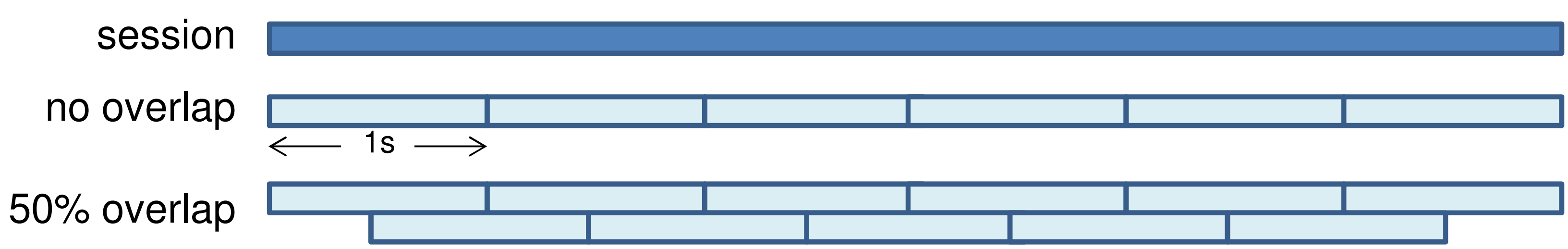
Advanced Diarization Methods: **Geometry**

1. **High level features**
2. Intra-speaker variability modeling
3. PLDA
4. PCA
5. Spectral clustering
6. Score normalization

High-Level Features

High-level features may be used to parameterize

- Segments
- Clusters
- Superframes
  - Audio is divided into evenly spaced segments (typically 1s)
  - Superframes may overlap



Kernel-PCA

- Define a set of anchor sessions $s_1,\ldots,s_n$
- Choose a kernel (e.g. the supervector-based kernel)

$$\varphi(X,Y)=\sum_g\left(\sqrt{w_g}\Sigma_g^{-\frac{1}{2}}\mu_g^X\right)^T\left(\sqrt{w_g}\Sigma_g^{-\frac{1}{2}}\mu_g^Y\right)$$

- Apply the Kernel-PCA framework to define high-level feature $\mathrm{T}(X)\in R^n$

$$\mathrm{T}: X \rightarrow R^n = V\begin{pmatrix}\ldots\\ \varphi(X,s_i)\\ \ldots\end{pmatrix}$$

where *V* stacks the eigenvectors of matrix $\Gamma:\Gamma_{i,j}=\varphi(s_i,s_j)$

H.A., "Trainable speaker diarization", in Proc. *Interspeech*, 2007

Supervectors

- Represent audio (segment/cluster/superframe) with a GMM supervector

$$T : X \rightarrow x$$

$$x_{g*D+d} = \sqrt{w_{g,d}^{UBM}} \frac{\mu_{g,d}^{X}}{\sigma_{g,d}^{UBM}}$$

- UBM may be trained on
  - Development data [H.A. 12']
  - Processed session [H.A. 10']

H.A., "Unsupervised Compensation of Intra-Session Intra-Speaker Variability for Speaker Diarization", in Proc. *Speaker Odyssey*, 2010

Speaker Factors

- Represent a superframe with speaker factors *y*

$$M = m + Vy$$

M: supervector for a given superframe

m : overall mean (UBM supervector)

V : rectangular matrix of low-rank (eigenvoices)

y : random vector with a standard normal prior (speaker factors)

- Typically 20 speaker factors are used
- FA model is trained on development data

F. Castaldo, et al., "Stream based speaker segmentation using speaker factors and eigenvoices," in Proc. *ICASSP*, 2008.

i-vectors

- Represent a superframe with an i-vector

$$M = m + T\phi$$

M : supervector for a given session

m : overall mean (UBM supervector)

T : rectangular matrix of low-rank (total variability matrix)

ϕ : standard normal random vector (**i-vector**)

- i-vector dimension is 100
- FA Model is trained on development data

S. Shum, "Unsupervised Methods for Speaker Diarization," S.M. Thesis, MIT Department of Electrical Engineering and Computer Science, 2011

Advanced Diarization Methods: **Geometry**

1. High level features
2. **Intra-speaker variability modeling**
3. PLDA
4. PCA
5. Spectral clustering
6. Score normalization

Intra-Session Intra-Speaker Variability

Intra-session intra-speaker variability (ISISV) is why speaker diarization is such a challenge

- Phonetic variability
- Energy level variability
- Acoustic (speaker intrinsic)
- Speech rate
- Non-speech rate (VAD errors)

ISISV can be modeled and compensated

H.A., "Trainable speaker diarization", in Proc. *Interspeech*, 2007

Modeling ISISV

High-level features of a given speaker assumed to distribute normally, with a shared covariance matrix

$$h_t^s \sim N\left(\mu_s, W\right)$$

Use of ISISV modeling in Diarization

- Define a distance measure using WCCN [H.A. 2007]

$$\varphi_{WCCN}\left(x, y\right) = x^{\mathrm{T}} W^{-1} y$$

- Compensation of ISISV using NAP [H.A. 2010]
  - A subspace is found applying PCA on W
  - Subspace is removed from high-level features
- Compensation of low-level features (fNAP) [H.A., 11']

Supervised Estimation of ISISV

Given a labeled dataset

i, j, s : session, superframe and speaker indices

$z_{i,j}^{s}$: high-level features

$\bar{z}_{i}^{s}$: mean of high-level features of speaker s in session i

$$\hat{W} = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s}\frac{1}{n_{s,i}}\sum_{j=1}^{n_{s,i}}\left(z_{i,j}^{s} - \bar{z}_{i}^{s}\right)\left(z_{i,j}^{s} - \bar{z}_{i}^{s}\right)^{T}$$

H.A., "Trainable speaker diarization", in Proc. *Interspeech*, 2007

Unsupervised Estimation of ISISV

Given a time series of high-level features $h_1,\ldots, h_T$

h_t is modeled as a sum of two components

$$h_t = \mu_{s_t} + n_t$$

where

μ_{s_t} : speaker dependent mean of the speaker at time *t*

n_t : intra-speaker inter-session variability; $n_t \sim N(0,W)$

H.A., "Unsupervised Compensation of Intra-Session Intra-Speaker Variability for Speaker Diarization", in Proc. *Speaker Odyssey*" 2010

Unsupervised Estimation of ISISV (2)

Consider the difference between two consecutive high level features:

$$h_t - h_{t-1} = \mu_{s_t} - \mu_{s_{t-1}} + n_t - n_{t-1}$$

B : between speaker covariance matrix

τ : mean speaker turn length

$$W = \tfrac{1}{2}\operatorname{cov}\left(h_t - h_{t-1}\right) - \frac{B}{\tau}$$

$$\cong \tfrac{1}{2}\operatorname{cov}\left(h_t - h_{t-1}\right)$$

$$\hat{W} = \frac{1}{2(T-1)} \sum_{t=2}^{T} \left(h_t - h_{t-1}\right)\left(h_t - h_{t-1}\right)^T$$

Advanced Diarization Methods: **Geometry**

1. High level features
2. Intra-speaker variability modeling
3. **PLDA**
4. PCA
5. Spectral clustering
6. Score normalization

PLDA for Speaker Clustering

PLDA scores are used to score cluster pairs

Experiments

- Data: COST278 multilingual pan-European BN database
  - Training: Two hours per language
  - Test: One hour per language
- GMM (256 Gaussians), i-vectors (400 dimensional), 200 speaker and channel factors
- 36% relative error reduction over a BIC-based AHC baseline

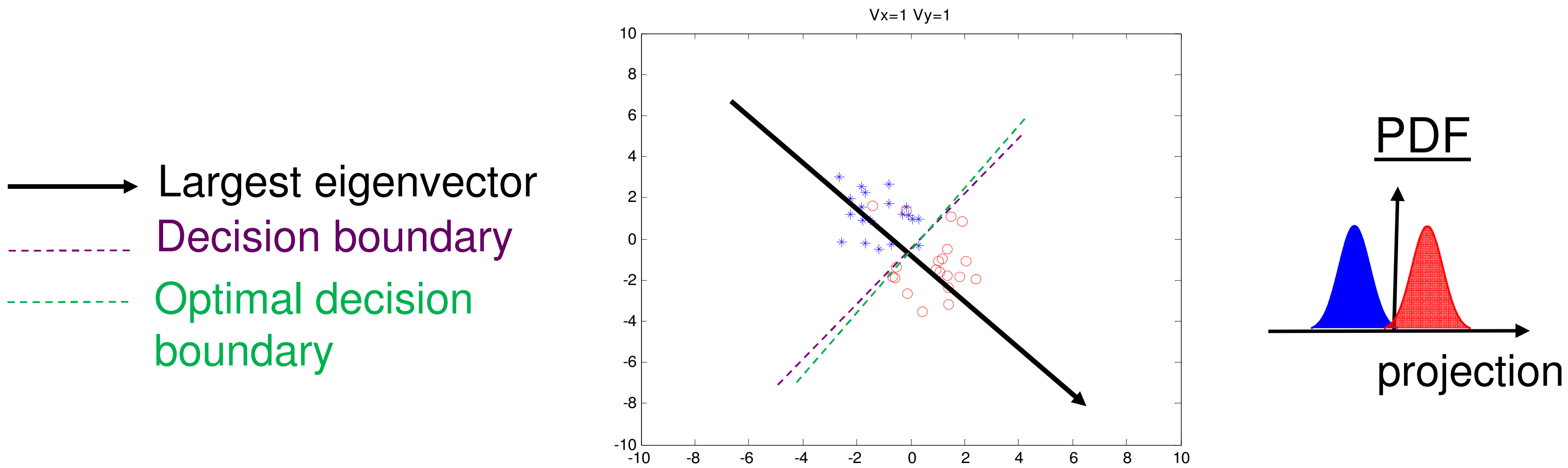
J. Silovsky, et al., "PLDA-based Clustering for Speaker Diarization of Broadcast Streams", in Proc. *Interspeech*, 2011

Advanced Diarization Methods: **Geometry**

1. High level features
2. Intra-speaker variability modeling
3. PLDA
4. **PCA**
5. Spectral clustering
6. Score normalization

PCA for Speaker Diarization

- PCA is applied on high-level features (superframes)
- Most suitable when number of speakers is known
  - Works very well for diarization of a 2-speaker conversation
- A pre-processing step of ISISV compensation is key



