Intra-class variability modeling for speech processing

Hagai Aronowitz
IBM Haifa Research Lab

Global Autonomous Language Exploitation (GALE)

- Largest ever DARPA funded project.
- Input languages: Arabic, Mandarin, English.
- Input resources: Broadcast TV, Radio, telephone conversations, web logs, newsgroups.
- Output: English text + information extraction and question answering.
- Technologies: Transcription, translation, information extraction, question answering.
- Budget: 200M $ for 5 years (2005-2010).
- Participants: **IBM**, **BBN**, **SRI**, ICSI, CMU, JHU, LIMSI, Cambridge, Stanford, UMD, UWS…

2

Outline

1. Introduction to speech classification.
2. Approximated cross entropy for GMM scoring.
3. Intra-speaker inter-session variability modeling.
4. Mapping speech segments into segment space.
5. Modeling in segment space.
6. More applications.

3

Outline

1. Introduction to speech classification.
2. Approximated cross entropy for GMM scoring.
3. Intra-speaker inter-session variability modeling.
4. Mapping speech segments into segment space.
5. Modeling in segment space.
6. More applications.

4

Speech Classification

Classify a segment of speech

- Speaker recognition.
- Language identification.
- VAD (voice activity detection).
- Speaker change detection.
- Speaker diarization ("who spoke when").
- Emotion recognition.
- Gender identification.
- Channel classification.

5

Speech classification (2)

- This presentation will focus on **speaker recognition**.
- Results can be generalized to other classification tasks.
  - Language identification [Noor&Aronowitz06]
  - VAD [Aronowitz06]
  - Speaker change detection [Aronowitz07]
  - Speaker diarization [Aronowitz07]
  - Channel classification [Aronowitz07]

6

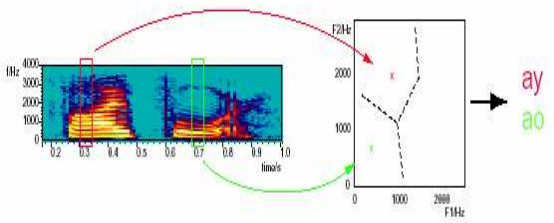
Classification

Given training instances from class + and class -, automatically classify unlabeled test instances.

Classic classification framework

- Represent instances in instance-space.
- Learn a classifier in instance-space
  - SVMs
  - NNs
  - Bayesian classifiers.
  - …

Classification (2)



Dan Ellis, Lecture notes

Classification of speech segments

Complex representation

- Time series
- Variable length
- High dimensional

→ How should instances we represented?

→ How can we learn a classifier?

Usually, the **classic classification framework** is partially **abandoned**.

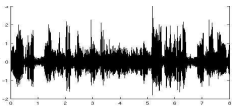
Our goal: stick to the **classic classification framework**.

Speaker Recognition

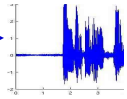
Definitions

S : target speaker

X : training session for speaker S

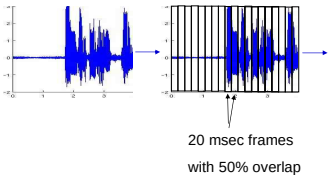


Y : test session



Goal: compute $\Pr(S|Y) = \frac{\Pr(Y|S)\Pr(S)}{\Pr(Y)} \propto \Pr(Y|S)$

Feature Extraction



$(x_1,\ldots,x_T)$

$x_i \in R^{26}$

T = segment length

A speech segment is represented as a sequence of 26-dimenional feature vectors (frames).

Speaker Recognition (2)

Definitions

S : target speaker

X : training session for speaker S

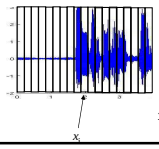
Y : test session

GMM based algorithm [Reynolds95]

(1) $\Pr(Y|S) \underset{①}{\cong} \Pr(Y|X) = \Pr(y_1,\ldots,y_T|X) \underset{②}{\cong} \prod_{t=1}^{T} \Pr(y_t|X) \underset{③}{\cong} \prod_{t=1}^{T} \Pr(Y_t|G_X)$

Assumptions

1. Training data represents speaker.
2. Test frames independence.
3. Training frames independence + Gaussian-Mixture-Model assumption.



