Machine Learning for Speaker Recognition

NIPS'08 Workshop on *Speech and Language: Learning-based Methods and Systems*

Andreas Stolcke

Speech Technology and Research Laboratory

SRI International

Joint work with:

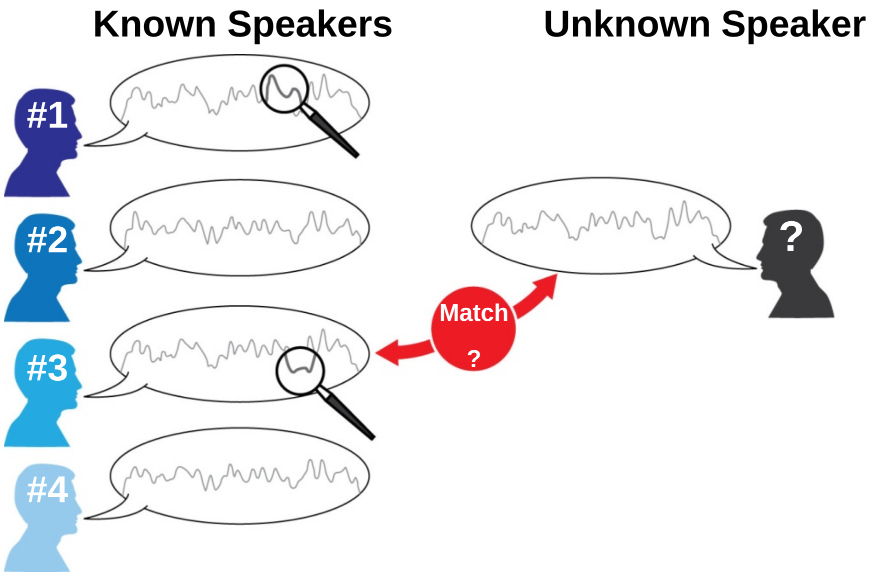
Luciana Ferrer, Sachin Kajarekar, Nicolas Scheffer, Elizabeth Shriberg, Robbie Vogt (QUT)

Outline

- What is speaker recognition?
- Feature extraction & normalization
- Modeling & classification
- System combination
- Open issues – future directions
- Summary

Speaker Recognition

- Speaker **identification**
  - Closed set of speakers
  - Test speaker one in set
  - 1-in-n classification



- Speaker **verification**
  - Single target speaker
  - Test speaker is target speaker or unknown
  - Binary classification (detection) task
  - Focus of this talk
    - more fundamental, widely researched

Speaker Verification - Metrics

- Equal error rate (EER)
  - False reject probability = false accept probability

- Detection cost function (DCF) =
  - $P(\text{FR})\, C(\text{FR})\, P(\text{target}) + P(\text{FA})\, C(\text{FA})\, (1-P(\text{target}))$
  - $C(\text{FR})$, $C(\text{FA})$, $P(\text{target})$