H.A., "Unsupervised Compensation of Intra-Session Intra-Speaker Variability for Speaker Diarization", in Proc. *Speaker Odyssey*" 2010

Advanced Diarization Methods: **Geometry**

1. High level features
2. Intra-speaker variability modeling
3. PLDA
4. PCA
5. **Spectral clustering**
6. Score normalization

Spectral Clustering

- Given a set of parameterized superframes $S=\{h_1,\ldots h_n\}$
- Form affinity matrix

$$A_{i,j} = \begin{cases} e^{-d^2(h_i,h_j)/2\sigma^2} & i \neq j \\ 0 & i = j \end{cases}$$

- Define diagonal matrix $D_{i,i} = \sum_k a_{i,k}$
- Form matrix $L = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$
- Stack k largest eigenvectors of L to form columns of matrix X
  - k can be chosen according to eigenvalues analysis
- Length normalized rows of X form new parameterization:

$$h_i \rightarrow X_i / \|X_i\|$$

S. Shum, et al., "On the Use of Spectral and Iterative Methods for Speaker Diarization," in Proc. *Interspeech*, 2012

Advanced Diarization Methods: **Geometry**

1. High level features
2. Intra-speaker variability modeling
3. PLDA
4. PCA
5. Spectral clustering
6. **Score normalization**

Score Normalization

- Score normalization is very effective for speaker recognition
- In [H.A. 07'] it was shown that score normalization is effective for speaker diarization
  - The need for normalization is more significant for non-BIC based approaches
  - Normalization was done using segments from a development set

Outline

1. Introduction
2. Speech processing
3. Voice activity detection
4. Classic diarization methods
5. Speaker recognition
6. **Advanced diarization methods**
   - Geometry
   - **Clustering**
   - Techniques

Advanced Diarization Methods: **Clustering**

1. K-means
2. Integer Linear Programming (ILP)
3. Fully Bayesian using Variational Bayes (VB)

Advanced Diarization Methods: Clustering

1. **K-means**
2. Integer Linear Programming (ILP)
3. Fully Bayesian using Variational Bayes (VB)

K-Means

- K-means is used to cluster superframes/clusters parameterized by high-level features
- Number of clusters
  - Known a-priori [H.A. 12']
  - Estimated using analysis of the eigenvectors obtained in spectral clustering [Shum 12']:
    Eigenvalues are modeled to exhibit exponential decay
    → Set k according to the first eigenvalue that deviates from the exponential model

Advanced Diarization Methods: **Clustering**

1. K-means
2. **Integer Linear Programming (ILP)**
3. Fully Bayesian using Variational Bayes (VB)

Integer Linear Programming (ILP)

Given an initial clustering of size N

- Clusters should ne highly pure
- Speakers may be divided into several clusters

Every cluster is parameterized with an i-vector

- Other high-level features may be used

Find optimal clustering of the i-vectors

- Minimize number of clusters
- Minimize the dispersion of the i-vectors

G. Dupuy et al., "Recent Improvements on ILP-based Clustering for Broadcast News Speaker Diarization", in Proc. *Speaker Odyssey*, 2014

Minimize: $$\sum_{k \in C} x_{k,k} + \frac{1}{\delta} \sum_{j \in C} \sum_{k \in K_j} d(k,j) x_{k,j}$$

Subject to:

$$x_{k,j} \in \{0,1\} \qquad k \in K_j, j \in C$$

$$\sum_{k \in K_j} x_{k,j} = 1 \qquad j \in C$$

$$x_{k,j} - x_{k,k} < 0 \qquad k \in K_j, j \in C$$

C : the set of i-vectors $C = \{1, \ldots, N\}$

$x_{k,k}$: a binary variable equal to 1 when i-vector *k* is a center

$d(k,j)$: distance between i-vectors k and j

δ : normalization factor

$x_{k,j}$: a binary variable equal to 1 when i-vector *j* is assigned to center k

K_j : all i-vector within a radius of δ from i-vector j

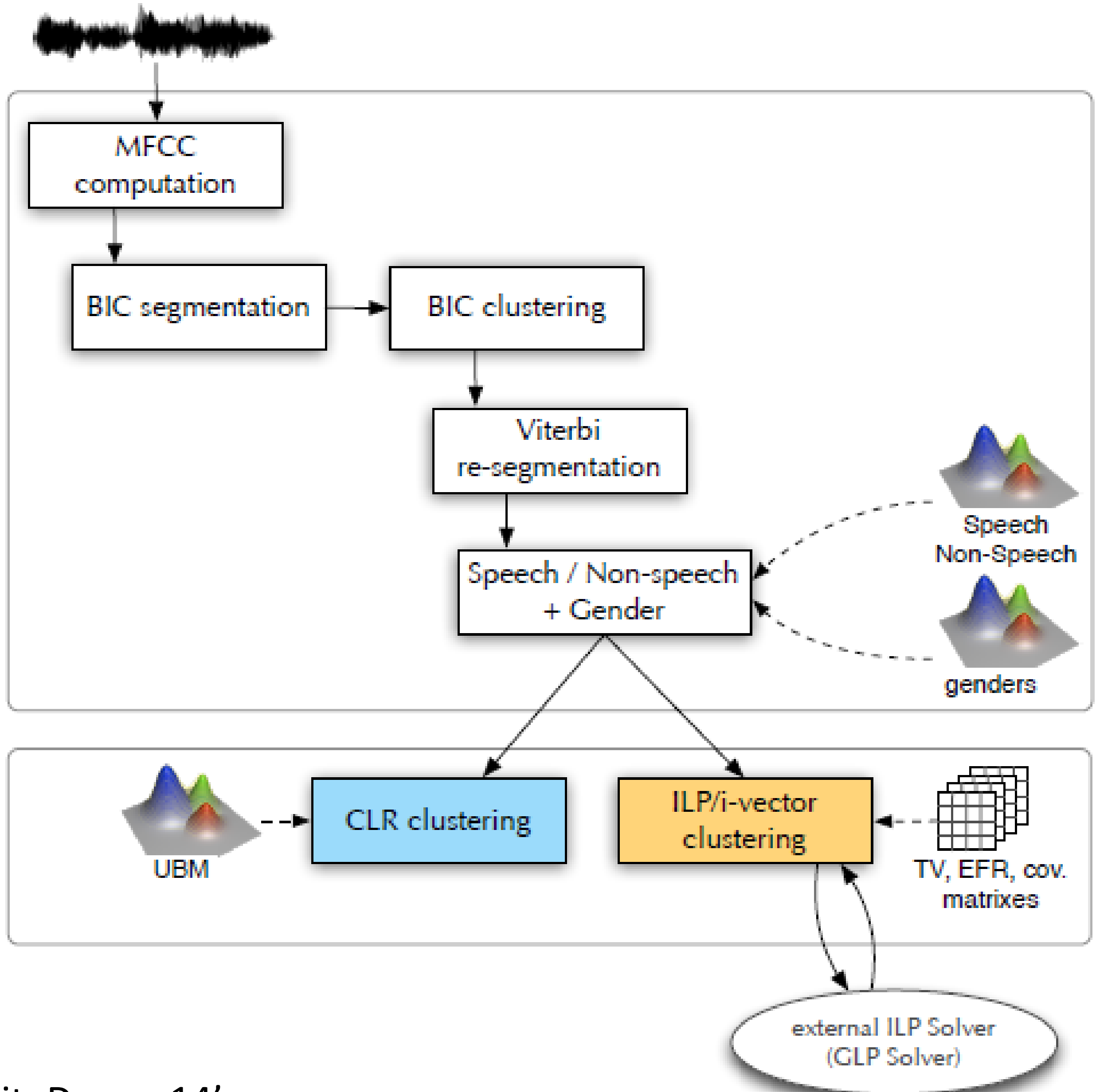


Image credit: Dupuy 14'

ILP: Results

Datasets

- REPERE 2012 French evaluation campaign
- 28 TV shows recorded from French TV

HAC/GMM		ILP/i-vector	
Threshold	*DER (%)*	*Distance δ*	*DER (%)*
0.0	19.55	75	17.01
-0.1	18.80	80	16.60
-0.2	19.76	85	15.94
-0.3	17.57	90	15.45
-0.4	17.69	95	15.45
-0.5	17.83	100	15.03
-0.6	17.70	**105**	**14.70**
-0.7	**16.22**	110	15.56
-0.8	17.26	115	15.46
-0.9	17.44	120	15.33
-1.0	18.29	125	16.18

G. Dupuy et al., “Recent Improvements on ILP-based Clustering for Broadcast News Speaker Diarization”, in Proc. *Speaker Odyssey*, 2014

Advanced Diarization Methods: **Clustering**

1. K-means
2. Integer Linear Programming (ILP)
3. **Fully Bayesian using Variational Bayes (VB)**

Bayesian Model Selection

- Given data X and model set $\{m\}$, optimal model is given by

$$\arg\max_m p(m|X) = \arg\max_m \frac{p(X|m)p(m)}{p(X)}$$

- Assuming $p(m)$ is uniform, maximize the marginal distribution, integrating over the parameters θ and hidden variables H :

$$p(X|m) = \int p(X,\theta,H|m)\,dH\,d\theta$$

- The marginal distribution is intractable
- Laplace approximation → BIC criterion

$$\log p(X|m)_{BIC} = \log p(X|\hat{\theta},m) - \tfrac{p}{2}\log N$$

Variational Bayes (VB)

Motivation

- Estimating $p(X|m)$ is useful for model selection but is usually intractable due to the integration on parameters θ and hidden variables H
- Maximizing the posterior distribution $p(\theta, H|X, m)$ is useful for optimizing the parameters of the model and finding optimal values for the hidden variable but is also usually intractable

F. Valente, "Variational Bayesian methods for audio indexing," Ph.D. dissertation, 2005

Variational Bayes: Method

- The posterior distribution $p(\theta, H|X, m)$ is approximated by the variational distribution $q(\theta, H) = q(\theta)q(H)$
- $F(\theta,H)$ is defined as the variational free energy:

$$F(\theta, H) = \int q(\theta, H)\log\frac{p(X, \theta, H|m)}{q(\theta, H)}dHd\theta$$

- For any distribution q:

$$\log p(X|m) = F(\theta, H) + \mathrm{KL}\left(q(\theta, H) \| p(\theta, H|X, m)\right)$$

- Variational learning aims at maximizing $F(\theta,H)$ which is a lower bound for the evidence $\log p(X|m)$

F. Valente, “Variational Bayesian methods for audio indexing,” Ph.D. dissertation, 2005

Variational Bayesian: EM Estimation

Maximizing free energy

Iterate:

