Extending ATVS Core:

description of ATVS'05 submission

Outline

1. Introduction
2. Description of acoustic systems
   - 2.1. GMM
   - 2.2. SVM
   - 2.3. KL-Tnorm
3. Description of higher level systems
   - 3.1. Phonetic
   - 3.2. Prosodic
4. Systems fusion
5. Conclusion

1. Introduction: ATVS 04/05 Submissions

2004		2005	
1conv/1conv	8conv/1conv	1conv/1conv	8conv/1conv
		KL-SVM	KL-SVM
GMM (LR-GMM)		KL-GMM	KL-GMM
			Prosodic
			Phone3gram
			SVM Fusion

Development Set

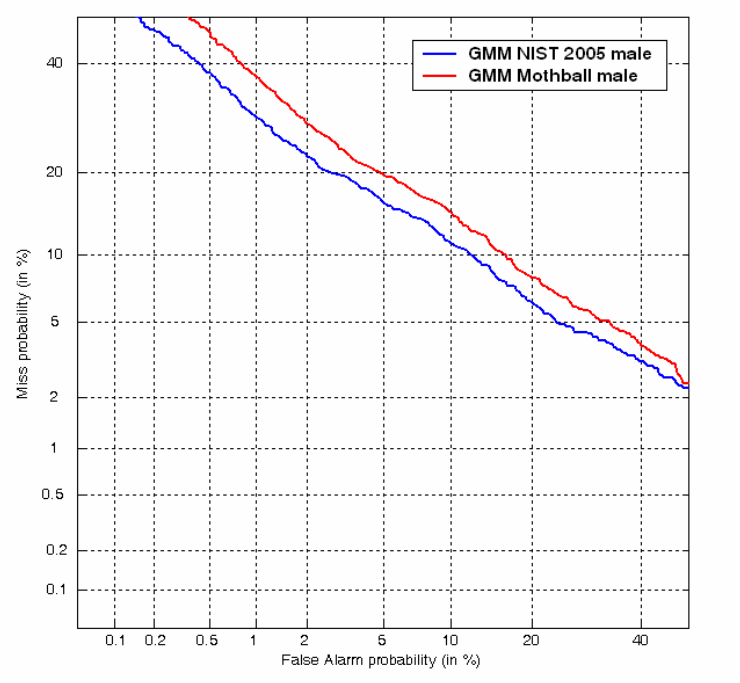
- One Development Set:
  - NIST SRE 2004 Corpus
- Two background and auxiliary data sets
  - NIST SRE 2004 corpus (top matching with DevSet)
  - Switchboard-I and past NIST SREs (mismatch with DevSet)
- All development trials: NIST SRE 2004 trial lists

2. Acoustic Systems

KL-GMM and KL-SVM

2.1. GMM System Evolution

- Mothball 2004 vs. 2005



- Improvements due to:
  - KL-TNorm
    - Automatic cohort selection method for TNorm
  - Database matching conditions (UBM, TNorm)
    - 2004: Swb-I and past NIST SREs
    - 2005: MIXER data

GMM Baseline System

- Feature Extractor:
  - 19 MFCC + delta = 38 features
  - CMN + Rasta + Feature Warping
- Model parameters
  - 1024 mixtures
  - MAP-UBM trained with NIST 2004 (MIXER) data
- TNorm and KL-TNorm
  - Cohorts: NIST 2004 target models
  - Gender-Dependent

2.2. SVM Baseline System

- GLDS-SVM with 2nd order explicit polynomial expansion
- "P" matrix channel compensation
  - null space of 3 dimensions (one per channel)
    - A. Solomonoff, C. Quillen and W.M. Campbell, *Channel compensation for SVM Speaker Recognition,* Proceedings Odyssey'04.
- Explicit normalized two degree polynomial expansion
  - V. Wan, W.M. Campbell, *Support vector machines for speaker verification and identification,* Proceedings 2000 IEEE Signal Processing Society workshop.
- Decomposed GLDS Kernel
  - W.M. Campbell, *Generalized linear discriminant sequence kernels for speaker recognition,* Proceedings ICASSP '02.
- The Speaker model is just a explicit W hyperplane
  - No need to store Support Vectors

SVM Baseline System

- KL-TNorm
  - Cohorts: NIST 2004 target models
  - Gender-Dependent
- GLDS fully new in Matlab: limited results due to:
  - 2nd order pol. exp. forced by Matlab memory limitations
  - Strong subsampling required
  - High control of the different stages in SVM system

2.3. KL-TNorm

- Fast Adaptive TNorm Cohort Selection Method
- Very fast distance computation method based on an upper-bound for Kullback-Leibler divergence
  - M. N. Do, “Fast Aproximation of Kullback-Leibler Distance for Dependence Trees and Hidden Markov Models”, Signal Pocessing Letters, Volume 10(4), April 2003, pp:115 – 118
- Valid only for HMM and Dependence Trees
  - Distances computed from GMM models and cohorts
  - Distances used for cohort selection both in GMM and SVM systems
    - D. Ramos-Castro et al., “KL-Tnorm: Speaker Verification Using Kullback Leibler Divergence-based Fast Adaptative Tnorm", Biometrics on the Internet, Third COST 275 Workshop, Hatfield, UK, 2005 (submitted).