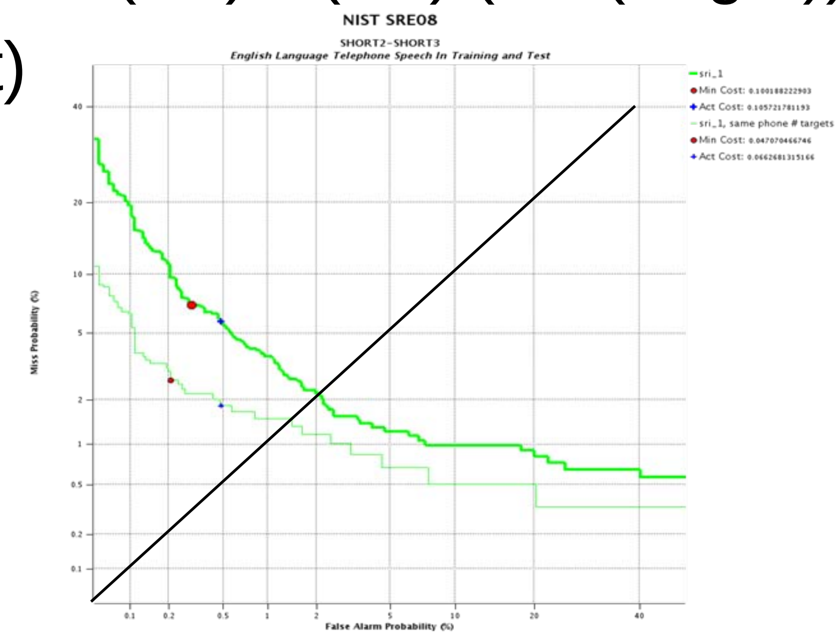
  application-dependent

- DET plots

  Detection

  Error

  Tradeoff



SRI International

High-level Structure of SR System

1. Audio data
2. Feature extraction
3. Modeling training $\Rightarrow$ target speaker model
4. Model testing: apply speaker model to test speaker features $\Rightarrow$ verification score s
5. Classification:

   $s > T \Rightarrow$ same speaker

   $s < T \Rightarrow$ different speaker (impostor)

Features for SR

- “Low-level” (classical approach)
  - Short-term spectral features (e.g., 25 ms)
  - No sequence modeling (beyond delta features)
  - Reflect vocal tract shape - **GOOD**
  - Highly dependent on channel, environment - **BAD**

- “High-level” (relatively recent)
  - Longer-term extraction region AND/OR
  - Based on linguistic units (words/syllables/phones)
  - Tend to reflect stylistic aspects of speech - **GOOD**
  - Requires complex features or ASR - **BAD**

Features - Examples

- Low-level:
  - Mel frequency or PLP cepstrum
  - Pitch
- High-level
  - Word/Phone conditioned low-level features
  - Pitch contours
  - Phone durations
  - Phone/word token sequences

Modeling of Speaker Features

- Generative models
  - Cepstral GMM-UBM
  - Language models
- Discriminative models
  - Support vector machines
  - Sequence kernels
  - Feature normalization

UBM-based Likelihood Ratios

- Estimate

$$\text{score} = \log \frac{P(\text{target} \mid D)}{P(\text{impostor} \mid D)} = \log \frac{P(\text{target})}{P(\text{impostor})} + \log \frac{P(D \mid \text{target})}{P(D \mid \text{impostor})}$$

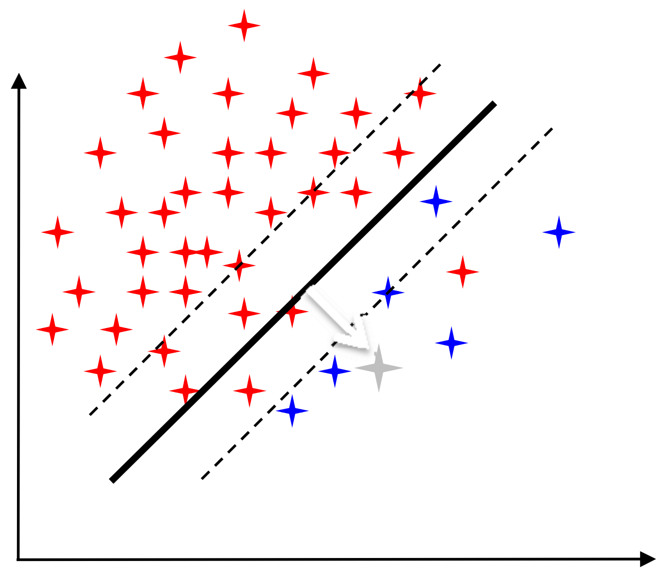
- $P(D \mid \text{target})$: target speaker model
- $P(D \mid \text{impostor})$: universal background model (UBM), trained on large population
- Normalize log-LR by utterance length to ensure comparability in thresholding
- Log prior odds add a constant offset to threshold

UBM-LR Examples

- Low-level:
  - Features = short-term cepstra
  - Likelihoods estimated by GMMs
  - State-of-the-art until recently [Reynolds et al. 2000]
- High-level:
  - Features = phone or word N-grams
  - Likelihoods estimated by N-gram LMs
- For robustness and normalization of LRs:
  - Target models derived from UBM by MAP-adaptation

Discriminative Modeling - SVMs

- Each speech sample generates a point in a derived feature space
- The SVM is trained to separate the target sample from the impostor (= UBM) samples
- Scores are computed as the Euclidean distance from the decision hyperplane to the test sample point
- SVMs training is biased against misclassifying positive examples (typically very few, often just 1)



Feature Transforms for SVMs

- SVMs have been a boon for SR research – allow great flexibility in the choice of features
- However, require a “sequence kernel”
- Dominant approach: transform variable-length feature stream into fixed, finite-dimensional feature space
- Then use linear kernel
- All the action is in the feature transform!