1. Fix $q(\theta)$ and optimize $q(H)$:

$$\ln q(H) = E_\theta[\ln p(X, \theta, H)] + \text{const}$$

where $E_\theta[\cdot] = \int \cdot\, q(\theta) d\theta$

2. Fix $q(H)$ and optimize $q(\theta)$:

$$\ln q(\theta) = E_H[\ln p(X, \theta, H)] + \text{const}$$

- The constants ensure integration to 1
- Convergence is guarantied
- Free energy increases in every step

VB for Speaker Diarization

Define

- $H=\{H_t\}$ where H_t indicates the speaker identity of superframe t
- $\theta=\{\theta_s\}$ where θ_s indicates the speaker factors under an eigenvoice factor analysis model $M = m + Vy$

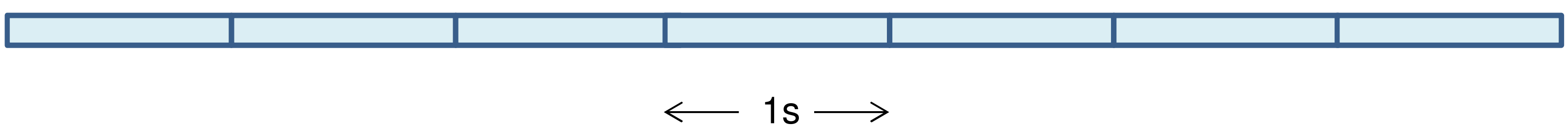
Alternate between estimating two types of posterior distributions until convergence

- Superframe–based posteriors for H (soft clustering)
- Speaker-based posteriors for θ

Kenny et al., "Diarization of Telephone Conversations using Factor Analysis", *IEEE Journal of Selected Topics in Signal Processing*, December 2010

Superframe Posteriors

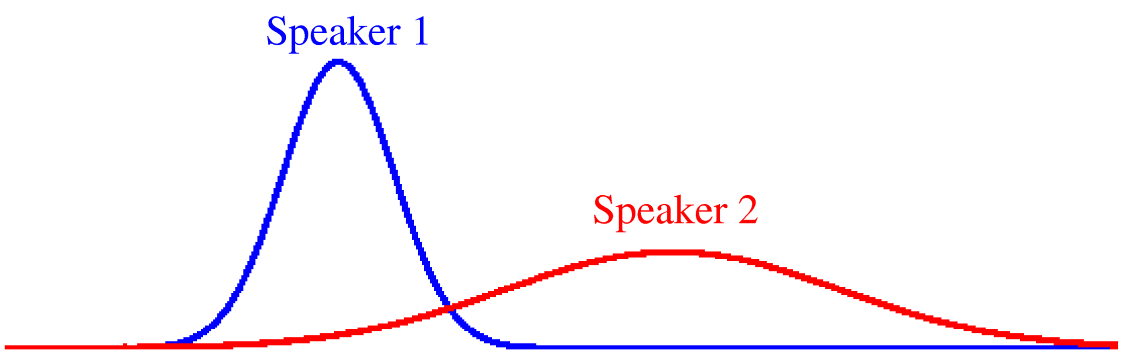
$$q_t^1, \ldots, q_t^S$$



q_t^s: Posterior probability that speaker s is talking at superframe t

$$q_t^s = p\left(H_t = s\right)$$

Speaker Posteriors (2 speakers)



Mean = Point estimate of speaker factors

Variance = Uncertainty

y_s : Speaker factors for speaker s

a_s : Mean of posterior distribution for speaker s

Λ_s^{-1} : Covariance of posterior distribution for speaker s

$$y_s \sim N\left(a_s, \Lambda_s^{-1}\right)$$

Image credit: Kenny 10'

VB for Speaker Diarization: **EM**

1. Given Superframe posteriors q_t^s estimate speaker posteriors (a_s , Λ_s^{-1})

$$\Lambda_s = I + V^* \left(\sum_{t=1}^{T} q_t^s N_t \right) \Sigma^{-1} V$$

$$a_s = \Lambda_s^{-1} V^* \Sigma^{-1} \left(\sum_{t=1}^{T} q_t^s \tilde{F}_t \right)$$

N_t : BW counts for superframe t

$\tilde{F}_t$: BW centralized (using UBM means) sums for superframe t

VB for Speaker Diarization: **EM** (2)

2. Given speaker posteriors and speaker priors π_s, estimate superframe posteriors

$$\tilde{q}_t^{\,s} = \pi_s \, p\left(x_t \middle| y_s = a_s\right) e^{-\frac{1}{2}\mathrm{tr}\left(V^* N_t \Sigma^{-1} V \Lambda_s^{-1}\right)}$$

$$q_t^{\,s} = \frac{\tilde{q}_t^{\,s}}{\sum_s \tilde{q}_t^{\,s}}$$

3. Given superframe posteriors estimate speaker priors

$$\pi_s = \tfrac{1}{T} \sum_{t=1}^{T} q_t^{\,s}$$

VB for Two-Speaker Diarization

1. Divide session into 1s superframes
2. Extract BW statistics from each superframe
3. EM estimation:
   - Initialize with random superframe posteriors
   - Iterate EM for estimating speaker and superframe posteriors
4. Construct GMMs for each speaker using hard decisions
5. Re-segment the data using Viterbi
6. Use resulting segments instead of superframes to run a 2^{nd} pass (steps 2-6)

Kenny et al., "Diarization of Telephone Conversations using Factor Analysis", *IEEE Journal of Selected Topics in Signal Processing*, December 2010

Experiments

Configuration

- Raw MFCCs (no feature warping)
- 1024 Gaussians
- 300 speaker factors

Results on NIST 08’ data

	DER (in %)	DER std (in %)
AHC +Viterbi resegmentation	6.8	12.3
AHC + Soft Viterbi resegmentation	3.5	8.0
VB	1.9	5.6
VB with multiple initializations	1.0	3.5

- Soft Viterbi resegmentation: use speaker posteriors instead of hard Viterbi decisions

Kenny et al., “Diarization of Telephone Conversations using Factor Analysis“, *IEEE Journal of Selected Topics in Signal Processing*, December 2010

VB: Choosing Number of Speakers

1. Apply VB for different number of speakers S and choose S that maximizes the evidence
2. Optimization of the number of speakers within the VB framework
   - Better than the latter approach [Valente 10']
3. Sticky HDP-HMM [Fox 11', Shum 13']
   - HDP: Hierarchical Dirichlet processes
   - HMM is used for temporal modeling

Outline

1. Introduction
2. Speech processing
3. Voice activity detection
4. Classic diarization methods
5. Speaker recognition
6. **Advanced diarization methods**
   - Geometry
   - Clustering
   - **Techniques**

Advanced Diarization Methods: **Techniques**

1. Exploiting a-priori acoustic information
2. Handling overlapping speech
3. Short sessions and online processing
4. Modeling speaker-turn dynamics

Advanced Diarization Methods: **Techniques**

1. **Exploiting a-priori acoustic information**
2. Handling overlapping speech
3. Short sessions and online processing
4. Modeling speaker-turn dynamics

Exploiting A-Priori Acoustic Information

Blind speaker diarization

- Speakers are unknown to the system
- No a-priori information on speaker space

A-priori knowledge

- Some speakers may be known in advance
- Broader acoustic information may be available (gender, SNR, accent, etc.)

H.A., "Speaker Diarization using A Priori Acoustic Information", in Proc. *Interspeech*, 2011

Tasks with Prior Information

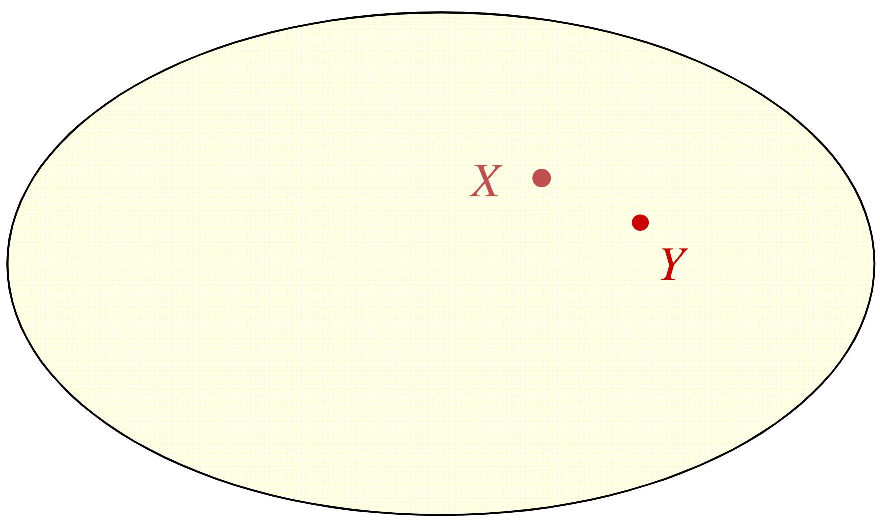
Two known speakers

- Standard approach: Train a GMM/HMM for each speaker and find Viterbi segmentation
- Problems: channel mismatch, inter-session variability
- Can we do better? Can we integrate blind-speaker diarization principles with the direct SID approach?

A mix of known and unknown speakers

- Call-centers: agent vs. customer
- Personal assistance system: user vs. other speakers
- BN: anchors, correspondents, politicians

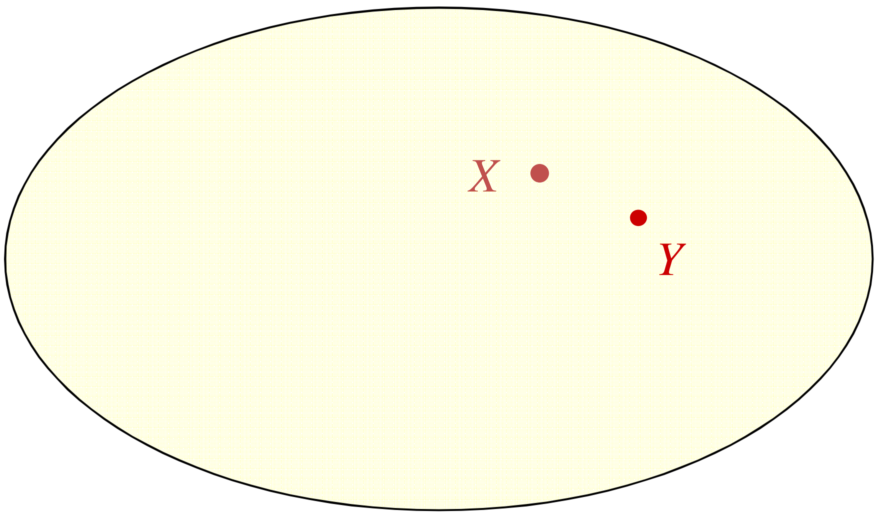
Are segments X and Y from the same speaker ?



