MLLR Transform and Constrained Cepstral Modeling

Winter School on Speech and Audio Processing
IIT Kanpur, January 2009

Andreas Stolcke

Speech Technology and Research Laboratory
SRI International, Menlo Park, Calif., U.S.A.
Joint work with:
E. Shriberg, T. Bocklet , S. Kajarekar, L. Ferrer, N. Scheffer,
M. Akbacak, R. Vogt (QUT)

Overview

- Higher-level Cepstral Modeling
- MLLR transform modeling
- ISV compensation
- Constrained cepstral modeling
- Combined results
- Summary

- Bonus feature: Nonnativeness detection

Higher-level Cepstral Modeling

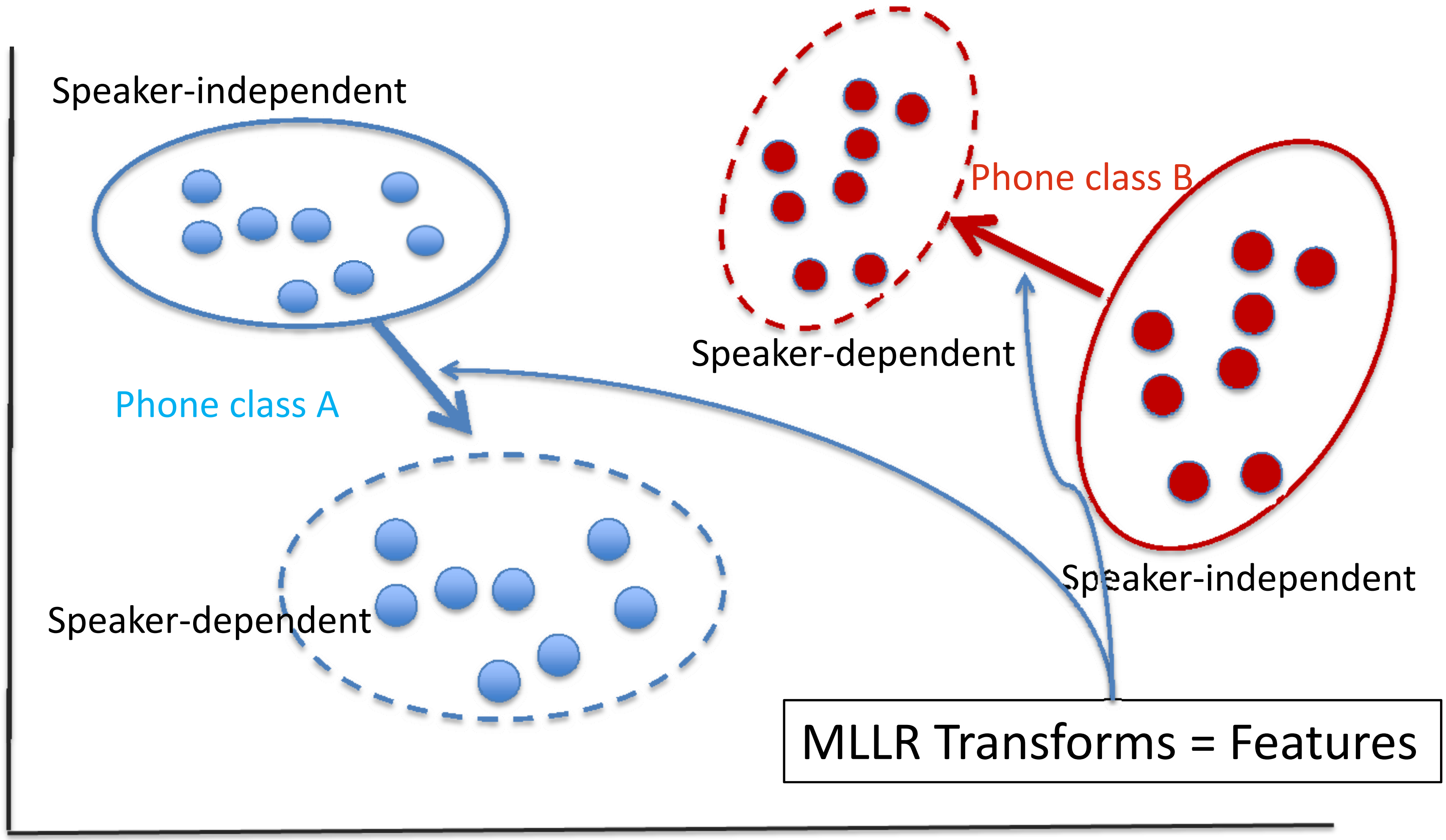
- How to augment low-level cepstral features with higher-level information?
- Rationale: remove variability due to phonetic content
- Allows text-dependent modeling in text-independent speaker recognition
- Main approach: condition (constrain) cepstral frames on specific linguistic units
  - Phone-conditioned cepstral models (survey in Park & Hazen '02; Kajarekar '05)
  - Word-conditioned cepstral models (Sturim et al. '02)
  - Syllable-conditioned (Baker et al. '05, Bocklet & Shriberg '09)
- Whole-word HMM modeling (Boakye & Peskin '04)
- MLLR transform modeling (Stolcke et al. '05, '07)

MLLR Transform Modeling

MLLR Transforms as Speaker Features

- How can we factor out what was said when comparing cepstral features?
  - Traditional approach: text-dependent speaker verification or text-conditioned cepstral features
  - But conditioning fragments the data
- Idea: use MLLR speaker adaptation parameters used by recognizer
  - Conditions features on what was said
  - But doesn't fragment the data, because transforms are shared among phone models