Cepstral Feature Transforms

- Polynomial expansion [Campbell 2002]
  - Expand each frame of features into polynomial vector:

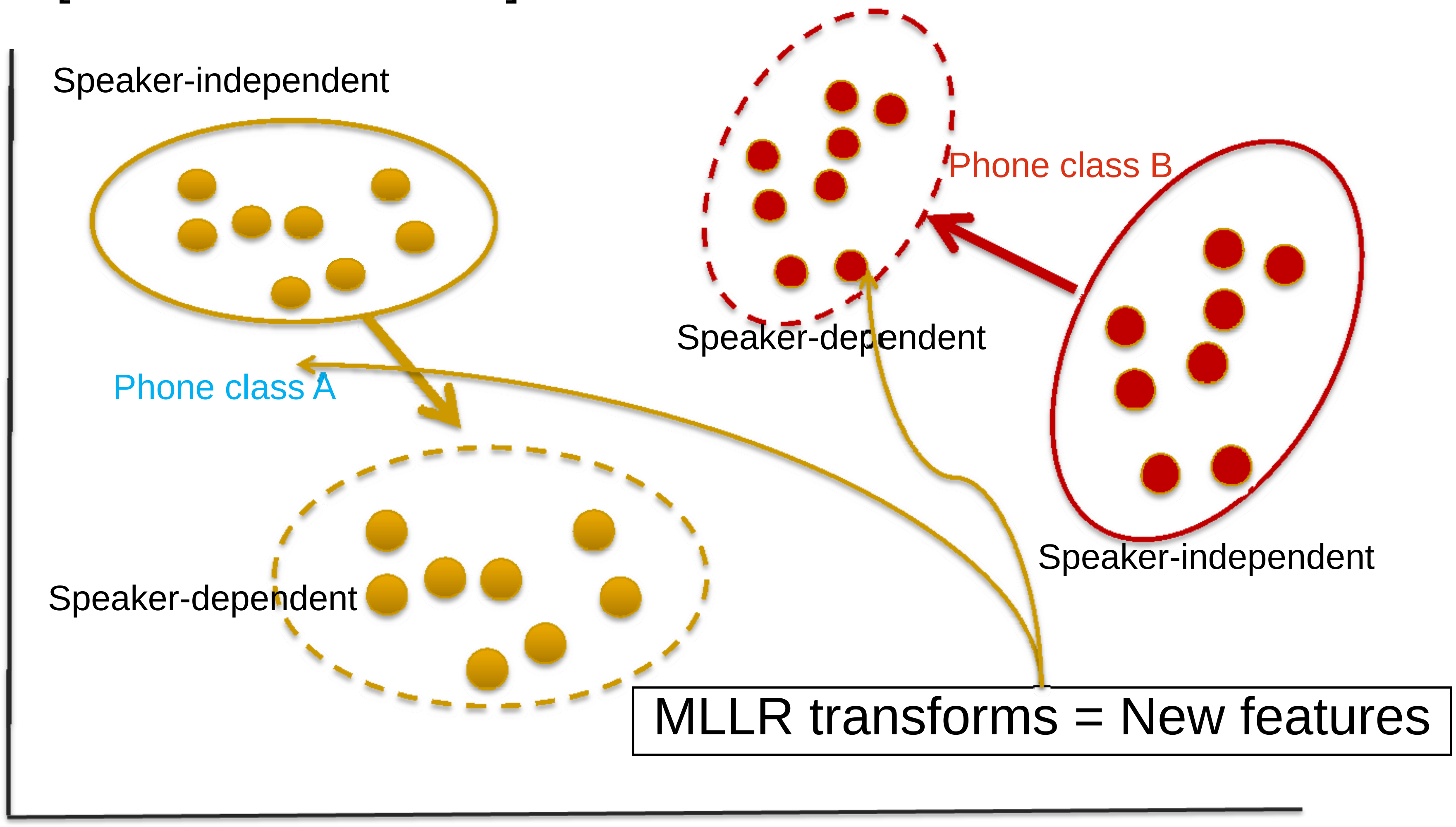
$$Y(X) = poly(X,2) = [X\ (x_1\ x_2 \ldots x_n)^2] \Rightarrow (X\ x_1^2\ \ x_1x_2\ \ x_1x_n\ \ \ldots.\ \ x_2^2\ \ \ldots x_n^2)$$