High level feature space

Are segments X and Y from the same speaker ?

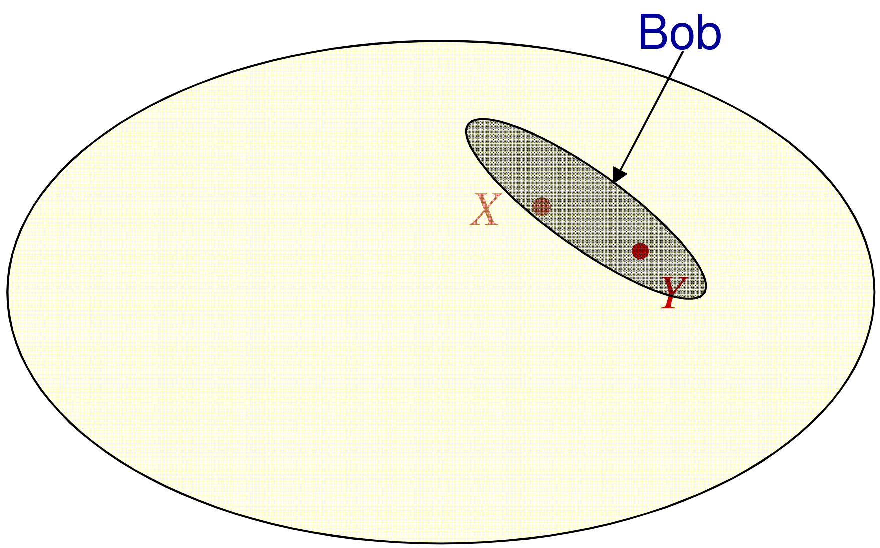
- In previous slides we have shown how to score the similarity between X *and* Y ("Geometry")



High level feature space

Are segments X and Y from the same speaker ?

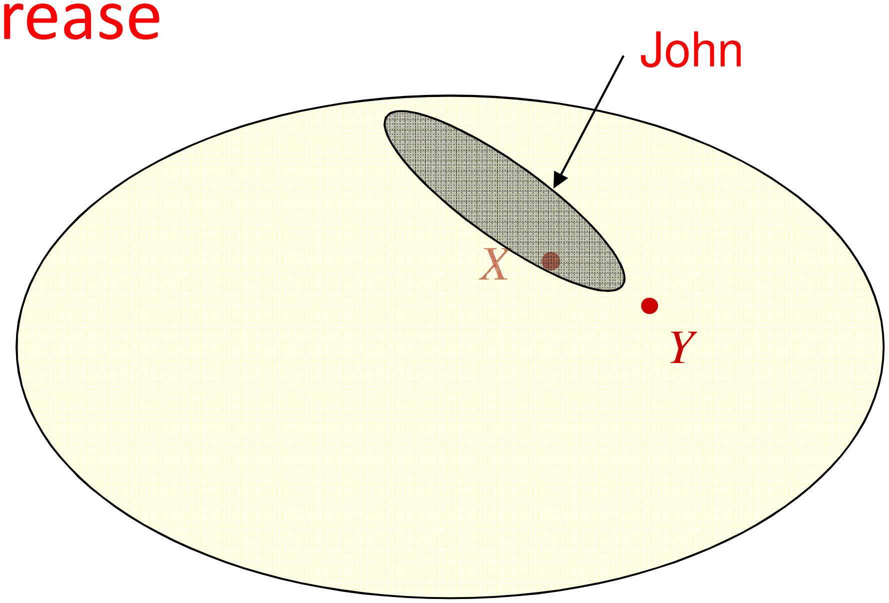
- In previous slides we have shown how to score the similarity between X *and* Y ("Geometry")
- Given an a priori speaker Bob, probability of match should increase



High level feature space

Are segments X and Y from the same speaker ?

- In previous slides we have shown how to score the similarity between X *and* Y ("Geometry")
- Given an a priori speaker John, probability of match should decrease



High level feature space

Outline of Method

Given a pair of segments *X* and *Y*

1. Score using acoustic similarity only is

$$f(X,Y) = \log \frac{p(Y|X)}{p(Y)}$$

2. Score with a priori information *Inf* is

$$f_{Inf}(X,Y) = \log \frac{p(Y|X,Inf)}{p(Y|Inf)}$$

3. Encode *Inf* in the feature domain: find transformation *T* for which

$$f_{Inf}(X,Y) = f(T(X),T(Y))$$

Scoring Acoustic Similarity

$$f(X,Y)=(x-u)^T\Sigma^{-1}(y-u)$$

with

- x and y are supervectors (or other high level features)
- u is the UBM supervector
- Σ is the inter-speaker intra-session covariance matrix

Integrating A Priori Information

1. Let $C_1,\ldots,C_k$ be a set of disjoint speakers
2. Assume supervectors of speaker C_i distribute normally with mean μ_i and covariance Λ (inter-session variability):

$$C_i \sim N(\mu_i, \Lambda)$$

3. Let C_0 be the complement of $\bigcup_i C_i$
   – approximated with the UBM: $C_0 \sim N(\mu_0, \Lambda_0)$
4. We obtain:

$$f_{Inf}(X,Y) = \log \frac{p(Y|X, Inf)}{p(Y|Inf)} = \log \frac{\sum_i p(C_i|X)\,p(Y|X, C_i)}{\sum_i p(C_i)\,p(Y|C_i)}$$

Integrating A Priori Information (2)

1. We assume that inter-session covariance matrix Λ is proportional to the intra-session covariance matrix Σ:

$$\Lambda = (\alpha - 1)\Sigma$$

2. It is shown in [H.A. 11'] that

$$f_{Inf} \cong (\hat{x} - u)^T \Sigma_\alpha^{-1} (y - u)$$

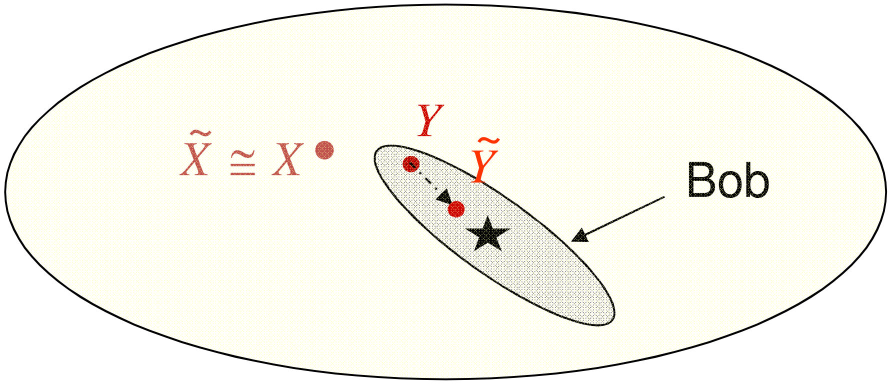
with $\hat{x} = \left(1 - \tfrac{1}{\alpha}\right)x + \tfrac{1}{\alpha}\sum_i p\left(\hat{C}_i \middle| x\right)\mu_i$

$$\Sigma_\alpha = \left(1 - \tfrac{1}{\alpha}\right)\Sigma$$

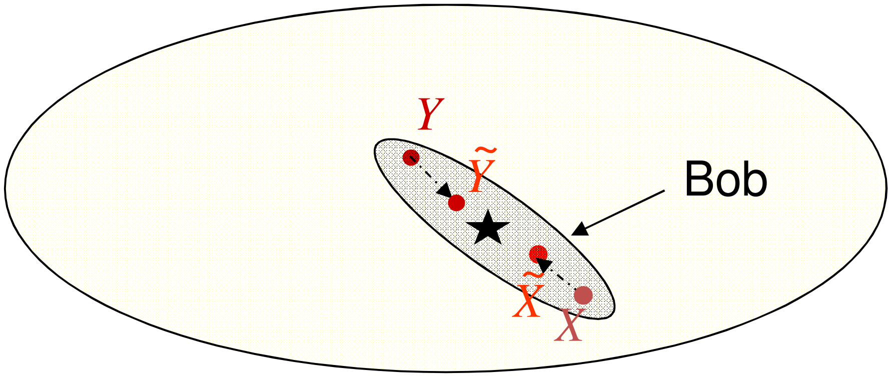
$$\hat{C}_i \sim N(\mu_i, \Lambda + \Sigma)$$

Illustration: A Single A-Priori Speaker

- Y is attracted to the center of Bob's distribution
- Distance between X and Y is increased



- Both X and Y are attracted to the center of Bob's distribution
- Distance between X and Y is decreased



Encoding the Information in the Feature Domain

- Instead of modifying the high-level features, we modify the low level features (MFCCs)
- Requirement for modified features $\tilde{X},\tilde{Y}$:

$$f_{Inf}\left(X,Y\right)\cong f\left(\tilde{X},\tilde{Y}\right)$$

- Solution [H.A. 11']:

  Use the fNAP [Vair 06'] method

$$\tilde{O}(t)=\left(1-\tfrac{1}{2\alpha}\right)O(t)+\tfrac{1}{2\alpha}\sum_i p\left(\hat{C}_i\middle|x(t)\right)\sum_m \gamma_m(t)\mu_{i,m}$$

- $x(t)$: supervector for the segment centered at time t
- $\gamma_m(t)$: UBM Gaussian m occupancy probability at frame t
- $\mu_{i,m}$: mean of the m-th Gaussian for speaker i

Experiments – Call Center

- Summed telephone conversations between a known agent and an unknown customer
- Agent GMMs are trained using 100 summed sessions (automatically diarized) per agent
- Test setup
  - 400 sessions
  - 25% shorter than 30 sec
  - Reference segmentation is errorful
  - Often a third (unlabeled) speaker exists

System	Baseline SER (in %)	Agent model SER (in %)
BIC-based	16.3	12.4
Supervector-based	14.1	10.2

Advanced Diarization Methods: **Techniques**

1. Exploiting a-priori acoustic information
2. **Handling overlapped speech**
   – HMM/GMM based [Boakye 11']
   – Convolutive Non-Negative Sparse Coding [Geiger 11']
3. Short sessions and online processing
4. Modeling speaker-turn dynamics