KL-TNorm

- Two GMM distributions: $f = \sum_{i=1}^{n} w_i f_i \quad \hat{f} = \sum_{i=1}^{n} \hat{w}_i \hat{f}_i$
- KL Distance:

$$D\left(\sum_{i=1}^{n} w_i f_i \middle| \sum_{i=1}^{n} \hat{w}_i \hat{f}_i\right) = \int \sum_{i=1}^{n} w_i f_i \log \frac{\sum_{i=1}^{n} w_i f_i}{\sum_{i=1}^{n} \hat{w}_i \hat{f}_i} \leq \int \sum_{i=1}^{n} \left[w_i f_i \log \frac{w_i f_i}{\hat{w}_i \hat{f}_i} \right] =$$

$$= \sum_{i=1}^{n} w_i \log \frac{w_i}{\hat{w}_i} + \sum_{i=1}^{n} \int f_i \log \frac{f_i}{\hat{f}_i}$$

Log-Sum Inequality

- KL Distance between 2 d-dimensional Gaussians:

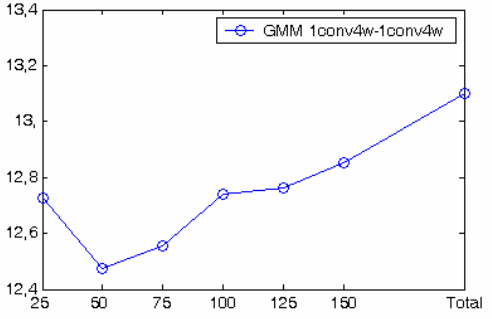
$$\int f_i \log \frac{f_i}{\hat{f}_i} = \frac{1}{2}\left[\log \frac{\det(\Sigma_i)}{\det(\hat{\Sigma}_i)} - d + tr\left(\hat{\Sigma}_i^{-1}\Sigma_i\right) + (\mu_i - \hat{\mu}_i)\hat{\Sigma}_i^{-1}(\mu_i - \hat{\mu}_i) \right]$$

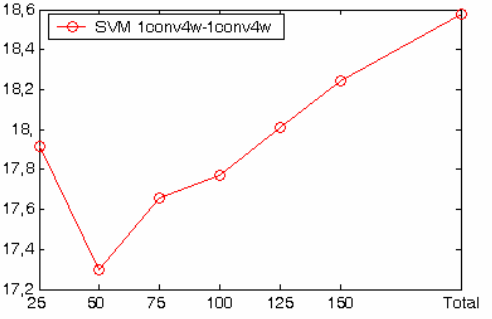
KL-TNorm: DevSet Experiments (I)

- Effect of cohort size (N vs. EER)