  - Mean and variance of expanded vectors is estimated over whole speech sample
  - Captures lower-order moments of feature distribution in a single vector

- GMM supervectors [Campbell et al. 2006]
  - MAP-adapt UBM-GMM to target speaker data
  - Stack all gaussian means into one “supervector”
  - Optional: Scale by variances
  - Use supervector as SVM feature vector
  - Can be interpreted as KL distance between GMMs

Feature Transforms via MLLR

[Stolcke et al. 2005]



Cepstral Model Comparison

- EER on NIST SRE'06

	1 train sample	8 train samples
GMM LLR	6.15	4.58
GMM-SV SVM	5.56	4.78
MLLR SVM	4.31	2.84

- Note: MLLR transform can leverage detailed ASR speech models and feature normalizations

Prosodic Modeling

- Syllable-based prosodic features
  [Shriberg et al. ‘05, Ferrer et al. ‘07]
  - Train global GMM that models observation vectors: pitch, energy, durations
  - Adapt mixture weights to speaker data
  - Use adapted weight vector as feature (a kind of Fisher kernel)
- Pitch and energy contours [Dehak et al. ‘07]
  - Fit Legendre polynomials
  - Use coefficients as feature vector

Token-Based Speaker Modeling

- Goal: model a phone [Andrews et al. '02] or word [Doddington '01] token stream
  - Captures pronunciation and idiolectal differences
  - Also, applicable to some prosodic features
- Compute N-gram frequencies from each sample, normalized by utterance length
- Frequencies of top-N n-gram types form (sparse) feature vector, suitable for SVM
- Requires proper scaling of feature dimensions (next slide)

Feature Scaling for SVMs

- SVMs are sensitive to scale of features
- Absent prior knowledge or explicit optimization [Hatch et al. '05], need to equate dynamic range of dimensions
- Proposed methods:
  - Variance normalization
  - TFLLR: kernel emulates LLR between N-gram models [Campbell NIPS'03]
  - TFLOG: similar to TF-IDF [Campbell '04]
  - Rank normalization
    - Maps feature space to uniform distribution
    - Distance between samples ≈ % population between them

Feature Scaling Comparison

- Comparison of feature scaling methods on a variety of features, modeled by SVMs [Stolcke et al. 2008]

- NIST SRE'06 EER

Feature	None	Variance	TFLLR	TFLOG	Rank norm
MLLR	5.29	3.94			**3.61**
Prosody	14.19	14.08			**13.65**
Phone N-ngrams	12.30	10.84	10.73		**10.30**
Word N-grams	22.98	31.07		**21.63**	23.19

- Note: TFLLR/TFLOG were proposed specifically for phone/word N-grams, respectively

- Rank norm seems to perform reasonably regardless of feature

Intra-Speaker Variability (1)

- Variability of the same speaker between recordings may overwhelm between-speaker differences

- Speaker recognition is the converse of Speech recognition

- Two old approaches:
  - Feature normalization [Reynolds et al. '03]
  - Score normalization: mean/variance normalization according to scores from
    - Other speaker models on same test data
    - Same speaker model on different test data

Intra-Speaker Variability in SVMs

- Nuisance Attribute Projection (NAP) [Solomonoff et al. '04]
  - Remove directions of the feature space that are dominated by intra-speaker variability
  - Estimate within-speaker feature covariance from a database of speaker with multiple recordings
  - Project into the complement of the subspace $\mathbf{U}$ spanned by the top-K eigenvectors:

$$\mathbf{y}' = \left(\mathbf{I} - \mathbf{U}\mathbf{U}^T\right)\mathbf{y}$$

  - Model with SVM's as usual

Factor Analysis with GMMs (1)