Handling Overlapped Speech

Motivation

- Particularly important for multiparty meetings
- Cause of missed speech errors
- Degrades the purity of speaker models

Framework

1. Detection
2. Exclusion from clustering stage
3. Labeling

HMM-based Overlapped Speech Detection

- HMM-GMM system with 3 classes:
  - Non-speech
  - Speech
  - Overlapped speech
- 3 states/class, 256 Gaussians
- HMMs are trained on ASR forced-aligned data

K. Boakye, et al., "Improved Overlapped Speech Handling for Speaker Diarization," in Proc. *Interspeech* 2011

Features for HMM-Based Detection

Features (in decreasing order of usefulness)

LPC residual energy, spectral flatness, RMS energy, harmonicity, modulation spectrogram features, MFCCs, harmonic energy ratio, diarization posterior entropy, kurtosis and zero-crossing rate

1. 50ms frames
2. Feature warping
3. PCA for decorrelation
4. Feature selection using KL-distance (selected)

K. Boakye, et al., "Improved Overlapped Speech Handling for Speaker Diarization," in Proc. *Interspeech,* 2011

Overlapped Speech Labeling

For each detected segment

1. Produce a speaker-ID score for each speaker
2. The two top-scoring speakers are selected for labeling
3. Exception:

   If

   Original segment label was associated with a third speaker

   then

   Third speaker and top scoring speaker are chosen

K. Boakye, et al., "Improved Overlapped Speech Handling for Speaker Diarization," in Proc. *Interspeech,* 2011

Overlapped Speech: Results

Data

- AMI Meeting Corpus (100 hrs)
- Single-channel far-field microphone signals
- Multi-site data
- Data contains roughly 15% overlapped speech

Overlapped speech detection

- Precision=55%, Recall=40%

Impact on DER

- 12% relative improvement (15% for oracle detection)
- Improvement mainly due to overlapped speech exclusion

K. Boakye, et al., "Improved Overlapped Speech Handling for Speaker Diarization," in Proc. *Interspeech,* 2011

Non-Negative Sparse Coding (CNSC)

A non-negative matrix $X \in R^{\geq 0}_{M \times N}$ is represented as:

$$X \approx WH$$

where $W \in R^{\geq 0}_{M \times K}$ and $H \in R^{\geq 0}_{K \times N}$ form the bases and base activations respectively

Optimization

$$(\hat{W}, \hat{H}) = \underset{W,H}{\operatorname{argmin}} \|X - WH\|_F^2 + \lambda \sum_{i,j} H_{i,j}$$

where λ controls the sparseness of the resulting representation

Shortcoming

NNSC fails to capture correlation between adjacent frames in X that is inherent in speech signals

J. T. Geiger, et al., "Speech overlap detection using convolutive non-negative sparse coding: New improvements and insights," in Proc. *EUSICPCO,* 2011

Convolutive Non-Negative Sparse Coding (CNSC)

- A non-negative matrix $X \in R_{M \times N}^{\geq 0}$ is represented as:

$$X \approx \sum_{p=0}^{p-1} W_p \overset{p\rightarrow}{H}$$

where $W_p \in R_{M \times K}^{\geq 0}$ and $H \in R_{K \times N}^{\geq 0}$ form the bases and base activations respectively

- P is the convolution range
- $\overset{p\rightarrow}{\cdot}$ is a column shift operator which shifts *p* columns of the matrix to the right. Vacated columns are filled with zeros

Optimization

- A non-convex optimization problem which is solved by an iterative optimization

J. T. Geiger, et al., "Speech overlap detection using convolutive non-negative sparse coding: New improvements and insights," in Proc. *EUSICPCO,* 2011

CNSC for Overlapped Speech Detection

- CNSC bases are learnt for individual speakers
- Interval of alleged overlapping speech is decomposed into speaker components

Base learning

- Features: spectral magnitude
- Learn CNSC base W for each speaker using pure speech
- Base patterns are concatenated to create a global basis W^G

Decomposition

- Spectral magnitude features are decomposed at the frame level with W^G kept fixed and H set to minimize the optimization criterion

J. T. Geiger, et al., "Speech overlap detection using convolutive non-negative sparse coding: New improvements and insights," in Proc. *EUSICPCO,* 2011

CNSC for Overlapped Speech Detection (2)

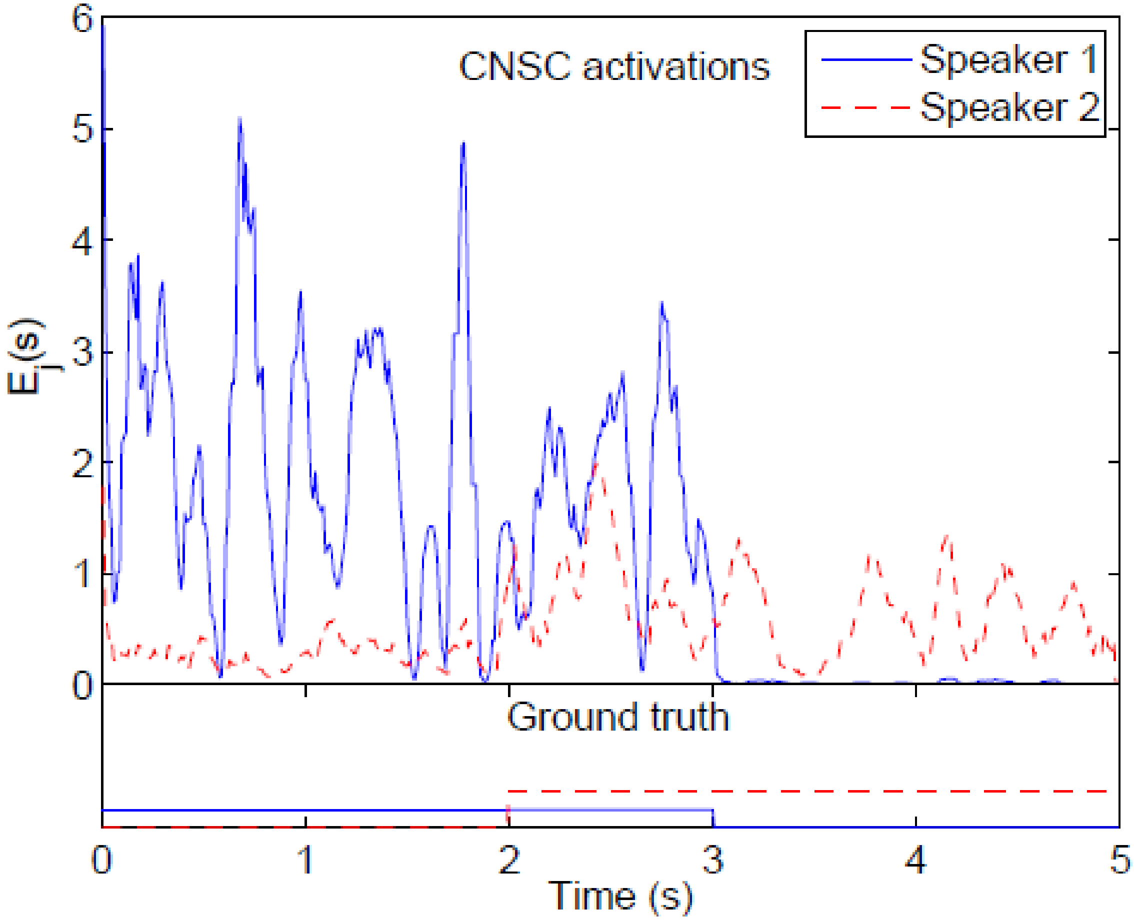


Image credit: Geiger 11’

Results: CNSC Compared to HMM/GMM

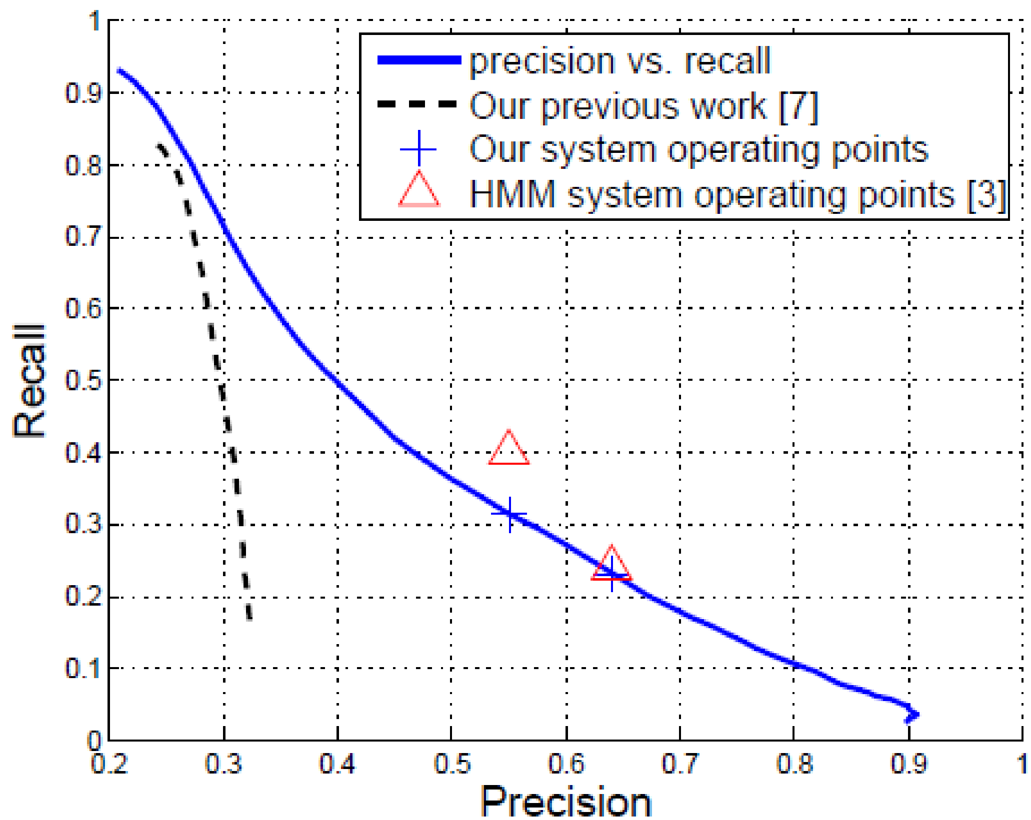


Image credit: Geiger 11'