Gaussian Mixture Models (GMMs)

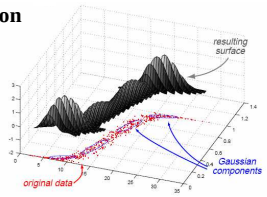
$\Pr(y_1,\ldots,y_T|X)=?$

Frame independence assumption

(2) $\Pr(y_1,\ldots,y_T|X)=\prod_{t=1}^{T}\Pr(y_t|X)$

Gaussian mixture modeling

(3) $\Pr(y_t|X)=\sum_{g=1}^{G}w_g N(y_t;\mu_g,\Sigma_g)$

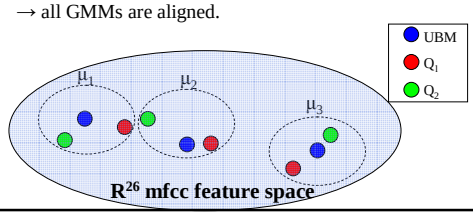


G : GMM order
w_g : The weight of the gth Gaussian of G_X
μ_g : The mean vector of the gth Gaussian of G_X
Σ_g : The (diagonal) covariance matrix of the gth Gaussian of G_X

13

Universal Background Model (UBM)

1. Train a universal GMM on data collected from many non-target speakers.
2. Train a GMM for target-speaker by applying MAP (Maximum A-Posteriori) adaptation to the UBM.
   → all GMMs are aligned.



14

GMM based speaker recognition

Train

- Train a universal GMM (UBM) using the EM algorithm.
- For every target speaker S:
  - Train a GMM G_S by applying MAP-adaptation.

Test : $\Pr(Y|S)=?$

- Compute $\Pr(Y|S)\cong\prod_{t=1}^{T}\Pr(y_t|G_S)=\prod_{t=1}^{T}\sum_{g=1}^{G}w_k^S N(y_t;\mu_g^S,\Sigma_g^S)$

15

Other algorithms using GMMs

- Language identification.
- VAD (voice activity detection).
- Speaker change detection.
- Speaker diarization ("who spoke when").
- Emotion recognition.
- Gender identification.
- Channel classification.

All use basically the same algorithm.

16

GMMs - issues

1. Frame independence assumption – invalid
   - Otherwise, GMMs trained on sufficiently long sessions of a single speaker should converge to each other.
   - What about channel, emotion, lexical variability, aging ?
2. Can we relax the independence assumption?
3. Can we model speaker variability without needing many training sessions of the same speaker?
4. GMM scoring is inefficient – O($SDGT$).
5. GMM scoring does not support indexing.

17

Outline

1. Introduction to speech classification.
2. Approximated cross entropy for GMM scoring.
3. Intra-speaker inter-session variability modeling.
4. Mapping speech segments into segment space.
5. Modeling in segment space.
6. More applications.

18

GMM approximation

Given: - $\boldsymbol{Y}$ – test segment
- $\boldsymbol{X}$ – target segment
- GMM $\boldsymbol{P}$ trained for Y
- GMM $\boldsymbol{Q}$ trained for X

Compute $\Pr(\boldsymbol{Y}|\boldsymbol{Q})$ using GMMs $\boldsymbol{P}$ and $\boldsymbol{Q}$ only:

- $\Pr(\boldsymbol{Y}|\boldsymbol{Q}) = f(\boldsymbol{P}, \boldsymbol{Q}) = ?$

Motivation: Efficient speaker retrieval & identification.

Efficient speaker clustering...

GMM approximation (2)

(1) $$\frac{1}{T}\log\Pr(Y|Q) = \frac{1}{T}\sum_{t=1}^{T}\log\Pr(y_t|Q) \xrightarrow[T\to\infty]{*} \int_x \Pr(x|P)\log(\Pr(x|Q))dx$$

$$= -H(P,Q)$$

= negative cross entropy

* Assuming Y was drawn from distribution P

For a finite T we approximate:

(2) $$\frac{1}{T}\log\Pr(Y|Q) \cong \int_x \Pr(x|P)\log(\Pr(x|Q))dx$$

Computing $\int_x \Pr(x|P)\log(\Pr(x|Q))dx$

Definitions

P_i , Q_i : The ith/ j^{th} Gaussian of distribution P/Q.

w_i^P, w_j^Q: The weight of the ith/ j^{th} Gaussian of P/Q.

G : Number of Gaussians in GMM P

(3) $$\int_x \Pr(x|P)\log(\Pr(x|Q))dx = \sum_{i=1}^{G} w_i^P \int_x \Pr(x|P_i)\log(\Pr(x|Q))dx$$

Approximating $\int_x \Pr(x|P_i)\log(\Pr(x|Q))dx$

1. Matching based lower bound [Aronowitz04]
2. Unscented-transform based approximation [Goldberger&Aronowitz05].
3. Others options in [Hershey07].