[Kenny et al. '05, Vogt et al. '05]

- An utterance h is best modelled by a GMM with mean supervector $\boldsymbol{\mu}_h(s)$, based on speaker and session factors

$$\boldsymbol{\mu}_h(s) = \boldsymbol{\mu}(s) + \mathbf{U}\mathbf{z}_h(s)$$

  - The **true speaker mean** $\boldsymbol{\mu}(s)$ is assumed to be independent of session differences.
  - **Session factors** exhibit an additional mean offset $\mathbf{z}_h(s)$ in a restricted, **low-dimensional subspace** represented by the transform $\mathbf{U}$
  - $\mathbf{U}$ is same as for NAP

Factor Analysis with GMMs (2)

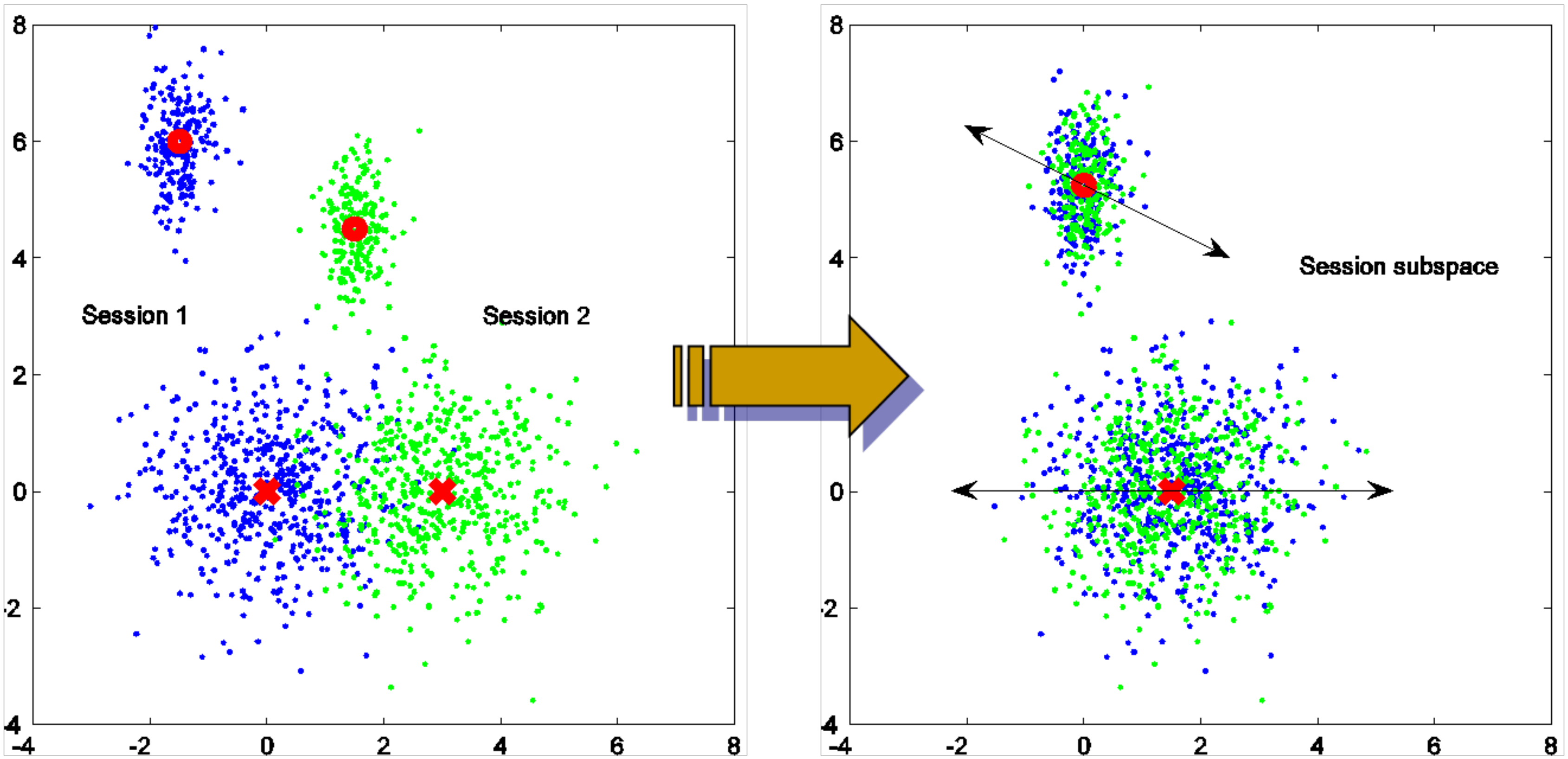
- Assuming $\mathbf{\mu}(s)$ is MAP adapted from the UBM mean $\mathbf{m}$,