EER DET KL-TNorm NIST 2004 DevSet

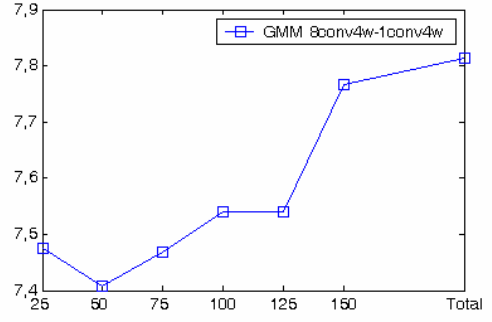
GMM 1c-1c

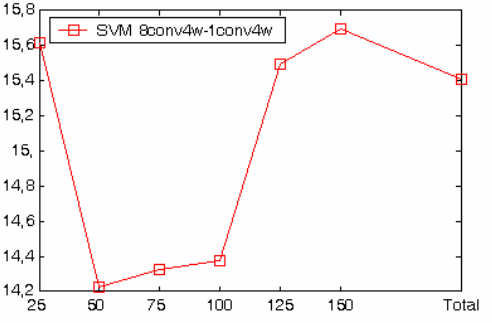




SVM 1c-1c

GMM 8c-1c

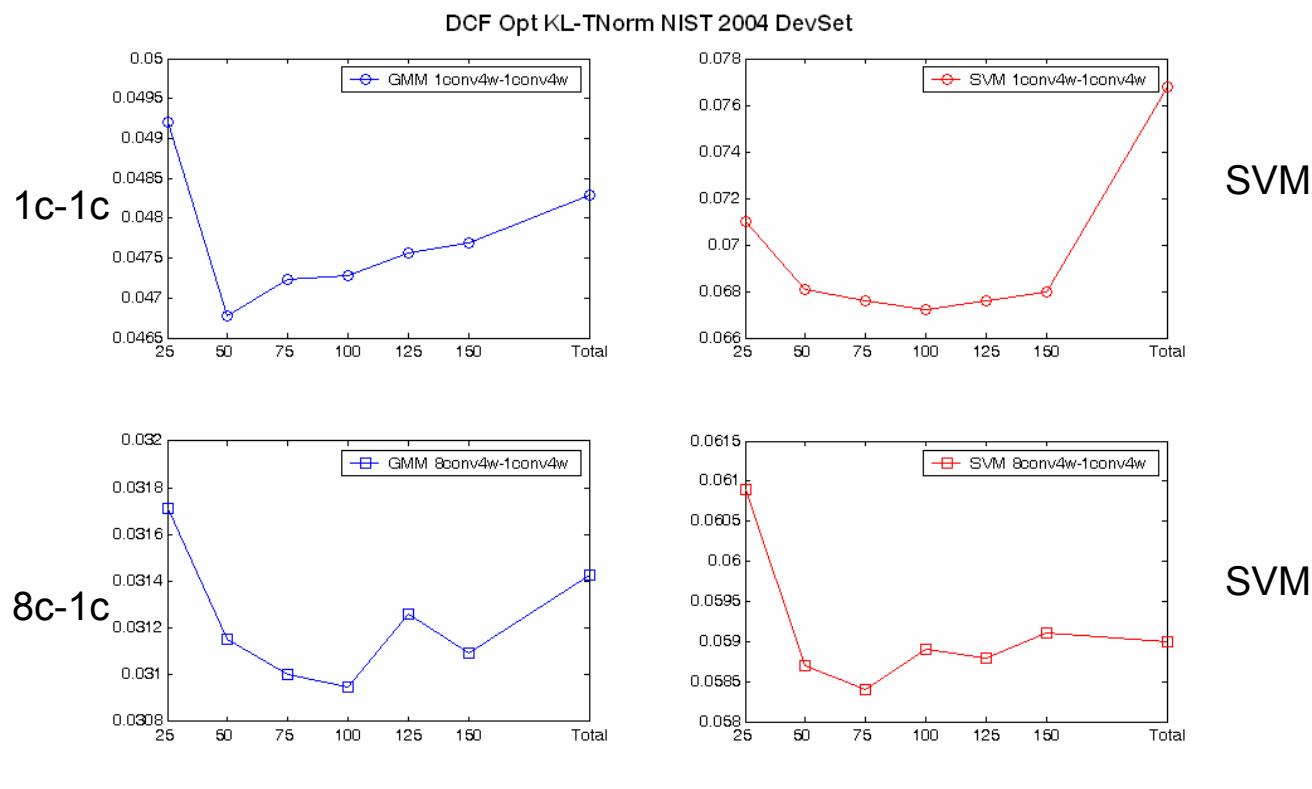




SVM 8c-1c

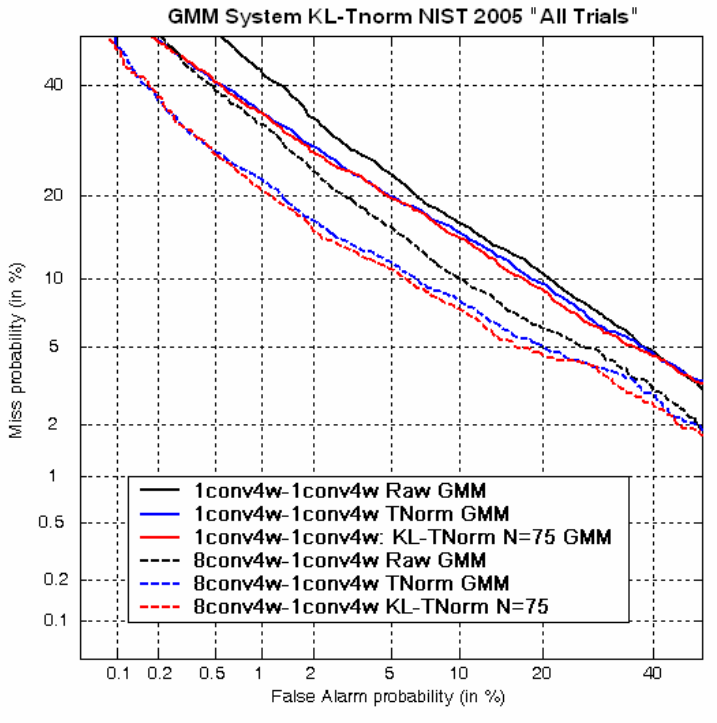
KL-TNorm: DevSet Experiments (II)