MLLR Adaptation Transforms



Maximum Likelihood Linear Regression

- Speaker adaptation in ASR
  - Affine mapping of Gaussian means turn speaker-independent into speaker-dependent models

$$\mu' = \mathbf{A}\mu + \mathbf{b}$$

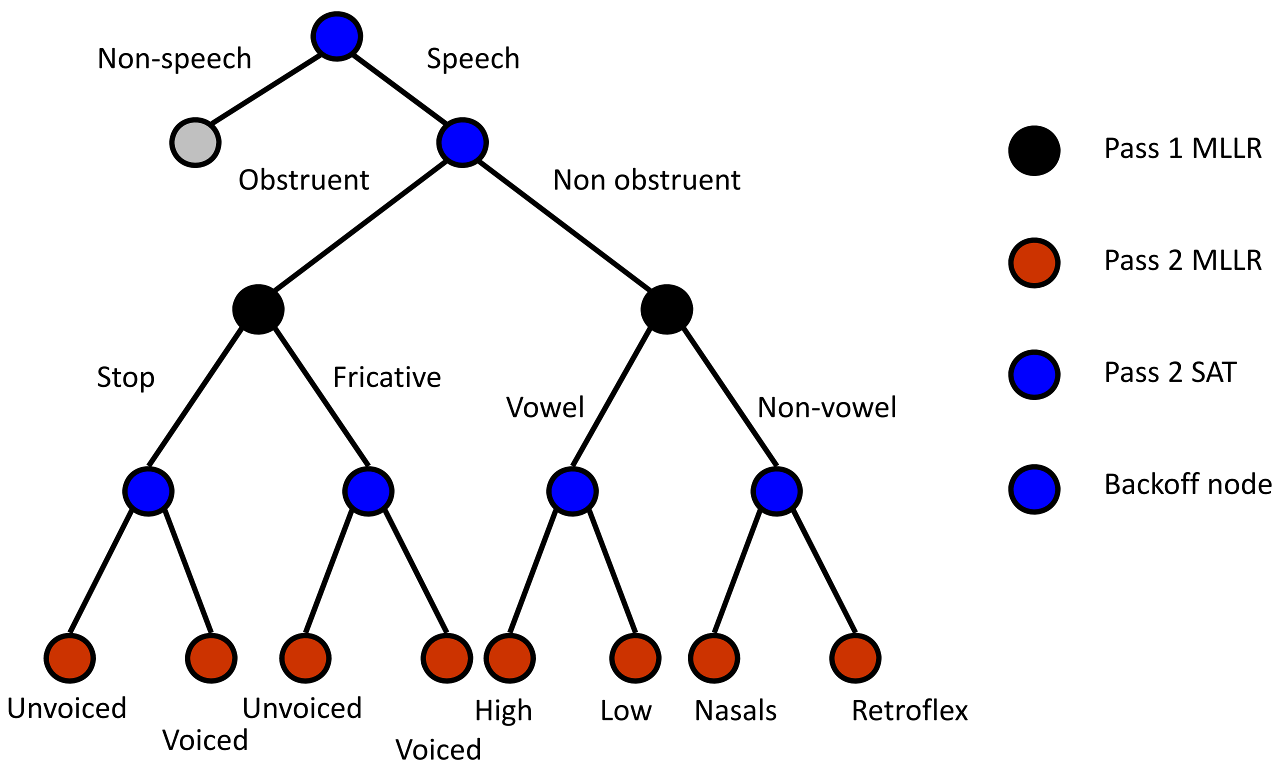
  - Estimated with maximum likelihood and EM
  - Two options for utterance model:
    - Phone-loop (doesn’t require word models, can be applied to any language)
    - Word hypothesis from prior recognition pass (language-dependent)

MLLR Computation Details

- Applied to 39-dim PLP features
  - reduced from 52-dim via HLDA
- ASR frontend normalizations:
  - Cepstral mean + variance normalization
  - Vocal tract length normalization
  - Feature transform estimated with constrained MLLR (speaker adaptive training)
- Acoustic models:
  - Trained on Switchboard 1 and other transcribed telephone speech
  - Gender-dependent
- 9 phone regression classes
  - 8 speech
  - 1 non-speech

MLLR Phone Classes

- All (tri)phones in one class share a transform
- 9 leaf nodes = 9 transforms per speaker
- Backoff tree used when not enough data per class/speaker



MLLR Feature Extraction

1. MLLR estimation
2. Concatenate **A** and **b** coefficient into one vector
3. Concatenate all speech transform vectors into one "supervector"
   - Discard nonspeech transform
4. Repeat 1-3 for the opposite gender-specific model, concatenate "male" and "female" supervectors
5. Rank-normalize each feature component [see 2nd lecture]

Feature dimensionality: (40 x 39) x 8 x 2 = 24960

MLLR Features: Miscellaneous Findings

- Combining male and female transforms reduces EER (SRE-04):

	1-side training	8-side training
Male transforms (8)	6.25	3.21
Female transforms (8)	6.54	3.21
Male + female transforms (16)	5.34	2.62

- 8 regression classes / transforms seems to be near optimal
  - Fewer or more classes give worse results
  - Probably dependent on ASR model and recognition accuracy
- Surprisingly, speaker normalizations in ASR frontend **help** system performance – **This needs further investigation!**
  - Leaving out VTLN hurts
  - Leaving out CMLLR transform hurts

MLLR-SVM and Cepstral GMM

- SRE-05 testset
- Neural network combiner trained on SRE-04

	1-side training
Cepstral GMM	7.22
MLLR SVM	5.91
Combined	4.84

- System complement each other
  - Different frontend features (MFCC vs. PLP)
  - Different modeling approaches

MLLR Features for Multiple Languages

- Speaker verification on Arabic data (Stolcke & Kajarekar '04)
  - Arabic conversations contained in SRE-04 and SRE-05 multilingual data
  - Background data: various dialectal Arabic corpora from LDC
- Tried two kinds of phone-loop MLLR reference models
  - English-trained, gender-dependent
  - Modern Standard Arabic, unisex (resampled to match phone channel)

	EER
Cepstral GMM	9.1
English MLLR SVM (male + female xform)	8.4
English MLLR SVM (female xform only)	9.6
Arabic MLLR SVM (unisex xform)	10.4

- English-trained MLLR works better, especially if dual-gender combination is exploited!