Matching based approximation

(4) $$\forall j: \int_x \Pr(x|P_i)\log(\Pr(x|Q))dx \geq \int_x \Pr(x|P_i)\log(\Pr(x|Q_j))dx$$

$$= \log w_j^Q - \sum_{d=1}^{D}\frac{(\mu_{i,d}^P - \mu_{j,d}^Q)^2}{2\sigma_{j,d}^{Q\,2}} - \sum_{d=1}^{D}\log\sigma_{j,d}^Q - \frac{1}{2}\sum_{d=1}^{D}\left(\frac{\sigma_{i,d}^P}{\sigma_{j,d}^Q}\right)^2 - \frac{D}{2}\log 2\pi$$

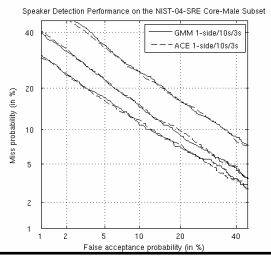
$\Rightarrow$ $$\int_x \Pr(x|P)\log(\Pr(x|Q))dx \geq$$

(5) $$\sum_{i=1}^{G} w_i^P \max_j\left\{\log w_j^Q - \sum_{d=1}^{D}\frac{(\mu_{i,d}^P - \mu_{j,d}^Q)^2}{2\sigma_{j,d}^{Q\,2}} - \sum_{d=1}^{D}\log\sigma_{j,d}^Q - \frac{1}{2}\sum_{d=1}^{D}\left(\frac{\sigma_{i,d}^P}{\sigma_{j,d}^Q}\right)^2\right\} + C$$

Matching based approximation (2)

Assuming weights and covariance matrices are speaker independent (+some approximation):

(6) $$\int_x \Pr(x|P)\log(\Pr(x|Q))dx \cong C - \sum_{i=1}^{G}\sum_{d=1}^{D} w_i \frac{(\mu_{i,d}^P - \mu_{j,d}^Q)^2}{2\sigma_{i,d}^2}$$



Outline

1. Introduction to speech classification.
2. Approximated cross entropy for GMM scoring.
3. Intra-speaker inter-session variability modeling.
4. Mapping speech segments into segment space.
5. Modeling in segment space.
6. More applications.

Motivation

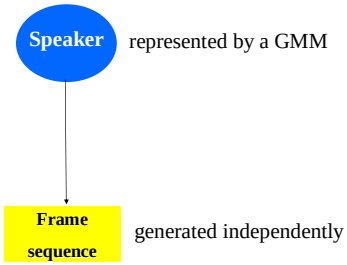
The classic GMM algorithm lacks explicit modeling of intra-speaker inter-session variability:

- Changing speaker characteristics – stress, emotion, aging.
- Noise
- Channel
- Language

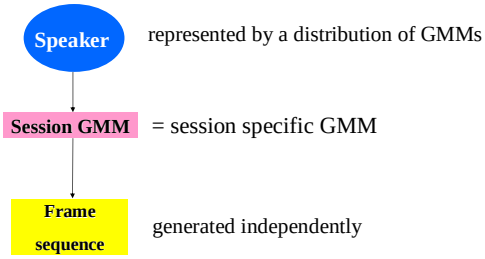
The frame independent assumption does not hold in these cases!

(1) $$\Pr(y_1,\ldots,y_T|S) \neq \prod_{t=1}^{T} \Pr(y_t|S)$$

Classic Generative Model



New Generative Model



Relaxing the independence assumption

(2) $$\Pr(Y|S) = \prod_{t=1}^{T} \Pr(y_t|S, f(Y)) = \prod_{t=1}^{T} \Pr(y_t|G_{S,f(Y)})$$

$f(Y)$ - session dependent attributes

- **stress, emotion, noise, channel, etc.**

$G_{S,f(Y)}$ - speaker and session dependent **pdf**

But

➢ session dependent attributes and $G_{S,f(Y)}$ are unknown

Solution: marginalization

(3) $$\Pr(Y|S) = \int_F \Pr(Y|G_{S,F}) \Pr(G_{S,F}|S) dF$$

Relaxing the independence assumption (2)

Solution: marginalization

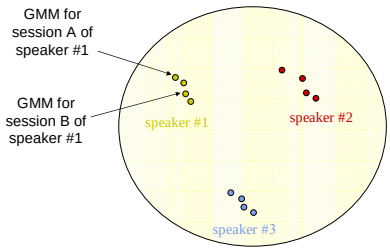
(3) $$\Pr(Y|S) = \int_F \Pr(Y|G_{S,F}) \Pr(G_{S,F}|S) dF$$

We assume that $\Pr(Y|G_{S,F})$ as a function of $G_{S,F}$ is concentrated sharply around the mode.