- Effect of cohort size (N vs. DCF)

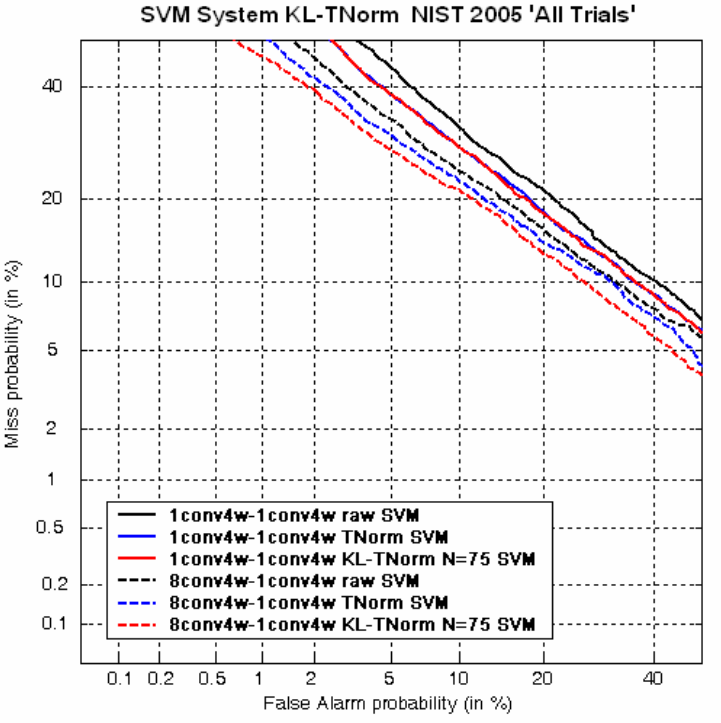


KL-TNorm: NIST 2005 Performance

- GMM system (N=75)



- SVM system (N=75)



3. ATVS High-Level Systems

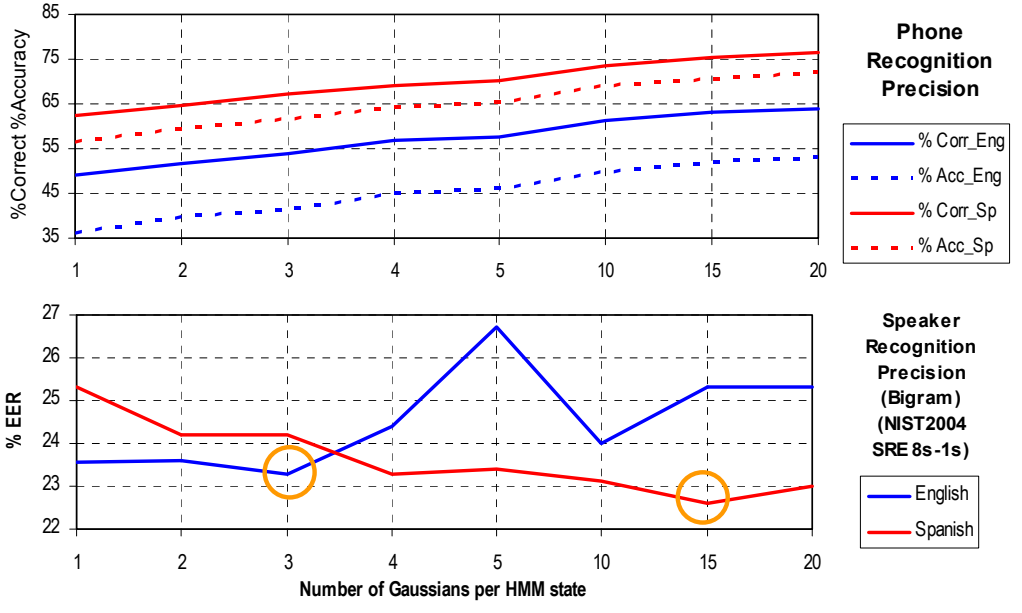
Phonetic & Prosodic Speaker Recognition

3.1. Phonetic Speaker Recognition
Acoustic-Phonetic Decoders

- 2 Languages:
  - English: TIMIT (8 kHz), 39 phones.
  - Spanish: ALBAYZIN (8 kHz), 23 phones.
- Advanced Distributed Speech Recognition Standard Front-End
  - ETSI ES 202 050: noise and channel robust, MFCC.
- Context, speaker & gender independent HMM phone models (HTK)
  - 3-state, left-to-right with no skips
  - 1 to 20 Gaussians/state
- N-gram (2/3/4) modeling and scoring
  - 1 UBPM per language (eng/sp), SRE'04 training data
  - 1 SPM per target (8conv), trained from scratch/adapted from UBPM