Other Work on MLLR Features

- MLLR features can be simplified
  - Use feature-level transform (CMLLR)
  - Use GMM instead of ASR-HMM as reference model for all frames
  - Not as powerful as ASR-based MLLR, but more convenient
  - Details in Ferras et al. (2007)

- Investigation of different SVM kernels based on MLLR transforms
  - For GMM-based MLLR, can define kernel that represents KL-divergence between speaker-adapted GMMs
  - Unfortunately results don't apply to HMM-based MLLR and rank-normed features (which is empirically the best approach)
  - Details in Karam & Campbell (2008)

Intra-Speaker Variability Compensation

Intra-Speaker Variability

- Variability of the same speaker between recordings may overwhelm between-speaker differences
- Speaker recognition is the converse of Speech recognition
- Two old approaches:
  - Feature mapping (Reynolds et al. ’03)
  - Score normalization: mean/variance normalization according to scores from
    - Other speaker models on same test data (Z-norm, H-norm)
    - Same speaker model on different test data (T-norm)
- Terminology:

  Intra-speaker variability = inter-session variability = ISV

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:

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

  - Optimize *K* on held-out data
  - Model with SVM’s as usual

Factor Analysis with GMMs

(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

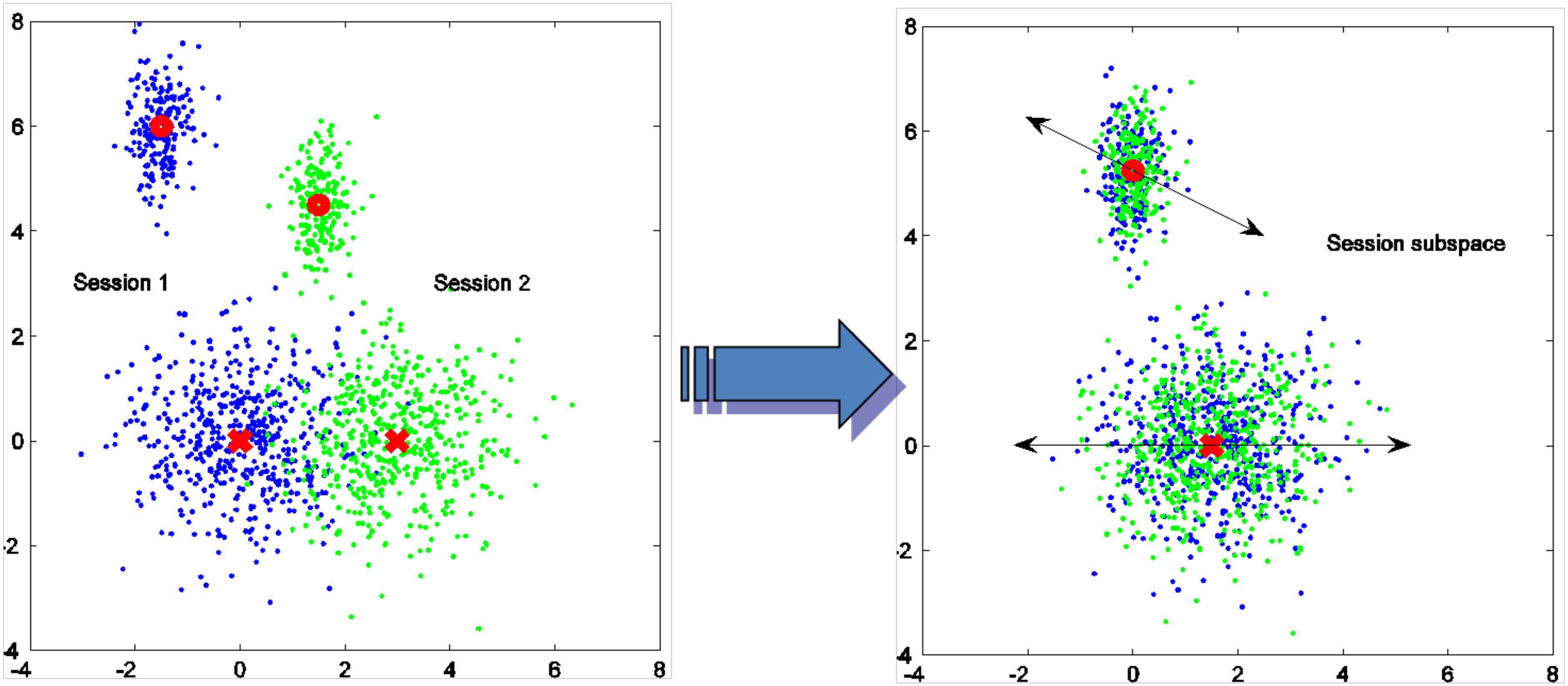
Factor Analysis with GMMs (cont.)