→

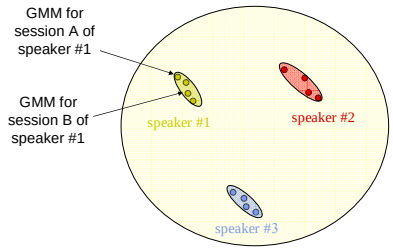
(4) $$\Pr(Y|S) = \Pr(Y) \Pr(\tilde{G}_{S,F}|S) \quad : \quad \tilde{G}_{S,F} = \underset{G_{S,F}}{\operatorname{argmax}} \Pr(Y|G_{S,F})$$

Session-GMM space



Session-GMM space

Session-GMM space



Session-GMM space

Prior distribution of GMMs

We assume all GMMs share a global covariance matrix:

$$\text{GMM} \leftrightarrow (\bar{\mu}_1,\ldots,\bar{\mu}_{N_g}, w_1,\ldots,w_{N_g}) \leftrightarrow (\hat{\mu},\hat{w})$$

Conditional independence of means and weights

(5) $\Pr(G=\hat{\mu},\hat{w}|S) \cong \Pr(\hat{\mu}|S)\Pr(\hat{w}|S)$

Modeling Pr($\hat{w}$|S)

We assume a multinomial distribution:

(6) $\Pr(\hat{w}\,|\,S) = \frac{n!}{(\hat{w}_1 n)!\cdots(\hat{w}_{N_g} n)!}\tilde{w}_{s,1}^{\hat{w}_1 n}\cdots\tilde{w}_{s,N_g}^{\hat{w}_{N_g} n}$

Prior distribution of GMMs (2)

Modeling Pr($\hat{\mu}$|S)

We assume a multivariate normal distribution:

(7) $\Pr(\hat{\mu}\,|\,S) = N(\tilde{\mu}_s, \tilde{\Sigma}_s)$

Assumption:

covariance matrix $\tilde{\Sigma}_s$ is speaker independent.

$\tilde{\Sigma}_s = \tilde{\Sigma}$

Training the prior distribution

Training $\tilde{w}_s$

1. Fit GMMs to the training session(s) of speaker.
2. Concatenate all GMM weights to super-vectors.
3. Compute average super-vector→ $\tilde{w}^s$

Training $\tilde{\mu}^s$

1. Fit GMMs to the training session(s) of speaker.
2. Concatenate all GMM means to super-vectors.
3. Compute average super-vector→ $\tilde{\mu}^s$

Training the prior distribution (2)

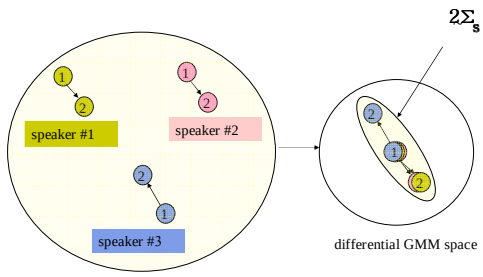
Training $\tilde{\Sigma}$

1. Fit separate GMMs to different sessions of a same speaker.
2. For every GMM g, concatenate all means → $\tilde{\mu}^{s,g}$
3. For all pairs of same speaker sessions compute:

   $\delta = \tilde{\mu}_{s,g_1} - \tilde{\mu}_{s,g_2}$
4. Estimate covariance matrix from δ vectors.

Training the prior distribution

Training covariance matrix $\tilde{\Sigma}_s$



GMM space

Computing Pr(Y | S)

1. Fit a GMM to test session Y
2. Concatenate all GMM weights → w
3. Concatenate all GMM means → μ
4. Compute

$$\Pr(Y \mid S) = \Pr(\mu \mid S)\Pr(w \mid S) = \Pr\left(N\left(\mu; \tilde{\mu}_s, \tilde{\Sigma}\right)\right) \frac{n!}{(w_1 n)! \cdots (w_{N_g} n)!} \tilde{w}_{s,1}^{w_1 n} \cdots \tilde{w}_{s,N_g}^{w_{N_g} n}$$

37

Estimating the covariance matrix

A typical dimension of $\tilde{\Sigma}_s$ is 50,000*50,000.

→ How can we estimate a full covariance matrix ?