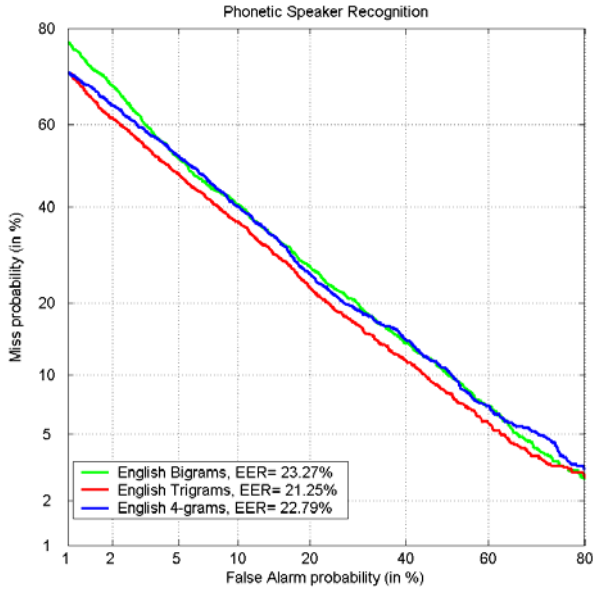
$$Score_i = \frac{1}{N}\log\left(\frac{P(X \mid SPM_i)}{P(X \mid UBPM)}\right)$$

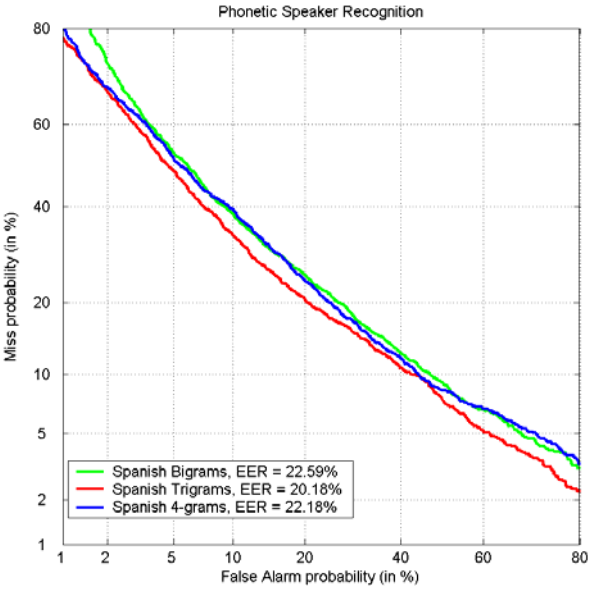
Gaussians per State in Phone HMMs



- HMM complexity and speaker recognition precision not clearly related
  - Very simple (and error-prone) phonetic decoders may be useful for SR
- English: 3 Gauss/state; Spanish: 15 Gauss/state

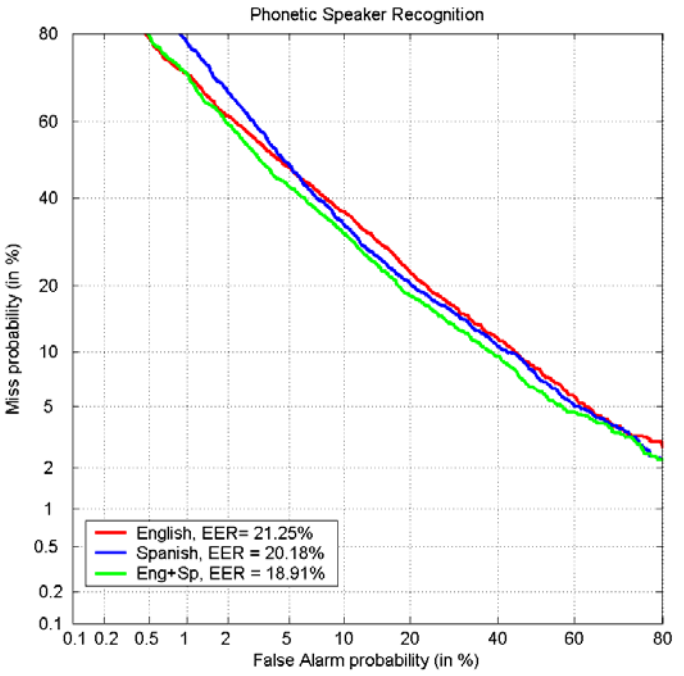
Bigrams, Trigrams or 4-grams





- SPM training method:
  - Bigrams: SPMs trained from scratch
  - Trigrams and 4-grams: SPMs adapted from UBPMs
- 4-grams perform worse than trigrams and similar to bigrams
- BEST MODELS: Trigrams adapted from the UBPM