$$\mathbf{\mu}(s) = \mathbf{m} + \mathbf{y}(s)$$

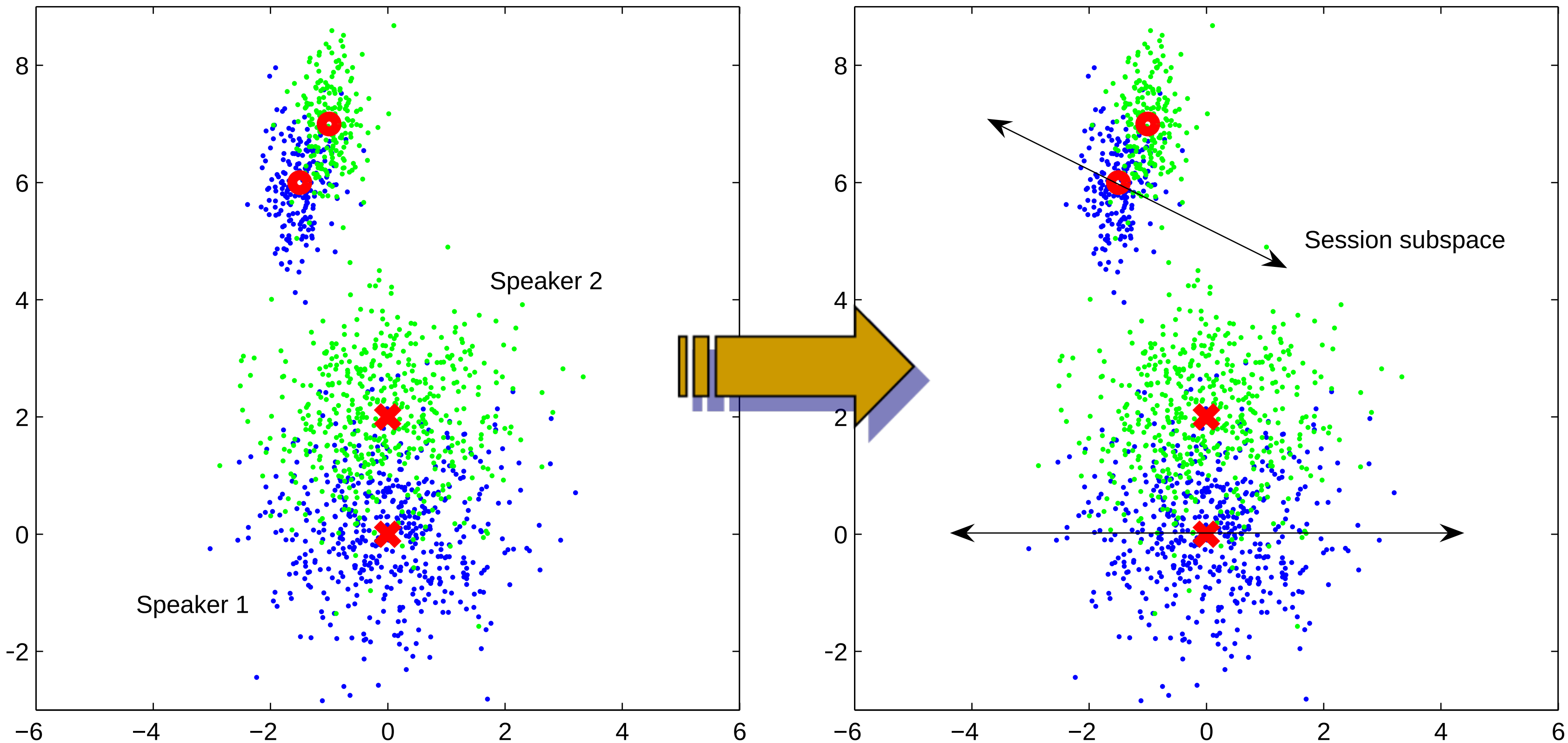
  - $\mathbf{y}(s)$ is the speaker offset from the UBM