Advanced Diarization Methods: **Techniques**

1. Exploiting a-priori acoustic information
2. Handling overlapped speech
3. **Short sessions and online processing**
4. Modeling speaker-turn dynamics

Short Sessions and Online Processing

Problem description

- Diarization is inherently **not** a sequential process
- Accuracy depends on session length
- Accuracy depends on amount of speech per speaker

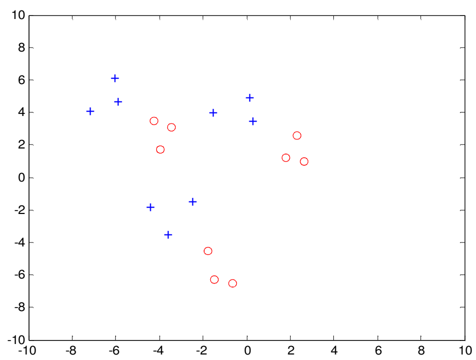
Reference System: Two-speaker Diarization [H.A. 2010]

1. Supervector parameterization of superframes
2. Intra-speaker variability compensation using NAP
3. PCA for soft classification of superframes into speakers
4. Viterbi-based segmentation
5. Viterbi resegmentation

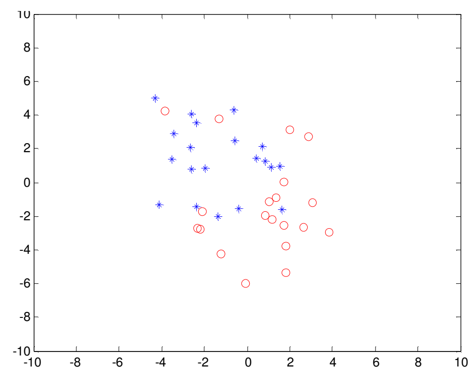
H. A. et al., "Online Two Speaker Diarization", in Proc. *Speaker Odyssey*, 2012

Offline System Overview

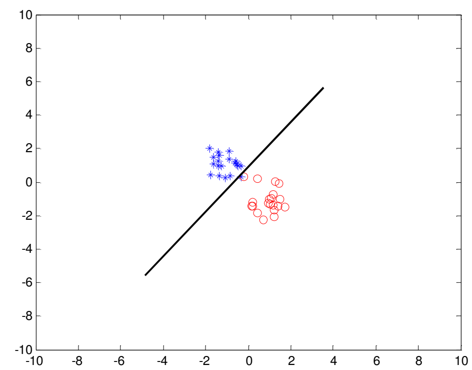
1. Speech is parameterized as a time-series of supervectors representing overlapping short segments (superframes)
   → PDF for each speaker is uni-modal
2. Intra-speaker variability is compensated



MFCC space



Supervector space



ISISV compensated supervector space

Offline System - Details

1. Audio parameterization
   - Train a session-dependent UBM
   - Define 1 second superframes (90% overlap)
   - Superframes → supervectors
2. ISISV modeling
   - Estimate ISISV covariance matrix from session
   - Compute NAP projection
   - Compensate supervectors

Offline System – Details (2)

3. Compute LLRs

Goal compute $\log p(x_t|s_i)$

x_t : supervector at time t

s_i : speaker i

- Compute covariance matrix of supervectors
- Find top eigenvector
- Project supervectors onto top eigenvector to obtain estimated LLRs

4. Viterbi segmentation

- Find an optimal smooth segmentation w.r.t. estimated LLRs

5. Viterbi re-segmentation in MFCC space

- Several iterations

Offline System - Shortcomings

Short sessions: accuracy degrades due to

- Insufficient data for estimation of UBM, ISISV, PCA
- Increased probability of an under-represented speaker

Inherently offline

- UBM estimation
- ISISV estimation
- PCA in supervector space
- Viterbi smoothing
- Viterbi re-segmentation

System Modification

UBM

- Train offline using dev-set

ISISV

- Estimate ISISV using dev-set
- Estimation is unsupervised (without speaker-turn labels)
- Estimation is done by pooling the difference supervector between each pair of adjacent superframes

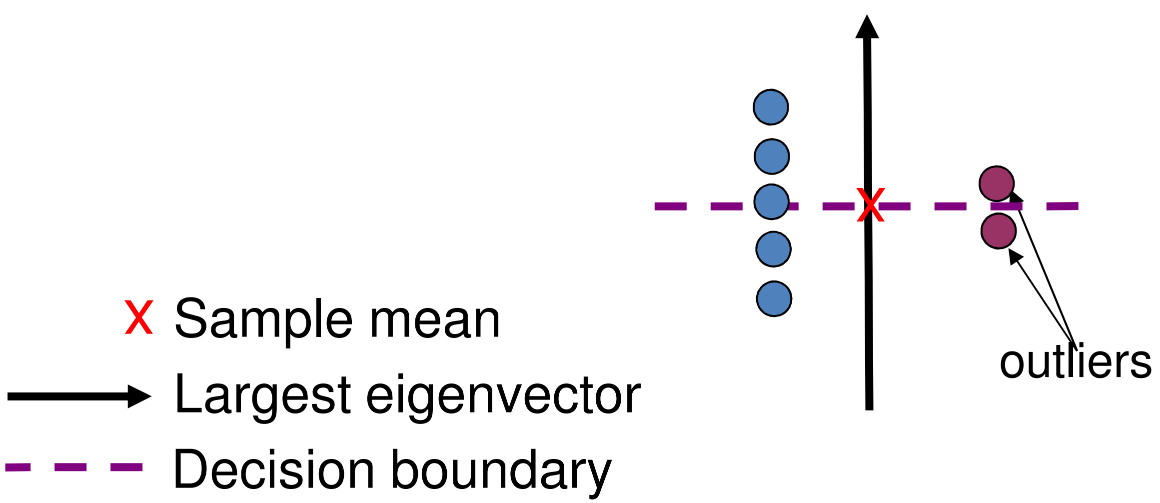
GMM order

- GMM order is reduced to 16 Gaussians

Robustness to Short Sessions

When speakers are highly unbalanced

- Top eigenvector may point to a wrong direction



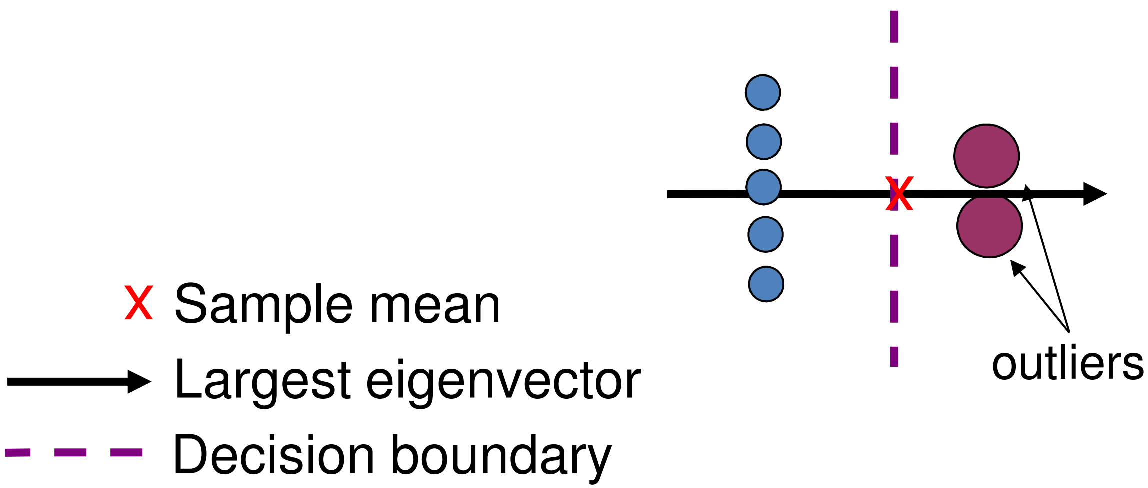
Robustness to Short Sessions (2)

When speakers are highly unbalanced

- Top eigenvector may point to a wrong direction

Outlier-emphasizing PCA

- Assign higher weights to outliers
- Outliers found by selecting top 10% supervectors with largest distance to the sample mean



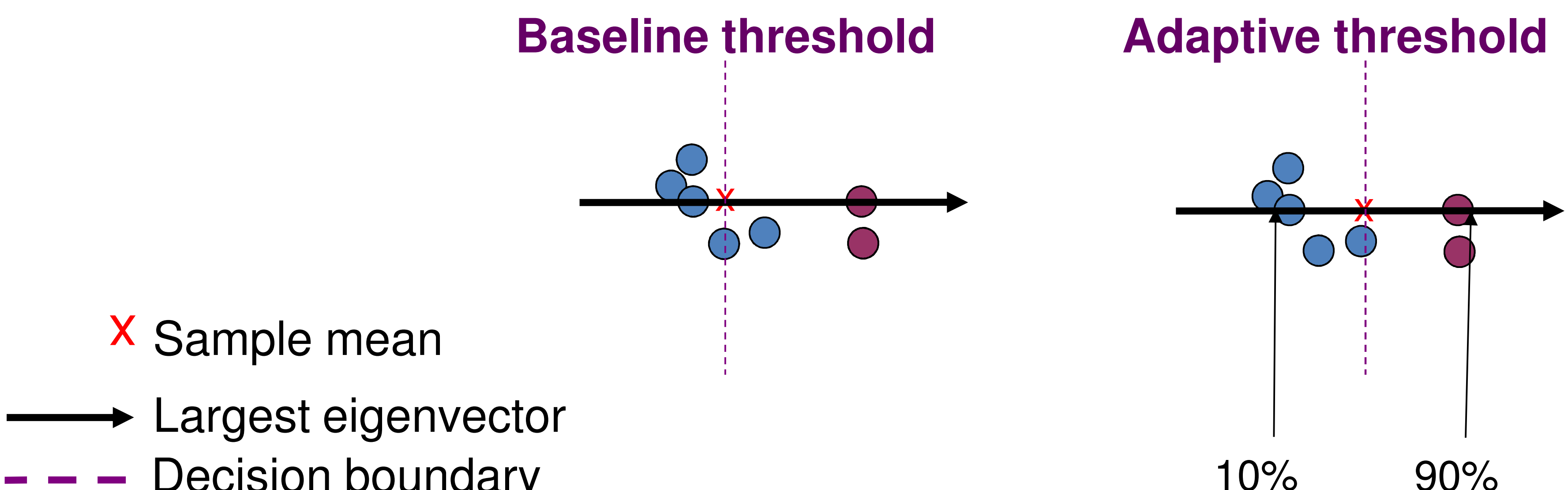
Robustness to Short Sessions (3)