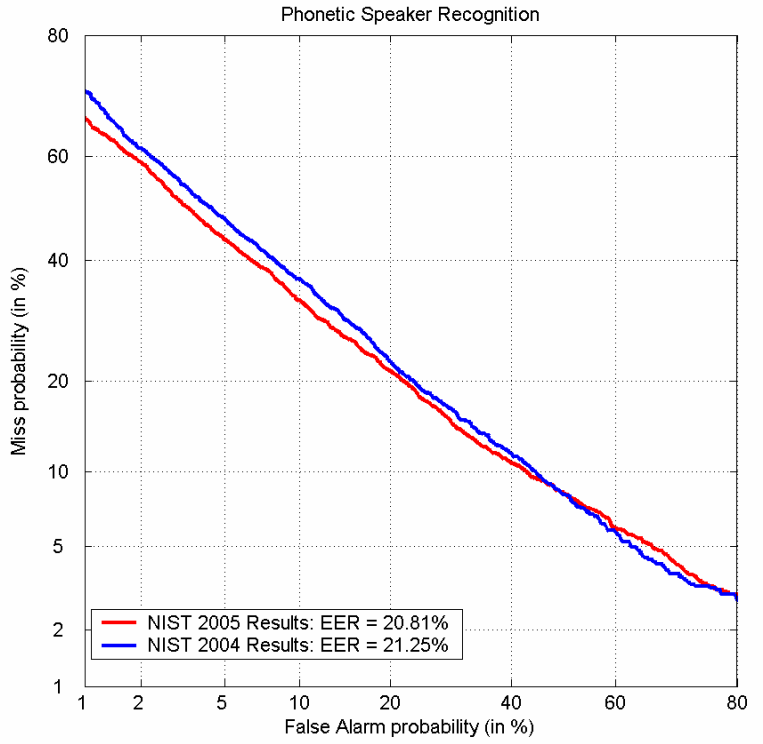
English, Spanish or Sum



- Spanish phonetic decoding
  - Better EER than English
  - Worse for ↓ False Alarm
- Sum fusion
  - Best EER
  - Similar to English for ↓ False Alarm
- We use only English decodings

- FINAL PHONETIC SPEAKER RECOGNITION SYSTEM:
  - English phonetic decodings
  - HMMs with 3 Gaussians/state
  - Trigram adapted from UBPM
  - Weight of UBPM: 0.7

Phonetic Results: Dev'04 vs. Eval'05

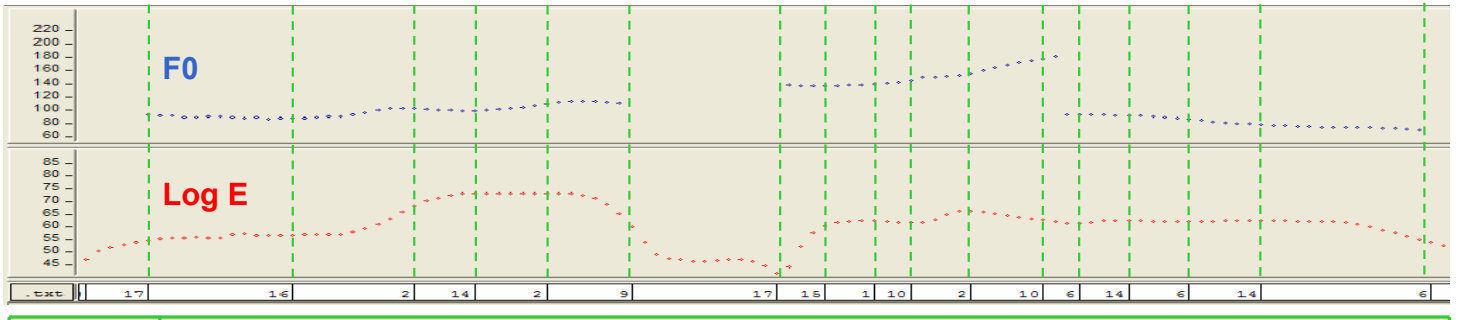


- UBPM trained on NIST'04 data
- System tested on NIST'04 and NIST'05 data
- Final results are:
  - Slightly better for NIST'05 data
    - No degradation dev → eval
  - Similar to NIST'04 state of the art phonetic speaker recognition subsystems

3.2. Prosodic Speaker Recognition

Four-Level Delta-based Tokenization*

1. Compute the delta features (50 ms) for F0 and energy contours
2. Detect the changes in the dynamics based on the delta features
3. Generate new segments using the detected points
4. Four-level quantization of each segment based on the slopes



TOKEN	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
FO	+F	+F	+S	+S	-F	-F	-S	-S	+F	+F	+S	+S	-F	-F	-S	-S	UV
E	+F	+S	+F	+S	-F	-S	-F	-S	-F	-S	-F	-F	+F	+S	+F	+S	*