- During target model training, $\mathbf{\mu}(s)$ and all $\mathbf{z}_h(s)$ are optimised **simultaneously**
  - $\mathbf{\mu}(s)$ using Reynolds' MAP criterion
  - $\mathbf{z}_h(s)$ using a MAP criterion with standard normal prior in the session subspace
  - Only the true speaker mean $\mathbf{\mu}(s)$ is retained

Intra-Speaker Variability: **Same** Speaker



Intra-Speaker Variability: **Different** Speakers



Cepstral Models with Intra-Speaker Variability Modeling

- EER on NIST SRE'06, 1-sample training

	Without ISV	With ISV
GMM LLR	6.15	4.75
GMM-SV SVM	5.56	4.21
MLLR SVM	4.31	3.61

- MLLR benefits the least because it already conditions-out variability due to phonetic content

Other Recent Developments (1)

- Joint factor analysis [Kenny et al. '06]
  - Constrain speaker means to vary in a low-dimensional subspace:

$$\boldsymbol{\mu}(s) = \mathbf{m} + \mathbf{V}\mathbf{x}(s) + \mathbf{y}(s)$$

  - $\mathbf{V}$ is subpace spanned by "eigenspeakers"
  - $\mathbf{y}(s)$ is the speaker residual and could be dropped if eigenspeaker space is good enough
  - Current the best-performing approach

- $\mathbf{x}(s)$ can be used as a (much lower-dimensional) feature vector

Other Recent Developments (2)

- Modeling of SVM weight correlation (prior) for SVM [Ferrer et al. '07]
  - Estimate weight covariance on well-trained speaker models
  - Prior folded into kernel function
- Decorrelating SVM classifier training for better system combination [Ferrer et al. '08a]
  - Train classifier A (any type)
  - Train SVM classifier B, penalized for score correlation with classifier A

Other Recent Developments (3)

- Constrained cepstral GMMs

  [Bocklet & Shriberg, 2009]

  - Ensemble of cepstral GMMs conditioned on syllable regions
  - Regions constrained by lexical and linguistics context (from ASR)
  - Syllables may be selected by multiple constraints, or not at all
  - Subsystems combined at score level (next slide)

System Combination

- Widely used for combining systems that differ either in features or modeling approach
- Methods used:
  - Neural net
  - SVM
  - Linear logistic regression
    - Works about as well as any anything else
- Conditioning combiner on auxiliary variables [Ferrer et al. ’08b]
  - On metadata: language, channel
  - Automatically extracted acoustic features (SNR)

Data Properties

- Typical NIST SRE task
  - Dimension of expanded feature space: 10k-100k
  - Positive sample size: 1, 3, or 8
  - Negative (impostor) sample size: 2-5k
  - 20k to 100k model-test sample pairings ("trials")
  - Sample duration: 5 minutes (2.5 min. of speech)
  - Challenging but doable with freely available SVM software [libSVM, SVMlight]

Research Issues

- Features
  - Preservation of sequence information in feature extraction
- Modeling
  - Coping with data mismatch
    - ISV model training on mismatched channel / style
  - Unsupervised training
  - Better feature/model combination
  - Discriminative training (in generative framework)
  - Graphical models?

Summary

- Dominant features: cepstral
- Dominant models: GMMs and SVM
- SVMs have opened door to many novel feature types – easy once feature transform into fixed-dim. linear space is defined
- Focus on modeling within-class (with-speaker) variability (NAP, JFA)
- Speaker recognition is a rich application field for ML research – **We need you!**

Questions

References (1)

W. D. Andrews, M. A. Kohler, J. P. Campbell, J. J. Godfrey, and J. Hernandez-Cordero (2002), Gender-dependent phonetic refraction for speaker recognition, *Proc. IEEE ICASSP*, vol. 1, pp. 149-152, Orlando, FL.