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

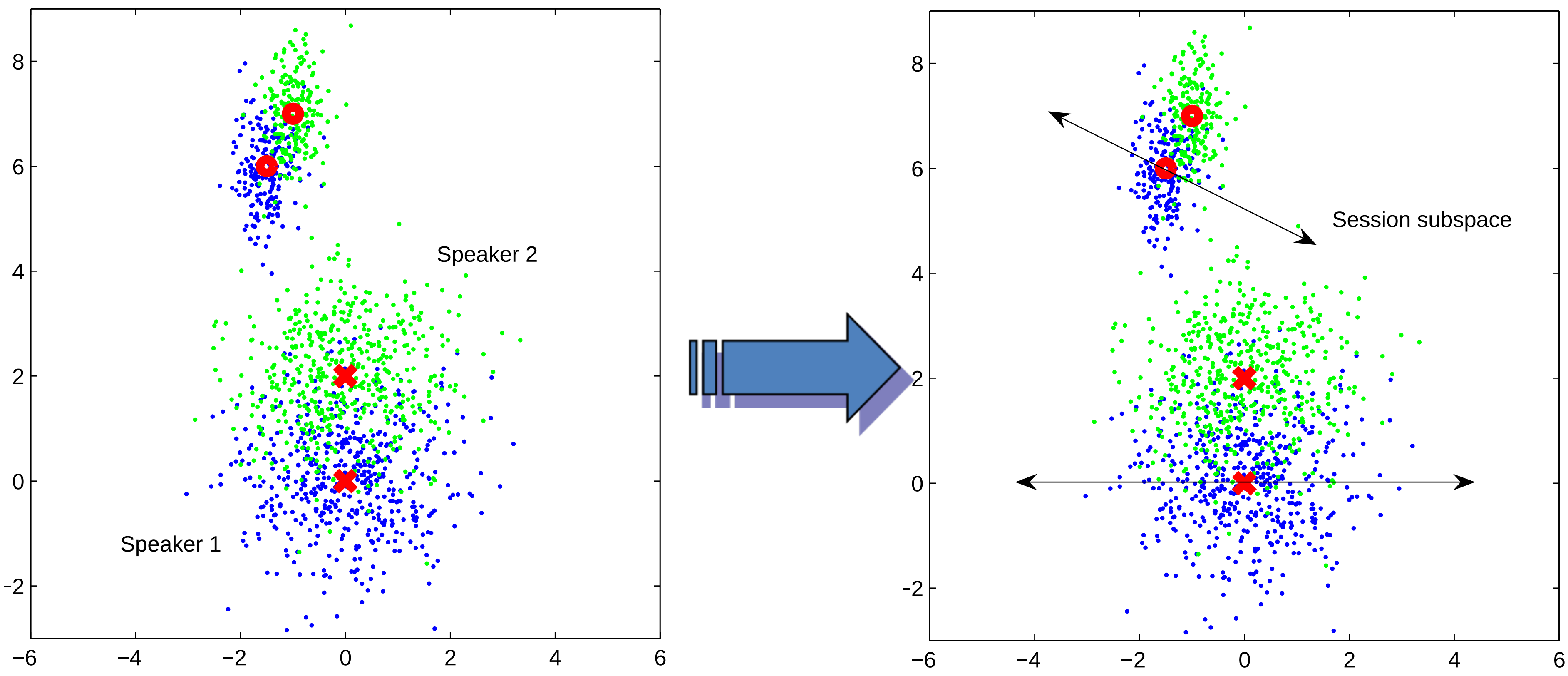
$$\boldsymbol{\mu}(s) = \mathbf{m} + \mathbf{y}(s)$$

  - $\mathbf{y}(\mathrm{s})$ is the speaker offset from the UBM
- During target model training, $\boldsymbol{\mu}(s)$ and all $\mathbf{z}_h(s)$ are optimized **simultaneously**
  - $\boldsymbol{\mu}(s)$ 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 $\boldsymbol{\mu}(s)$ is retained

Intra-Speaker Variability: **Same** Speaker



Intra-Speaker Variability: **Different** Speakers



ISV Compensation Results

- Compared three cepstral systems
- One system is cepstral "supervector" SVM (Campbell et al. '06)
- SRE'06 test data

	ISV Method	1-side training		8-side training	
		No ISV	ISV	No ISV	ISV
Cepstral GMM	FA	6.15	4.75	4.58	2.79
Supervector SVM	NAP	5.56	4.21	4.78	3.33
MLLR SVM	NAP	4.31	3.61	2.84	2.64

- Cepstral GMM and supervector SVM improve more with ISV, especially for 8-side training
- MLLR ISV has smaller number of nuisance dimensions
  - Phone conditioning already removes some ISV

Constrained Cepstral Modeling

Constrained Cepstral Modeling: Motivation

- Two reasons for constraining cepstral features:
  - Reduce intra-speaker variability
  - Capture regions of high inter-speaker variability, i.e.,
  - Emphasize words/syllables/phones where speakers "sound more like themselves"

- Unlike previous word- or phone-conditioned cepstral systems:
  - Uses automatic syllabification of phone output from ASR
  - Model does not cover all frames, and subsets can reuse frames