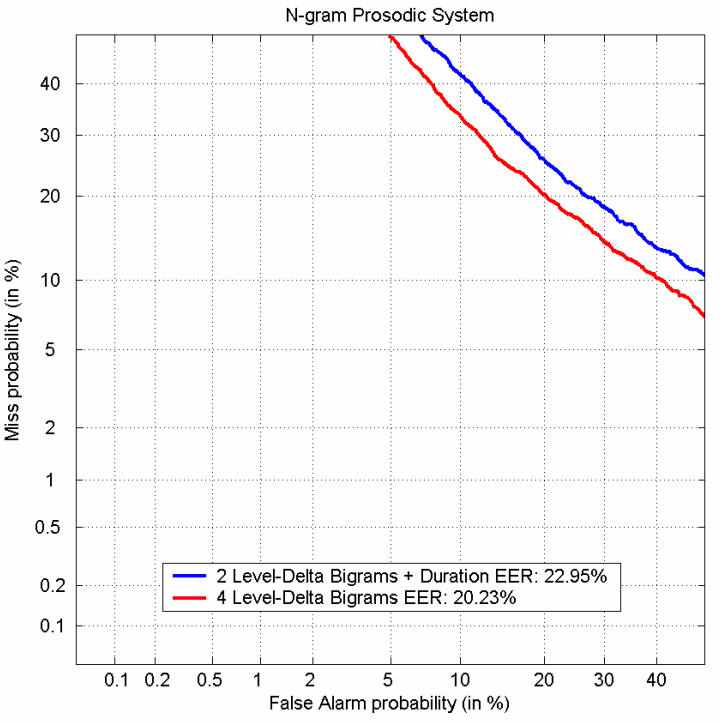
+F=Fast-rising; +S=Slow-rising; -F=Fast-falling; -S=Slow-falling; UV=Unvoiced

*Adapted from: Adami, A. "Modeling Prosodic Differences for Speaker and Language Recognition ", Ph.D Thesis, OGI 2004

N-gram Modeling and Score

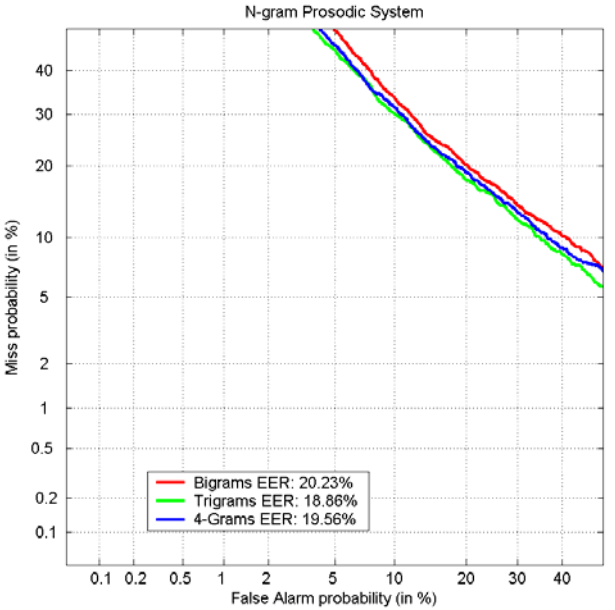
- Gender-dependent UBMs trained using HTK LM tools
  - Data from NIST 2004 Extended-data task
- Speaker models created by interpolation:
  - Spk_model=0.2 UBM + 0.8 Spk_data_model
  - Speaker models include the “general knowledge” of the UBM and the “specific knowledge” of the speaker data
- Scores: conventional log-likelihood ratio test
- No score normalization techniques applied

Quantization selection



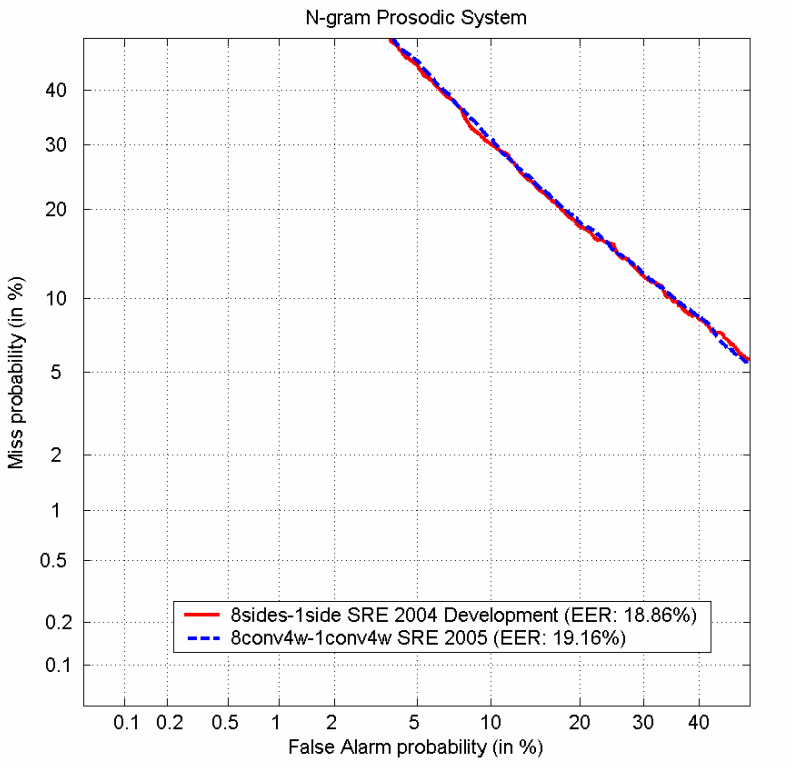
- Much better performance for the 4 Level-Delta tokenization
- Over-segmentation of the contours affects the duration information