T. Bocklet & E. Shriberg (2009), Speaker Recognition Using Syllable-Based Constraints for Cepstral Frame Selection, *Proc. IEEE ICASSP*, Taipei, to appear.

W. M. Campbell (2002), Generalized Linear Discriminant Sequence Kernels for Speaker Recognition, *Proc. IEEE ICASSP*, vol. 1, pp. 161-164, Orlando, FL.

W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek (2004), Phonetic Speaker Recognition with Support Vector Machines, in *Advances in Neural Processing Systems 16*, pp. 1377-1384, MIT Press, Cambridge, MA.

W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek (2004), High-level speaker verification with support vector machines, *Proc. IEEE ICASSP*, vol. 1, pp. 73-76, Montreal.

W. M. Campbell, D. E. Sturim, D. A. Reynolds (2006), Support vector machines using GMM supervectors for speaker verification, *IEEE Signal Proc. Letters* 13(5), 308-311.

N. Dehak, P. Dumouchel, and P. Kenny (2007), Modeling Prosodic Features With Joint Factor Analysis for Speaker Verification, *IEEE Trans. Audio Speech Lang. Proc.* 15(7), 2095-2103.

G. Doddington (2001), Speaker Recognition based on Idiolectal Differences between Speakers, *Proc. Eurospeech*, pp. 2521-2524, Aalborg.

References (2)

L. Ferrer, E. Shriberg, S. Kajarekar, and K. Sonmez (2007), Parameterization of Prosodic Feature Distributions for SVM Modeling in Speaker Recognition, *Proc. IEEE ICASSP*, vol. 4, pp. 233-236, Honolulu, Hawaii.

L. Ferrer, K. Sonmez, and E. Shriberg (2008a), An Anticorrelation Kernel for Improved System Combination in Speaker Verification. *Proc. Odyssey Speaker and Language Recognition Workshop*, Stellenbosch, South Africa.

L. Ferrer, M. Graciarena, A. Zymnis, and E. Shriberg (2008b), System Combination Using Auxiliary Information for Speaker Verification, *Proc. IEEE ICASSP*, pp. 4853-4857, Las Vegas.

L. Ferrer (2008), Modeling Prior Belief for Speaker Verification SVM Systems, *Proc. Interspeech*, pp. 1385-1388, Brisbane, Australia.

A. O. Hatch, A. Stolcke, & B. Peskin (2005), Combining Feature Sets with Support Vector Machines: Application to Speaker Recognition. *Proc. IEEE Speech Recognition and Understanding Workshop*, pp. 75-79, San Juan, Puerto Rico.

P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel (2005), Factor Analysis Simplified, *Proc. IEEE ICASSP*, vol. 1, pp. 637-640, Philadelphia.

P. Kenny, G. Boulianne, P.Ouellet, and P. Dumouchel (2006), Improvements in Factor Analysis Based Speaker Verification, *Proc. IEEE ICASSP*, vol. 1, pp. 113-116, Toulouse.

References (3)

D. A. Reynolds, T. F. Quatieri, and R. B. Dunn (2000), Speaker Verification Using Adapted Gaussian Mixture Models, *Digital Signal Processing* 10, *181-202.*

D. Reynolds (2003), Channel Robust Speaker Verification via Feature Mapping, *Proc. IEEE ICASSP*, vol. 2, pp. 53-56, Hong Kong.

E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke (2005), Modeling prosodic feature sequences for speaker recognition, *Speech Communication* 46(3-4), 455-472.

A. Solomonoff, C. Quillen, and I. Boardman (2004), Channel Compensation for SVM Speaker Recognition, *Proc. Odyssey Speaker Recognition Workshop*, pp. 57-62, Toledo, Spain.

A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman (2005), MLLR Transforms as Features in Speaker Recognition. *Proc. Eurospeech*, Lisbon, pp. 2425-2428.

A. Stolcke, S. Kajarekar, and L. Ferrer (2008), Nonparametric Feature Normalization for SVM-based Speaker Verification, *Proc. IEEE ICASSP*, pp. 1577-1580, Las Vegas.

R. Vogt, B. Baker, and S. Sridharan (2005), Modelling Session Variability in Text-independent Speaker Verification, *Proc. Eurospeech*, pp. 3117-3120, Lisbon.