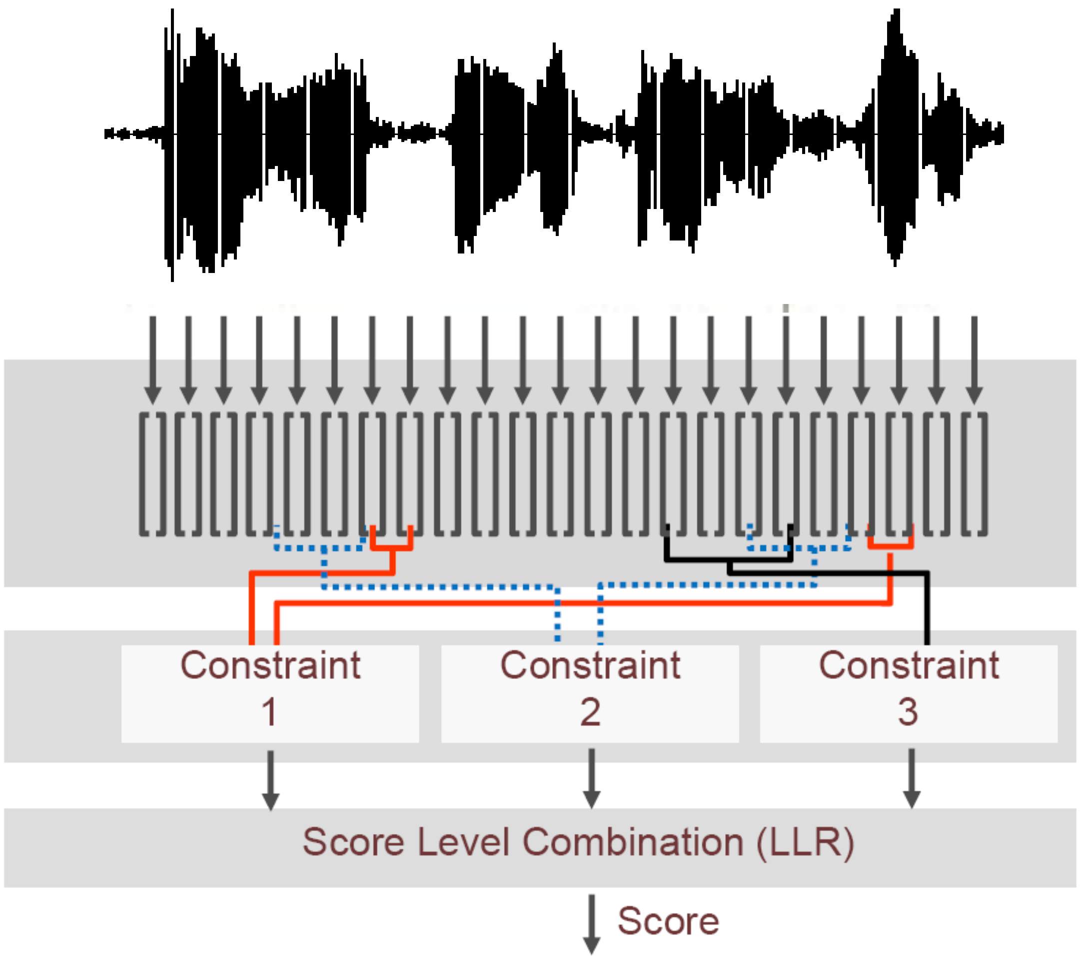
- First employed in SRI 2008 SRE submission – to be published in ICASSP '09 (Bocklet & Shriberg , 2009)

Constrained Cepstral GMM

Speech

Features (MFCCs)

Constrained GMMs



Constrained GMMs

- Feature extraction conditioned/restricted to 4 syllable based, 1 word based and 3 phone based constraints
  - Based on syllabification of phone alignments from ASR

- Syllable/word based constraints:
  1.-3. Syllable onset / nucleus / coda
  4. Syllables following pauses
  5. Monosyllabic words
- Phone based constraints:
  6. Phone [T]
  7. Any of the phones [B,P,V,F]
- Modeling
  - GMMs, background models trained on SRE04, no altmic data
  - ISV: 50 eigenchannels trained on SRE04+05 altmic data
  - Score combination via linear logistic regression
  - ZT-Norm used for score normalization (trained on SRE04)

Constrained Cepstral GMMs: Results

- Results on SRE08 English data
- 4 or 5 constraints give similar performance to 8
- Best systems include nucleus, onset, and [N]-in-syllable constraints

Constraint/System	EER
Syl. onset	5.70
Syl . nucleus	4.48
Syl. coda	8.07
Post-pause	8.80
Monosyllabic words	4.40
Syl. with [N]	10.99
Syl. with [T]	9.53
Syl. with [B,P,V,F]	12.05
All Constraints combined	**2.77**
Unconstrained GMM	**2.91**

All System Results

- Results (EER) on SRE'08 English dataset
- All systems use ISV compensation (FA or NAP)

Systems (gray = ASR-dependent)	**1-side training**	**8-side training**
Constrained cepstral GMM	2.769	0.658
Cepstral GMM	2.914	1.277
Cepstral (PLP) GMM Supervector	3.419	1.095
Cepstral (MFCC) GMM Supervector	3.683	1.312
MLLR	4.154	1.312
Phone-loop MLLR	4.154	1.972
Prosodic w/ASR	10.016	3.502
State-in-phone Durations	14.820	9.208
Prosodic w/o ASR (poly)	17.180	10.253
Prosodic w/o ASR (supervector)	17.765	12.282
Phone-in-word durations	19.626	8.113
Word N-gram	20.685	7.714

Combined Results

- 4 most important systems (incrementally selected):

  1. Constrained GMM, 2. PLP-SV, 3. Prosody, 4. MLLR

- 4-BEST combination gives result as good as all-system combination
- 4-CEP: combination of ASR-independent cepstral systems:

  Unconstrained GMM, PLP-SV, MFCC-SV, Phone-loop MLLR

Systems (gray = ASR-dependent)	**1-side training**
Constrained cepstral GMM	2.769
Cepstral GMM	2.914
4-BEST	1.954
4-CEP	2.199

- 29% error reduction over single best system
- 11% over cepstral system combination