When speakers are highly unbalanced

- Setting a proper classification threshold is a challenge

Adaptive threshold setting

- Averaging the values for the 10 and 90 percentiles of the projected supervectors

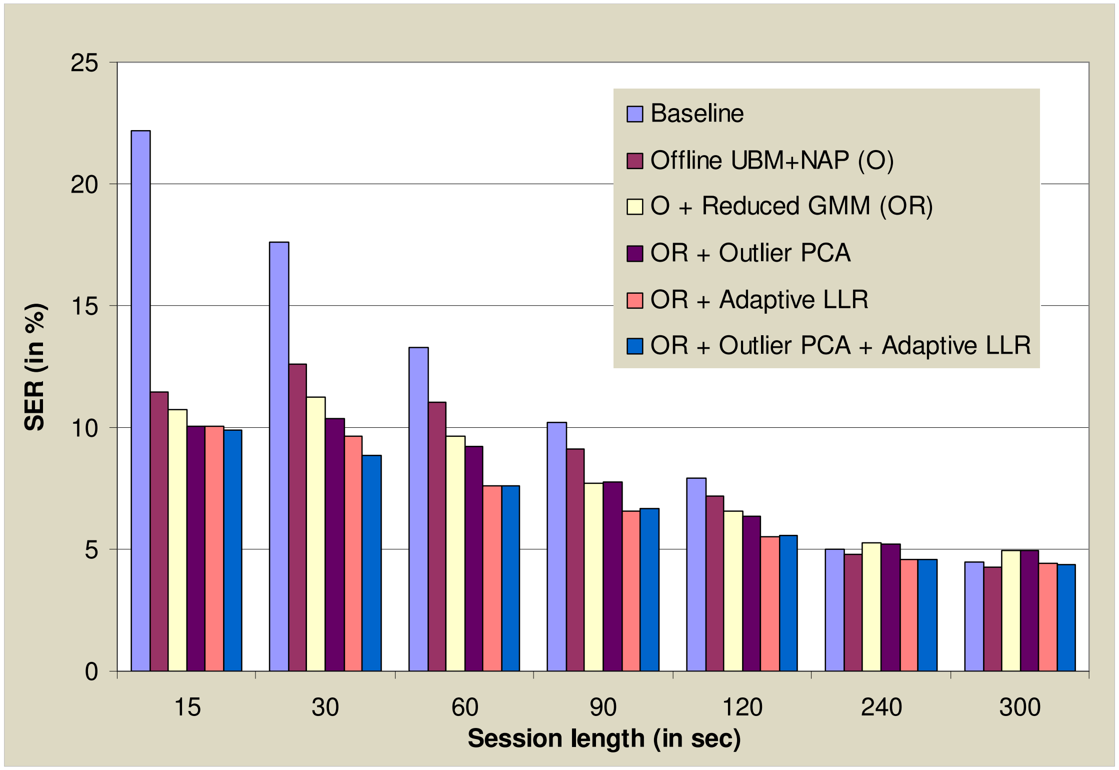


Datasets and Protocol

- NIST-2005 SRE
- Stereo phone calls artificially summed
- Ground-truth derived from ASR transcripts provided by NIST
- **No** forgiveness collar
- Short sessions with less than 3s per speaker are removed

  → Results for 15s sessions may be better than those for 30s sessions because sessions are more speaker balanced

Robustness to Short Sessions: **Results**



Session length (in s)	15	30	60	90	120	240	300
Baseline SER (in %)	22.2	17.6	13.3	10.2	7.9	5.0	4.4
Improved SER (in %)	9.9	8.8	7.6	6.7	5.6	4.6	4.4
Error reduction (in %)	55	50	43	35	30	9	0

Online Diarization : Scheme

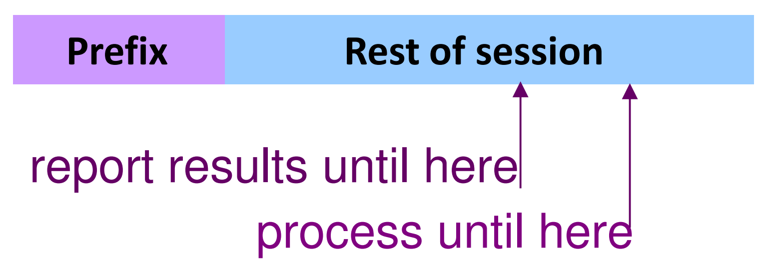
Prefix | **Rest of session**

Offline processing

- Prefix must be short
- Prefix length may be adaptive
- Outcome:
  - Speaker models
  - Other parameters

Online processing

- Initialize models from prefix processing
- Update models periodically
- Online processing with a delay



Online VAD

- VAD is energy-based
- Threshold is set according to energy histogram
- Viterbi is used for smoothing

Prefix | **Rest of session**

Offline processing

- Prefix length is 15s
- Energy histogram is computed
- Energy threshold is set using energy histogram
- VAD is computed for prefix

Online processing

- Histogram taken from prefix (updated periodically)
- Viterbi forward table is computed online
- Partial backtracking is used for decoding with a latency (0.1s)

Online Segmentation & Clustering

Online front-end

- MFCC extraction
- Supervectors extraction
- Intra-speaker variability compensation

Prefix	Rest of session

Offline processing

- PCA is computed
- Supervectors are projected
- Viterbi smoothing
- Viterbi re-segmentation

Online processing

- PCA statistics are accumulated
- PCA periodically recomputed
- Viterbi forward table is computed online
- Partial backtracking is used for decoding with a latency

Online Diarization: Results

SER as a function of the prefix length

- Delay parameter = 0.2s
- Sessions with an under-represented speaker in prefix (<3s) are excluded

Prefix (in sec)	15	30	45	60	90	120	300
SER (n %)	6.4	5.7	5.6	5.5	5.1	4.8	4.4

Conclusions

- Good accuracy for >3s per speaker in prefix
- Latency is 1.3s

Time Complexity

- 50xRT for the offline system
- 30xRT for the online system

Advanced Diarization Methods: **Techniques**

1. Exploiting a-priori acoustic information
2. Handling overlapped speech
3. Short sessions and online processing
4. **Modeling speaker-turn dynamics**

Modeling Speaker-Turn Dynamics

Motivation

- Different speakers have different roles
- Role affects
  - Distribution of speaker-turn durations
  - Interaction patterns

Goal

Use role detection for

- Setting a speaker dependent minimum duration constraint
- Training a social role n-gram model for use as prior information on speaker interaction patterns

F. Valente et al., “Speaker diarization of meetings based on speaker role n-gram models,” in Proc. *ICASSP*, 2011

Conversation Analysis

- Roles are stable behavioral patterns that speakers exhibit during the conversations and influence the way people take-turns in the conversation
- Types of roles:
  - Formal roles: the chairperson in a meeting or the moderator in a debate
  - Functional roles: the function that each speaker has in a spontaneous conversation, e.g., Information provider, Information seeker, Orienter, etc.
  - Social roles: the way each speaker relates to others in the discussion, e.g., Progatonist, Supporter, Gatekeeper, etc.

F. Valente et al., "Speaker diarization of meetings based on speaker role n-gram models," in Proc. *ICASSP*, 2011

Social Role Dataset

Corpus

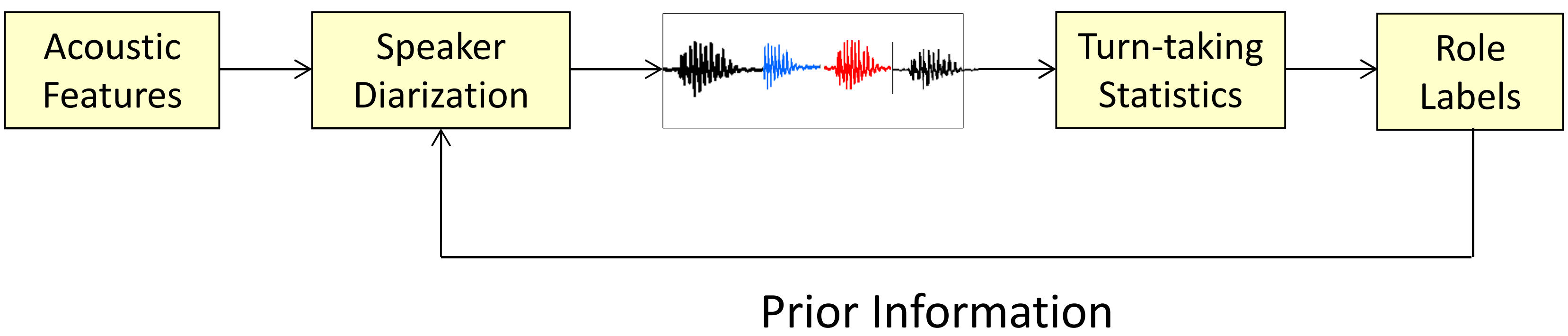
- AMI Meeting Corpus: scenario meetings subset
- 4 participants play the role of a design team

Roles

- Protagonist - drives the conversation, asserts its authority and assumes a personal perspective
- Supporter - shows a cooperative attitude demonstrating attention and acceptance and provides technical and relational support
- Neutral - passively accepts other speaker's ideas
- Gatekeeper - acts like group moderator

Outline

1. Speaker diarization is used to extract features for role recognition
2. Role recognition is performed
3. Recognized roles are used to improve diarization