Bigrams, Trigrams or 4-grams



- Speaker model training method:
  - Bigrams: Speaker models trained from scratch
  - Trigrams and 4-grams: Speaker models adapted from UBMs
- Trigrams slightly better than 4-grams and bigrams
- BEST MODELS: Trigrams adapted from the UBM

Prosodic Results: Dev'04 vs. Eval'05



- Final results are similar for NIST'04 and NIST'05 data
  - No degradation dev → eval
- Also similar to NIST'04 state of the art prosodic speaker recognition subsystems

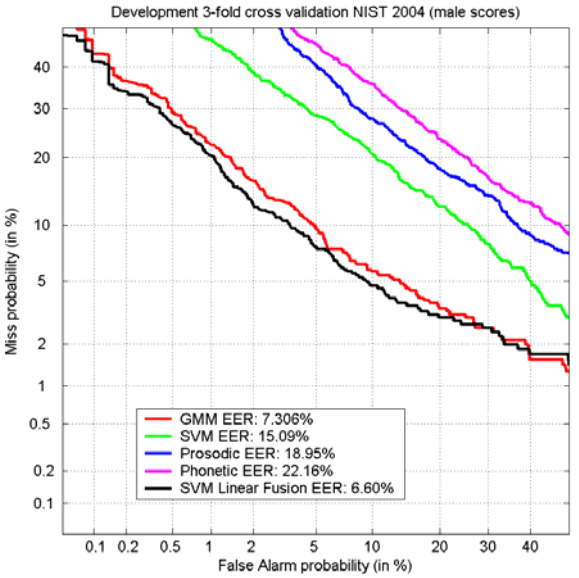
4. System Fusion

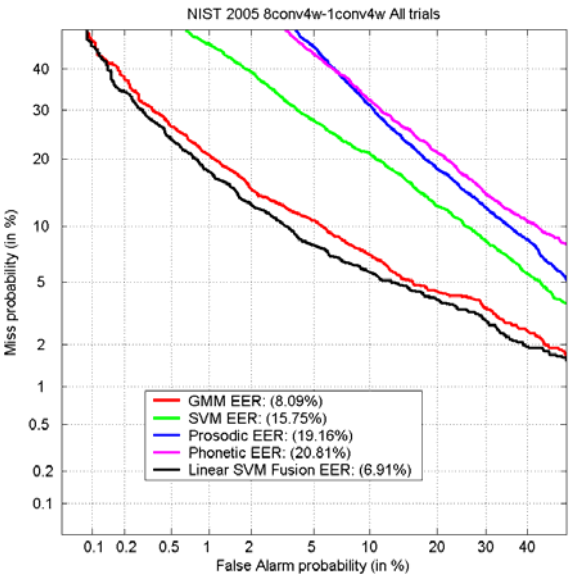
8conv4w-1conv4w task

4.1. SVM Fusion
SVM configuration

- Linear kernel:
  - FR errors 10 times more costly than FA
  - Compensate for the amount of target and impostor trials
- Fusion of four systems
  - (KL-GMM, KL-SVM, Prosodic Trigram, Phone Trigram)
- Threshold selection based on NIST 2004 development data
  - 3-fold cross validation from 8sides-1side task (male trials)

Dev'04 and Eval'05 Results





- Overall system performance dominated by the GMM system
- Consistency between Development and Evaluation data
- Need to improve High-level systems performance to increase their contribution

Conclusion

Conclusions

- GLDS-SVM system: from scratch in Matlab
  - Good process control but memory limitations
  - Submitted system: just 2nd order feature expansion
  - Now porting Matlab-based GLDS-SVM to ATVS C++ core
- KL-Tnorm excellent performance (GMM & SVM):
  - Accuracy and speed
  - Specially in 8conv-1conv task
- Phonetic and prosodic performed excellent
- Succesful submission of three ATVS brand-new core technologies (SVM, Phonetic and Prosodic)