Summary

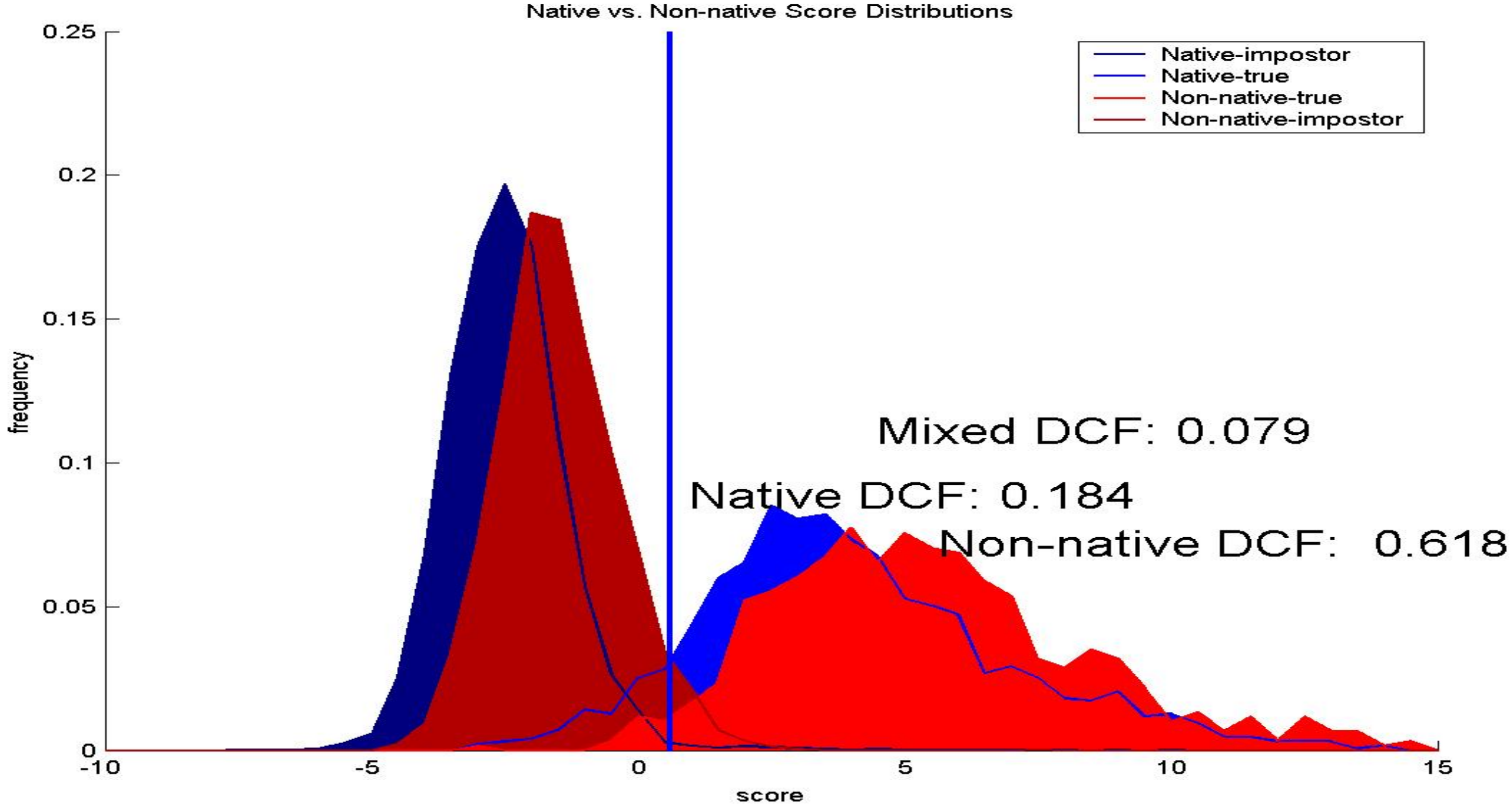
- Presented two very different ways to incorporate higher-level information into cepstral models
  - MLLR feature transforms
  - Conditioning on linguistic units
- Both approaches give excellent results
- MLLR compares very favorably with cepstral GMM and supervector SVM models prior to ISV compensation
- GMM-based systems have improved dramatically with recent factor analysis ISV modeling approach
- New syllable-constrained system currently best cepstral system
- Prosodic and MLLR systems among the 4-best systems selected from over a dozen low- and high-level systems

Nonnativeness Detection

Nonnativeness Detection

- Task: Given speech sample, is talker speaking in his/her native language?
  - This is NOT dialect recognition, but related
- Original motivation: nonnatives show systematic bias in speaker verification scores (next slide)
  - Have since found automatic nonnativeness estimates can reduce speaker id EER by up to 15% (Ferrer et al. ’08b)
- Additional motivations:
  - Intelligence applications
  - Speech recognition (reduce model mismatch)
  - Scientific: effects of L1 on L2
- Results reported in Shriberg et al. (2008)

Nonnativeness and Speaker Verification Scores



- Nonnativeness introduces systematic bias (shift) in scores
- Introduces calibration error in testing

Nonnativeness ID Data Sets

- Fisher-1 English database [broad range of L1s]
  - Extracted balanced native/nonnative subsets
  - 749 nonnatives, 741 natives
  - 1.9 conversations per speaker
  - 10 minutes per conversation (≈ 5 per speaker)
- NIST SRE-06 Mixer [L1= mainly Chinese]
  - Listened to a large subset to find nonnatives
  - 280 native speakers (1604 sides)
  - 315 nonnative speakers (986 sides)
  - 5 minutes per conversation (≈ 2.5 per speaker)

L1 Distribution by Corpus

L1	Fisher (%)	SRE06 (%)
Spanish	17.90	-
Chinese/Mandarin	14.64	**82.77**
Russian	8.05	9.82
Hindi	8.05	0.48
German	3.99	-
Cantonese	3.39	-
Korean	3.33	0.48
French	3.06	-
Arabic	2.59	0.64
Other	1.26	5.79