Solution #1 (10% EER reduction)

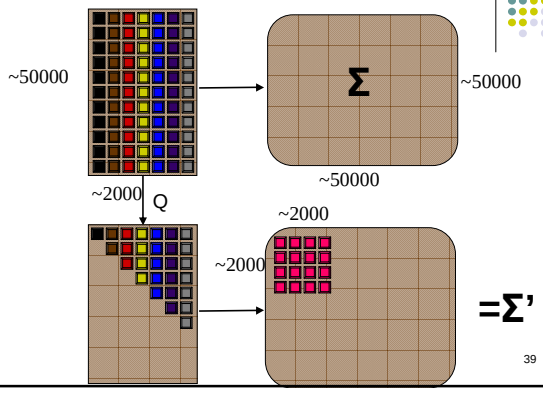
Assume Σ is diagonal.

Solution #2 (38% EER reduction)

Estimate Σ robustly using PCA + regularization.

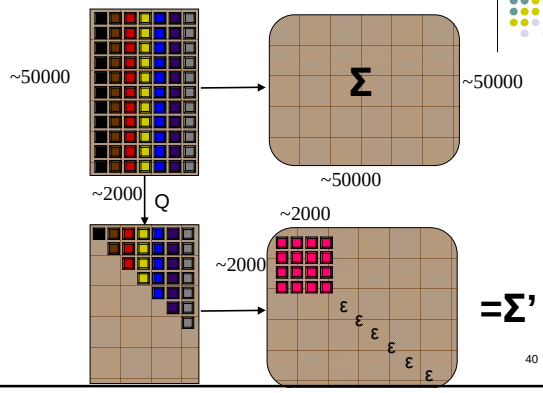
38

Solution #2 : 38% EER reduction



39

Solution #2 : 38% EER reduction



40

Evaluation data set

$\tilde{\Sigma}_s$ is estimated from the NIST-2006-SRE.

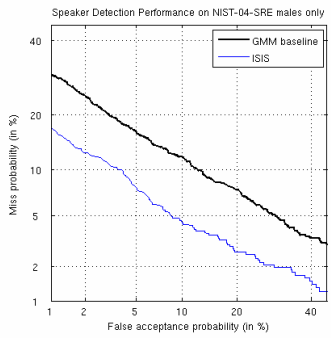
Evaluation: NIST-2004-SRE.

Experimental Setup

- ETSI MFCC (13-cep + 13-delta-cep).
- Energy based voice activity detector.
- CMS.
- 2048 Gaussians.
- Target models are adapted from GI-UBM.
- ZT-norm.

41

Results on the NIST-2004 Eval



42

Summary

- We introduced a novel model for generation of speech by speakers.
- We model intra-speaker inter-session variability by rotating the GMM super-vector space in a way that most variability is contained in a low-dimensional subspace.
- EER was reduced by 38% on the NIST-2004 SRE.
- Using kernel-PCA EER was reduced by 44%.

43

Outline

1. Introduction to speech classification.
2. Approximated cross entropy for GMM scoring.
3. Intra-speaker inter-session variability modeling.
4. Mapping speech segments into segment space.
5. Modeling in segment space.
6. More applications.

44

Mapping speech segments into segment space

Given an audio classification task

- Identify basic components (sessions , segments)
- Map training and test data into session/segment space
- Model each class in session/segment space
- Classify test data in session/segment space

45

Mapping techniques

- Anchor modeling projection [Sturim et al., 2001]
- GMM supervectors [Aronowitz et al. 2004]
- MLLR transofrms [Stolcke et al., 2005]
- Kernel-PCA [Aronowitz 2007]

46

Anchor modeling projection

Given: anchor models $\lambda_1,\ldots,\lambda_n$ and session $X= x_1,\ldots,x_F$

Projection: $X \rightarrow \left[\hat{s}(X|\lambda_1),\ldots,\hat{s}(X|\lambda_n)\right]$

$\hat{s}(X|\lambda_i) = \frac{1}{F}\log\frac{\Pr(X|\lambda_i)}{\Pr(X|\lambda_{UBM})}$ = average normalized log-likelihood

- Speaker indexing [Sturim et al., 2001]
- Intersession variability modeling in projected space [Collet et al., 2005]
- Speaker clustering [Reynolds et al., 2004]
- Speaker segmentation [Collet et al., 2006]
- Language identification [Noor and Aronowitz, 2006]

47

Kernel-PCA based mapping

- MDS (Multidimensional Scaling)