F. Valente et al., “Speaker diarization of meetings based on speaker role n-gram models,” in Proc. *ICASSP*, 2011

Social Role Recognition

Feature extraction

- Acoustic features
- Turn-taking patterns
- Turn duration
- Total speaking time

A high-dimensional feature vector is created per speaker

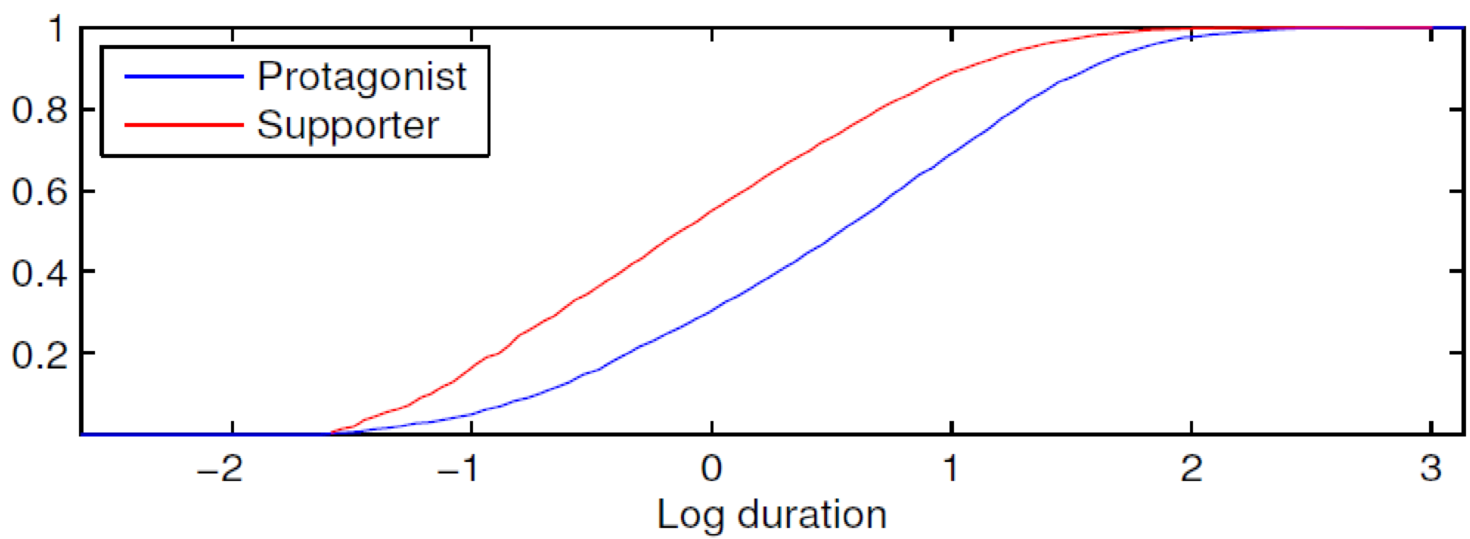
Classification

- A linear support vector machine (SVM)

Role-based Speaker Diarization

Speaker-turn length

- Distribution of turn length is role dependent
- Role-dependent minimal turn-length is estimated empirically



Role-based Speaker Diarization (2)

Speaker-turn N-Gram

- A Trigram model was found to yield lowest perplexity

	Unigram	Bigram	Trigram
Perplexity	4.4	3.5	2.9

- Trigrams are estimated from the training data
- Viterbi segmentation is modified to support the trigram model

Results

Datasets

- AMI meetings
- RT 07’, RT 09’ meetings

Dataset	Baseline SER (in %)	Role-based SER (in %)	Relative Error Reduction (in %)
AMI	17.6	14.8	16
RT 07’,09’	10.2	8.9	13

Summary

1. Voice activity detection
   - HMM, phoneme recognizer, segmental, DNNs
2. Classic diarization methods
   - BIC, AHC, Agglomerative SID clustering, Viterbi resegmentation
3. Advanced diarization methods
   - Geometry
     - High level features
     - Intra-speaker variability modeling
     - PLDA
     - PCA
     - Spectral clustering
     - Score normalization

Summary

3. Advanced diarization methods (Cont.)
   - Clustering
     - K-means
     - Integer Linear Programming (ILP)
     - Fully Bayesian using Variational Bayes (VB)
   - Techniques
     - Exploiting a-priori acoustic information
     - Handling overlapped speech
     - Short sessions and online processing
     - Modeling speaker-turn dynamics

References

VAD

- M. Sahidullah, et al., “Comparison of Speech Activity Detection Techniques for Speaker Recognition”, arXiv e-preprint, 2012
- P. Schwarz, "Phoneme Recognition based on Long Temporal Context”, PhD Thesis, 2009
- H. Aronowitz, “Segmental modeling for speech segmentation”, in Proc. *ICASSP* 2007
- G. Saon et al., “The IBM Speech Activity Detection System for the DARPA RATS Program”, in Proc. *Interspeech* ,2013

Score normalization

- C. Auckenthaler, et al. "Score Normalization for Text-independent Speaker Verification Systems," *Digital Signal Processing*, vol. 10 No 1-3, 2000
- H. Aronowitz et al., "Modeling Intra-Speaker Variability for Speaker Recognition", in Proc. *Interspeech* 2005
- S. Shum, et al., “Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification,” in Proc. *Speaker Odyssey*, 2010

References

Speaker recognition

- D. Reynolds, et al., “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Processing, vol. 10,* pp. 19–41, 2000
- H. Aronowitz et al., "Speaker indexing in audio archives using test utterance Gaussian mixture modeling", in Proc. *ICSLP*, 2004
- H. Aronowitz et al.,"Modeling Intra-Speaker Variability for Speaker Recognition", in Proc. *Interspeech* 2005
- W. Campbell, et al., “SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation”, in Proc. *ICASSP*, 2006
- A. O. Hatch et al. , “Within-Class Covariance Normalization for SVM-based Speaker Recognition”, in Proc. *ICSLP*, 2006

References

Speaker recognition (cont.)

- Kenny, P., et al., P. "A Study of Inter-Speaker Variability in Speaker Verification" IEEE TASLP, 2008
- O. Glembeck et al., “Comparison of scoring methods used in speaker recognition with Joint factor analysis”, in Proc. *ICASSP*, 2009
- N. Dehak, et al., “Front-end factor analysis for speaker verification”, in IEEE TASLP, 2011
- P. Kenny, “Bayesian Speaker Verification with Heavy-Tailed Priors”, in Speaker Odyssey, 2010.
- D.G. Romero, et al., “Analysis of i-vector Length Normalization in Speaker Recognition Systems “, in Proc. *Interspeech*, 2011

References

Classic Diarization Methods

- H. Gish, M. Siu, R. Rohlicek, “Segregation of speakers for speech Recognition and Speaker Identification,” in *Proc. ICASSP*, 1991
- M. Siegler, U. Jain, B. Raj and R. Stern, “Automatic Segmentation, Classification, and Clustering of Broadcast News Audio,” Proc. DARPA Speech Recognition Workshop, pp. 97-99, 1997
- S. S. Chen and P. S. Gopalakrishnam, “Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion,” in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, 1998
- D. Reynolds, E. Singer, B. Carlson, G. O’Leary, J. McLaughlin and M. Zissman, “Blind clustering of speech utterances based on speaker and language characteristics,” Proc. *ICSLP*, 1998
- M. Cettolo, M. Vescovi, and R. Rizzi, “Evaluation of BIC-based algorithms for audio segmentation,” *Computer Speech and, Language,* 2005

References

Classic Diarization Methods (Cont.)

- C. H. Wu and C. H. Hsieh, “Multiple Change-Point Audio Segmentation and Classification Using an MDL-based Gaussian Model,” IEEE Transactions on Audio, Speech and Language Processing, 2006
- C. Barras, et al., “Multi-Stage Speaker Diarization of Broadcast News,” IEEE TASLP, vol. 14, no. 5, pp. 1505–1512, 2006

References

Modern Diarization Methods

- F. Valente, “Variational Bayesian methods for audio indexing,” Ph.D. dissertation, 2005
- M. Collet, et al., “Speaker tracking by anchor models using speaker segment cluster information,” in Proc. *ICASSP*, 2006
- C. Vair et al.., “Channel factors compensation in model and feature domain for speaker recognition,” in Proc. *Odyssey*, 2006
- H. Aronowitz, “Trainable speaker diarization”, in *Proc*. *Interspeech*, 2007
- P. Smaragdis, “Convolutive Speech Bases and Their Application to Supervised Speech Separation,” *IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 1, pp.* 1–12, 2007
- F. Castaldo, et al., “Stream based speaker segmentation using speaker factors and eigenvoices,” in Proc. *ICASSP,* 2008

References

Modern Diarization Methods (Cont.)

- J. Žibert and F. Mihelič, “Novel Approaches to Speaker Clustering for Speaker Diarization in Audio Broadcast News Data”, Speech Recognition, InTech, 2008
- F. Valente, P. Motlícek, D. Vijayasenan, “Variational Bayesian speaker diarization of meeting recordings”, in *ICASSP*, 2010
- H. Aronowitz, "Unsupervised Compensation of Intra-Session Intra-Speaker Variability for Speaker Diarization", in Proc. *Speaker Odyssey,* 2010
- P. Kenny, D. Reynolds, and F. Castaldo, “Diarization of Telephone Conversations using Factor Analysis”, in IEEE *Journal of Selected Topics in Signal Processing*, 2010
- F. Valente, D. Vijayasenan, and P. Motlicek, “Speaker diarization of meetings based on speaker role n-gram models,” in Proc. *ICASSP*, 2011
- S. Shum, "Unsupervised Methods for Speaker Diarization," S.M. Thesis, MIT Department of Electrical Engineering and Computer Science, 2011

References

Modern Diarization Methods (Cont.)

- T. Stafylakis and V. Katsouros, “A Review of Recent Advances in Speaker Diarization with Bayesian Methods”, Speech and Language Technologies, InTech, 2011.
- H. Aronowitz, "Speaker Diarization using A Priori Acoustic Information", in Proc. *Interspeech*, 2011
- J. Silovsky, J. Prazak, P. Cerva, J. Zdansky, J. Nouza,” PLDA-based Clustering for Speaker Diarization of Broadcast Streams", in Proc. *Interspeech*, 2011
- E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, “A sticky HDP-HMM with application to speaker diarization,” *Ann. Appl. Statist., vol. 5, no. 2A, pp. 1020–1056, Jun.* 2011
- H. Aronowitz, Y. Solewicz, O. Toledo-Ronen, "Online Two Speaker Diarization", in Proc. *Speaker Odyssey*, 2012

References

Modern Diarization Methods (Cont.)

- S. Shum, N. Dehak, and J. Glass, “On the Use of Spectral and Iterative Methods for Speaker Diarization,” in Proc. *Interspeech*, 2012
- G. Dupuy, M. Rouvier, S. Meignier, and Y. Esteve, “i-Vectors and ILP Clustering Adapted to Cross-Show Speaker Diarization,” in Proc. *Interspeech*, 2012
- S. Shum, N. Dehak, R. Dehak and J. R. Glass, “Unsupervised Methods for Speaker Diarization: An Integrated and Iterative Approach”, in *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 21, no. 10, 2013