- Fisher-1 has L1 information
- SRE06 required listening and inference from non-English data

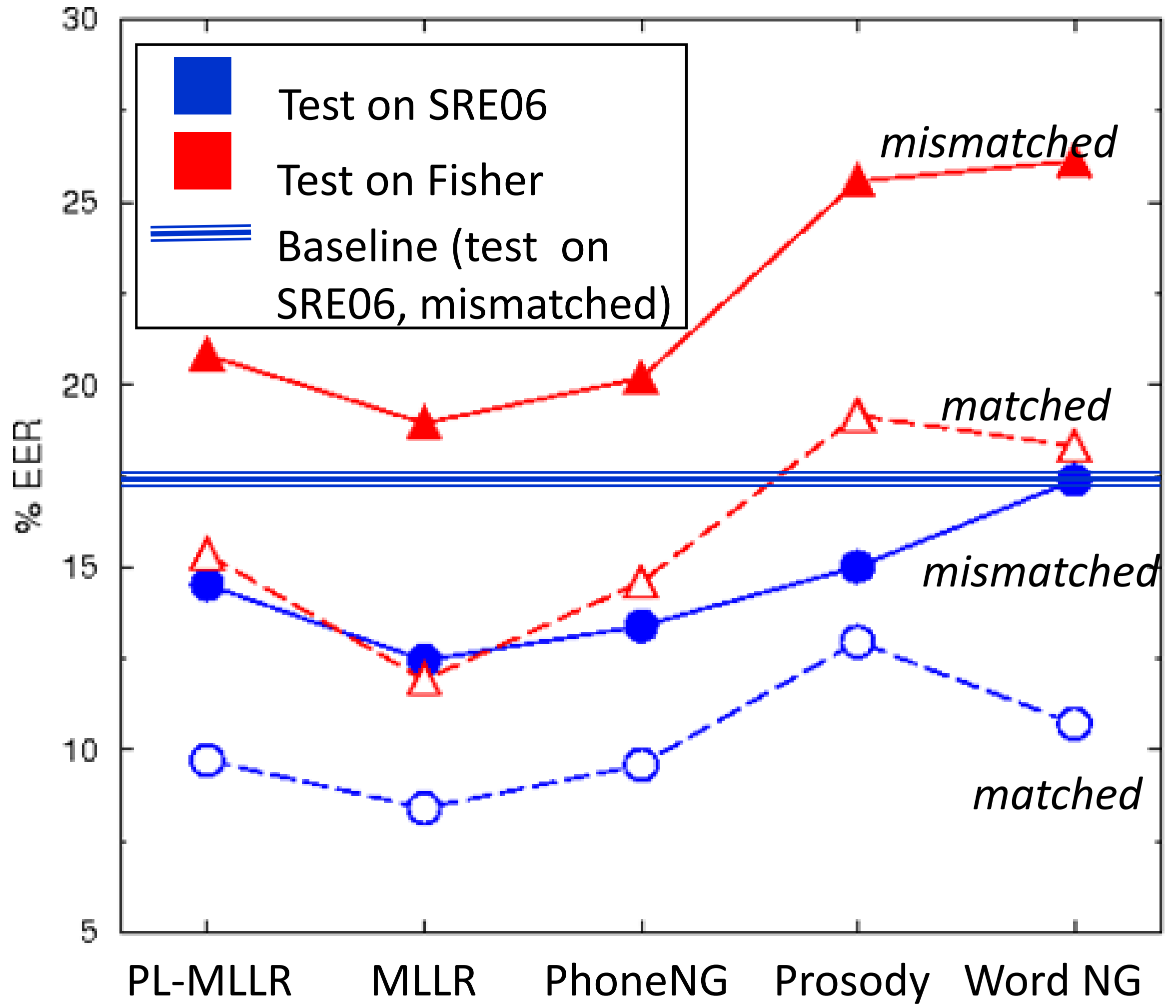
Experiments

- Train binary nativeness classifiers on training set, test on independent test set
- Matched training/test:
  - Training and test from same corpus
  - Speakers divided into 10 partitions
  - Train on 9 and test on 1 partition (round-robin)
- Mismatched training/test:
  - Train on Fisher, test on SRE06, and vice-versa
  - More realistic for real-world applications

Nativeness Detection Models

- Baseline: 1-best phone N-gram LMs (PRLM)
  - Commonly used for language and dialect ID
- SRI SID systems ("out of the box")
  - Lattice-based phone N-gram SVM: *models pronunciation*
  - Phone-loop MLLR SVM: *pronunciation*
  - Word-based MLLR SVM: *pronunciation*
  - SNERF SVM: *prosody (pitch, pause, duration, energy)*
  - Word N-gram SVM: *lexical choice, idioms, grammar*
- No ISV compensation, no score normalization
- Combined system
  - Score-level neural network combiner

Nonnativeness: Results for Individual Systems



- Train and test corpus makeup (in L1s) matter
- Need range of L1s in training
- SID systems perform better or equal to LID baseline
- Combination yields further gains (next)

Nonnativeness Detection: Combination Results

Systems	EER %
Baseline (phone n-gram LM)	17.3
Single best SID system (MLLR)	12.5
2-best combination (MLLR + Prosody)	10.4
3-best combination (MLLR + Prosody + Word-Ngram)	9.3
All 4 (MLLR + Prosody + Word-Ngram + Baseline)	8.6

- Mismatched condition: trained on Fisher, test on SRE06
- Phone N-grams are largely redundant with MLLR system
- Prosody system is most complementary to acoustic models

Nonnativeness Detection: Conclusions

- Speaker modeling techniques work well for nonnativeness ID
- Results mirror those in speaker recognition
  - Relative performance of individual systems
  - Contributions to system combination
  - However: for nonnativeness ID, stylistic models closer to acoustic in absolute performance
- Large effect of corpus mismatch
  - Distribution of test L1s in training is important
- Future work:
  - Inter-speaker variability compensation (NAP or factor analysis)
  - Detect L1 or L1 family
  - Detect speaker's proficiency in L2

Thank you – Questions?