  Given a set of objects and pairwise distances, finds an optimal mapping of the objects into R^n.
- Kernel-PCA

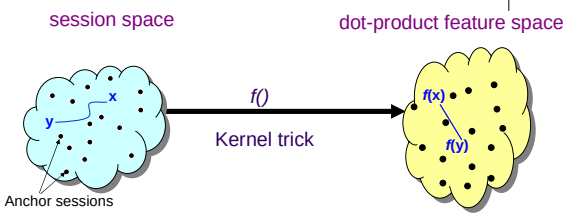
  A kernelized version of PCA (Principal Component Analysis) .
- Kernel-PCA based mapping

  A kernelized version of MDS:

  Given a set of objects and a kernel function, finds an optimal mapping of objects into R^n.

48

Kernel-PCA based mapping (2)



Goals: - Map sessions into feature space

- Model in feature space

Kernel-PCA based mapping (3)

Given - kernel K

- n anchor sessions $A_1,\ldots,A_n$

Find an orthonormal basis for $span\{f(A_1),\ldots,f(A_n)\}$

Method

1) Compute eigenvectors of the centralized kernel-matrix $k_{i,j} = K(A_i,A_j)$.

Normalize eigenvectors by square-roots of corresponding eigenvalues $\rightarrow \{v_i\}$

$\{f_i\}$ for $f_i = (f(A_1),\ldots,f(A_n))v_i$ is the requested basis.

Kernel-PCA based mapping (4)

Common speaker subspace - $C = span\{f(A_1),\ldots,f(A_n)\}$

Speaker unique subspace - $U = F / span\{f(A_1),\ldots,f(A_n)\}$

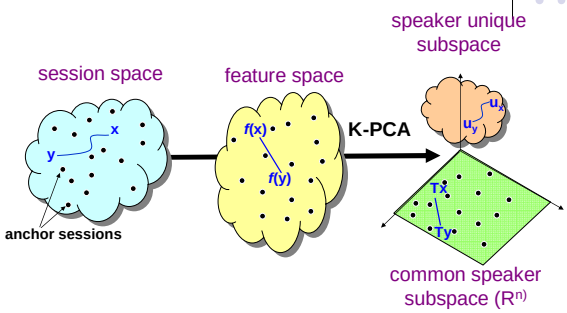
Given sessions x, y, $f(x), f(y)$ may be uniquely represented as:

$f(x) = c_x + u_x$ and $f(y) = c_y + u_y$ with $c_x, c_y \in C$ and $u_x, u_y \in U$

$T: x \rightarrow \begin{pmatrix} v_1 \\ \ldots \\ v_n \end{pmatrix} \begin{pmatrix} K(x,A_1) \\ \ldots \\ K(x,A_n) \end{pmatrix}$ is a mapping $x \rightarrow R^n$ with the property:

$$\|T(x) - T(y)\|^2 = \|c_x - c_y\|^2$$

Kernel-PCA based mapping (5)



Outline

1. Introduction to speech classification.
2. Approximated cross entropy for GMM scoring.
3. Intra-speaker inter-session variability modeling.
4. Mapping speech segments into segment space.
5. Modeling in segment space.
6. More applications.

Modeling in segment space

Modeling techniques

- Parametric distributions: GMMs
- SVMs

Curse of dimensionality

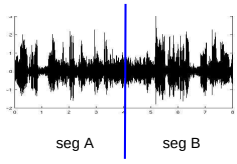
- PCA
- LDA
- Regularization
- Factor Analysis

Outline

1. Introduction to speech classification.
2. Approximated cross entropy for GMM scoring.
3. Intra-speaker inter-session variability modeling.
4. Mapping speech segments into segment space.
5. Modeling in segment space.
6. More applications.

55

Speaker change detection



Hyp I: seg A, seg B share the same speaker identity
Hyp II: seg A, seg B originate from different speakers

→ This is equivalent to the speaker identification task!
→ Same techniques should apply

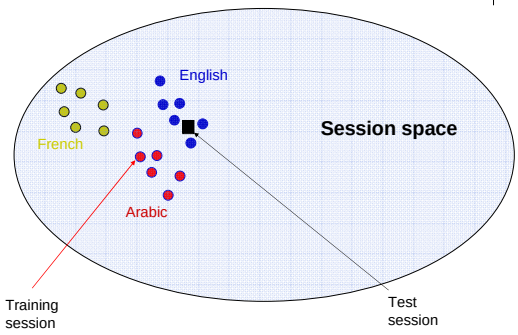
Results: Reduction of more than 50% in SER (speaker error rate)