References (1)

A. G. Adami, R. Mihaescu, D. A. Reynolds, and J. J. Godfrey (2003), Modeling Prosodic Dynamics for Speaker Recognition, *Proc. IEEE ICASSP*, vol. 4, pp. 788-791, Hong Kong.

W. D. Andrews, M. A. Kohler, and J. P. Campbell (2001), Phonetic Speaker Recognition, *Proc. Eurospeech, pp. 149–153, Aalborg.*

B. Baker, R. Vogt, and S. Sridharan (2005), Gaussian Mixture Modelling of Broad Phonetic and Syllabic Events for Text-Independent SpeakerVerification, *Proc. Eurospeech*, pp. 2429–2432, Lisbon.

K. Boakye and B. Peskin (2004), Text-Constrained Speaker Recognition on a Text-Independent Task, *Proc. Odyssey Speaker and Language Recognition Workshop*, pp. 129-134, Toledo, Spain.

T. Bocklet and 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 (2004a), 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 (2004b), 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)

M. Ferras, C. C. Leung, C. Barras, and J.-L. Gauvain (2007), Constrained MLLR for Speaker Recognition, *Proc. IEEE ICASSP*, vol. 4, pp. 53-56, Honolulu.

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.

V. R. R. Gadde (2000), Modeling word duration, Proc. ICSLP, pp. 601-604, Beijing.

A. O. Hatch, B. Peskin, and A. Stolcke (2005a), Improved Phonetic Speaker Recognition using Lattice Decoding, *Proc. IEEE ICASSP*, vol. 1, pp. 169-172, Philadelphia.

A. O. Hatch, A. Stolcke, and B. Peskin (2005b), 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.

L. Heck et al. (1998), SRI System Description, NIST SRE-98 evaluation.

S. Kajarekar, L. Ferrer, K. Sonmez, J. Zheng, E. Shriberg, and A. Stolcke (2004), Modeling NERFs for Speaker Recogniition, *Proc. Odyssey Speaker Recognition Workshop*, pp. 51-56, Toledo, Spain.

S. S. Kajarekar (2005), Four Weightings and a Fusion: A Cepstral-SVM System for Speaker Recognition. *Proc. IEEE Speech Recognition and Understanding Workshop*, pp. 17-22, San Juan, Puerto Rico.

Z. N. Karam and W. M. Campbell (2008), A Multi-class MLLR Kernel for SVM Speaker Recognition, *Proc. IEEE ICASSP* pp. 4117-4120, Las Vegas.

References (3)

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.

D. Klusacek, J. Navrátil, D. A. Reynolds, and J. P. Campbell (2003), Conditional pronunciation modeling in speaker detection, *Proc. IEEE ICASSP*, vol. 4, pp. 804-807, Hong Kong.

J. Navrátil, Q. Jin, W. D. Andrews, and J. P. Campbell (2003), Phonetic Speaker Recognition Using Maximum-Likelihood Binary-Decision Tree Models, *Proc. IEEE ICASSP*, vol. 4, pp. 796-799, Hong Kong.

A. Park and T. J. Hazen (2002), ASR Dependent Techniques for Speaker Identification, *Proc. ICSLP*, pp. 1337–1340, Denver.

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.

E. E. Shriberg (2007), Higher Level Features in Speaker Recognition, in C. Müller (Ed.) *Speaker Classification I.* Volume 4343 of Lecture Notes in Computer Science / Artificial Intelligence. Springer: Heidelberg / Berlin / New York, pp. 241-259.

E. Shriberg and L. Ferrer (2007), A Text-Constrained Prosodic System for Speaker Verification, *Proc. Eurospeech*, pp. 1226-1229, Antwerp.

E. Shriberg, L. Ferrer, S. Kajarekar, N. Scheffer, A. Stolcke, and M. Akbacak (2008), Detecting Nonnative Speech Using Speaker Recognition Approaches. *Proc. Odyssey Speaker and Language Recognition Workshop*, Stellenbosch, South Africa.

References (4)

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

K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub (1998), Modeling Dynamic Prosodic Variation for Speaker Verification, *Proc. ICSLP*, pp. 3189-3192, Sydney.

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

A. Stolcke, S. Kajarekar, L. Ferrer, and E. Shriberg (2007), Speaker Recognition with Session Variability Normalization Based on MLLR Adaptation Transforms, *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7), 1987-1998.

A. Stolcke and S. Kajarekar (2008), Recognizing Arabic Speakers with English Phones. *Proc. Odyssey Speaker and Language Recognition Workshop*, Stellenbosch, South Africa.

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

D. E. Sturim, D. A. Reynolds, R. B. Dunn, and T. F. Quatieri (2002), Speaker Verification Using Text-Constrained Gaussian Mixture Models, *Proc. IEEE ICASSP*, vol. 1, pp. 677-680, Orlando.

G. Tur, E. Shriberg, A. Stolcke, and S. Kajarekar (2007), Duration and Pronunciation Conditioned Lexical Modeling for Speaker Recognition, *Proc. Eurospeech*, pp. 2049-2052, Antwerp.

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

M. A. Zissman and E. Singer (1994), Automatic language identification of telephone speech messages using phoneme recognition and N-gram modeling, *Proc. IEEE ICASSP*, vol. 1, pp. 305-308, Adelaide.