56

Language Identification

Using GMM super-vectors

LID in session space



58

Algorithm

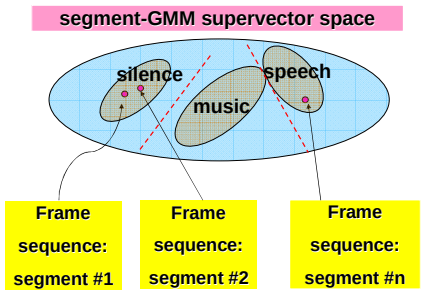
1. Front end: shifted delta cepstrum (SDC).
2. Represent every train/test session by a GMM super-vector.
3. Train a linear SVM to classify GMM super-vectors.

Results

- EER=4.1% on the NIST-03 Eval (30sec sessions).

59

Modeling in Segment-GMM supervector space



60

Segmental Modeling for Audio Segmentation

Goal

- Segment audio accurately and robustly into speech / silence / music segments.

Novel idea

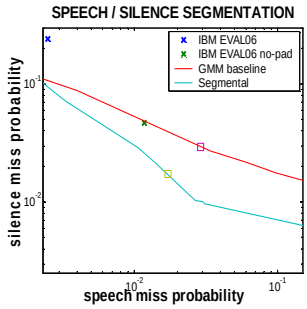
- Acoustic modeling is usually done on a frame-basis.
- Segmentation/classification is usually done on a segment-basis (using smoothing).

Why not **model** explicitly whole segments?

Note: speaker, noise, music-context, channel (etc.) are constant during a segment.

61

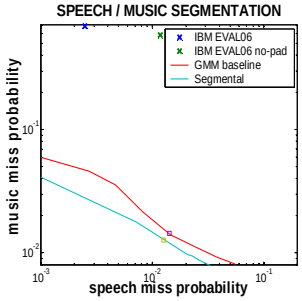
Speech / silence segmentation - results



System	EER	FA @ FR=0.5%	FR @ FA=1%
EVAL06	FA=24.2% @ FR=0.25%		
GMM baseline	2.9%	7.9%	29.6%
Segmental	1.7%	5.1%	2.7%
Error reduction	**41%**	**35%**	**91%**

62

Speech / music segmentation - results



System	EER	FA @ FR=0.5%	FR @ FA=1%
EVAL06	FA=69% @ FR=0.25%		
GMM baseline	1.43%	3.4%	3.2%
Segmental	1.27%	2.0%	1.9%
Error reduction	**11%**	**41%**	**41%**

63

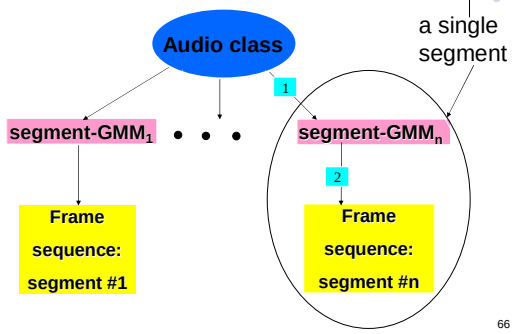
THE END

64

Segmental Modeling for Audio Segmentation

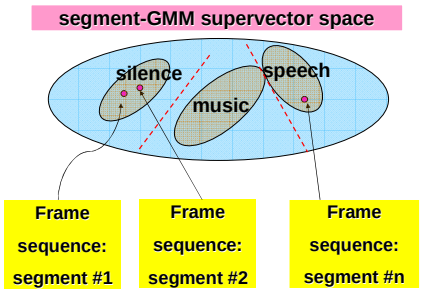
65

Suggested model for segment generation



66

Modeling in Segment-GMM supervector space



Segment modeling

Usual frame independence assumption:

(1) $\Pr(Y = y_1,..,y_n \mid C_j) = \prod_{i=1}^{n} \Pr(y_i \mid C_j)$

Our model:

(2) $$\frac{\Pr(Y \mid C_i)}{\Pr(Y \mid C_j)} = \frac{\int_{f_Y} \Pr(Y \mid f_Y, C_i)\Pr(f_y \mid C_i)df_Y}{\int_{f_Y} \Pr(Y \mid f_Y, C_j)\Pr(f_y \mid C_j)df_Y} = \frac{\int_{f_Y} \Pr(Y \mid f_Y)\Pr(f_y \mid C_i)df_Y}{\int_{f_Y} \Pr(Y \mid f_Y)\Pr(f_y \mid C_j)df_Y} \cong \frac{\Pr(f_Y^* \mid C_i)}{\Pr(f_Y^* \mid C_j)}$$

f_Y – True distribution of segment Y

f_Y^* – Estimated distribution of segment Y

Our frame independence assumption: $\Pr(Y \mid f_Y) = \prod_{i=1}^{n} \Pr(y_i \mid f_Y)$

We assume f_Y is a GMM and estimate it using MAP-adaptation

Training

1. Train a universal GMM (UBM) from entire train-set.
2. For each class:
   A. Define evenly spaced overlapping segments of length L (300ms)
   B. For each segment:
      I) Estimate a **segment-GMM** by adapting the UBM.
      II) Construct **segment-supervector** from **segment-GMM**.
3. Estimate **class-GMMs** using **segments-supervectors** as training data.

   or

   Normalize **supervectors**' covariance and train an SVM using an RBF-kernel.

Segmentation

- Define evenly spaced overlapping segments of length L.
- For each segment:
  1. Estimate a **segment-GMM**
  2. Construct a **supervector-GMM**
  3. Compute classification scores using either **class-GMMs** or SVMs
  4. Smooth scores with adjacent segments.
  5. Compare scores to pre-tuned thresholds and classify.

Segment supervectors

Segment GMM order	Segment supervector	Results
1	mean + log diagonal covariance	Successful
3	mean + log diagonal covariance + log weights	Successful
128	log weights	Unsuccessful (*yet*)

Datasets

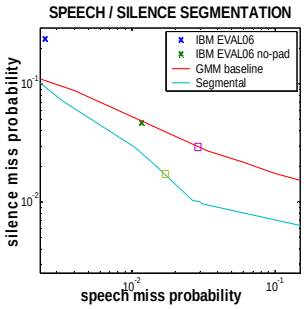
Train

- 4 manually annotated shows (from Arabic training set).

Test

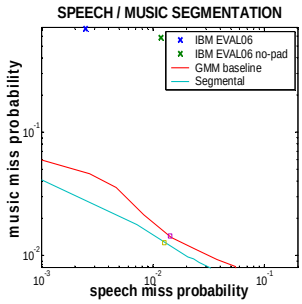
- A manually annotated subset of BNAT05 (Arabic). (a random portion of each show was annotated).
- SNR: 10-40db

Speech / silence segmentation - results



System	EER	FA @ FR=0.5%	FR @ FA=1%
EVAL06	FA=24.2% @ FR=0.25%		
GMM baseline	2.9%	7.9%	29.6%
Segmental	1.7%	5.1%	2.7%
Error reduction	**41%**	**35%**	**91%**

Speech / music segmentation - results



System	EER	FA @ FR=0.5%	FR @ FA=1%
EVAL06	FA=69% @ FR=0.25%		
GMM baseline	1.43%	3.4%	3.2%
Segmental	1.27%	2.0%	1.9%
Error reduction	**11%**	**41%**	**41%**

More details in:

[1] H. Aronowitz, D. Burshtein, A. Amir, "Speaker indexing in audio archives using test utterance Gaussian mixture modeling", in Proc. ICSLP, 2004.

[2] H. Aronowitz, D. Burshtein, A. Amir, "A session-GMM generative model using test utterance Gaussian mixture modeling for speaker verification", in Proc. ICASSP, 2005.

[3] H. Aronowitz, D. Irony, D. Burshtein, "Modeling Intra-Speaker Variability for Speaker Recognition ", in Proc. Interspeech, 2005.

[4] J. Goldberger and H. Aronowitz, "A distance measure between GMMs based on the unscented transform and its application to speaker recognition" , in Proc. Interspeech 2005.

[5] H. Aronowitz, D. Burshtein, "Efficient Speaker Identification and Retrieval", in Proc. Interspeech 2005.

More details in:

[6] E. Noor, H. Aronowitz, "Efficient language Identification using Anchor Models and Support Vector Machines," in Proc. ISCA Odyssey Workshop, 2006.

[7] H. Aronowitz, "Segmental modeling for audio segmentation" in Proc. ICASSP, 2007.

[8] H. Aronowitz, D. Burshtein, "Efficient Speaker Recognition Using Approximated Cross Entropy (ACE)", in IEEE Trans. on Audio, Speech & Language Processing, September 2007.

[9] H. Aronowitz, "Speaker Recognition using Kernel-PCA and Intersession Variability Modeling", in Proc. Interspeech, 2007.

[10] H. Aronowitz, "Trainable Speaker Diarization", in Proc. Interspeech, 2007.