

---

# **Advances in Speaker Recognition: Getting to Know You**

**Joseph P. Campbell, et al.**  
j.campbell@ieee.org

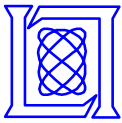
**MIT Lincoln Laboratory  
Lexington, MA**

**Biometric Consortium Conference  
Arlington, VA**

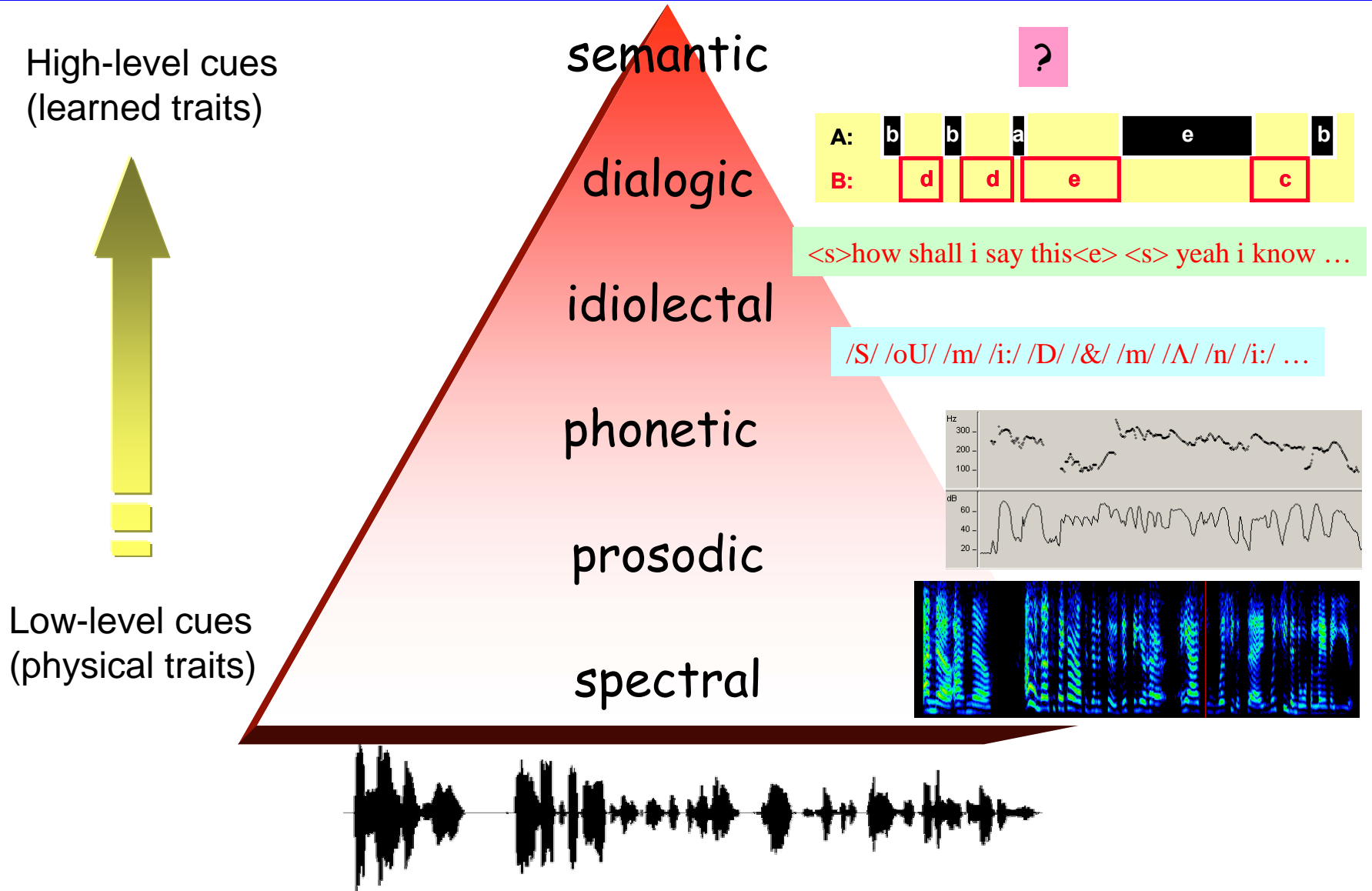
**23 September 2003**

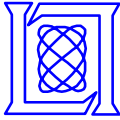
---

This work was sponsored by the Department of Defense under Air Force contract F19628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.



# Exploiting Higher-Levels of Information





# Growing area of research

---

- **SuperSID Johns Hopkins U. CLSP 2002 Summer Workshop**
  - NSF sponsored international team from research labs & academia
  - <http://www.clsp.jhu.edu/ws2002/groups/supersid/>
- **Third year in NIST Speaker Recognition Evaluation**
  - Extended data task
  - More data is desperately needed to fully pursue this research
  - <http://www.nist.gov/speech/tests/spk/>



# Outline

---

spectral

- Gaussian mixture model: GMM-UBM
- Support vector machine (SVM)
- Text Constrained GMM-UBM

prosodic

- Pitch and energy distributions
- Pitch and energy gestures

- GMM mixture tokens

phonetic

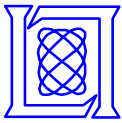
- Phone n-gram
- Phone SVM
- Conditional pronunciations

idiolectal

- Word n-gram

dialogic

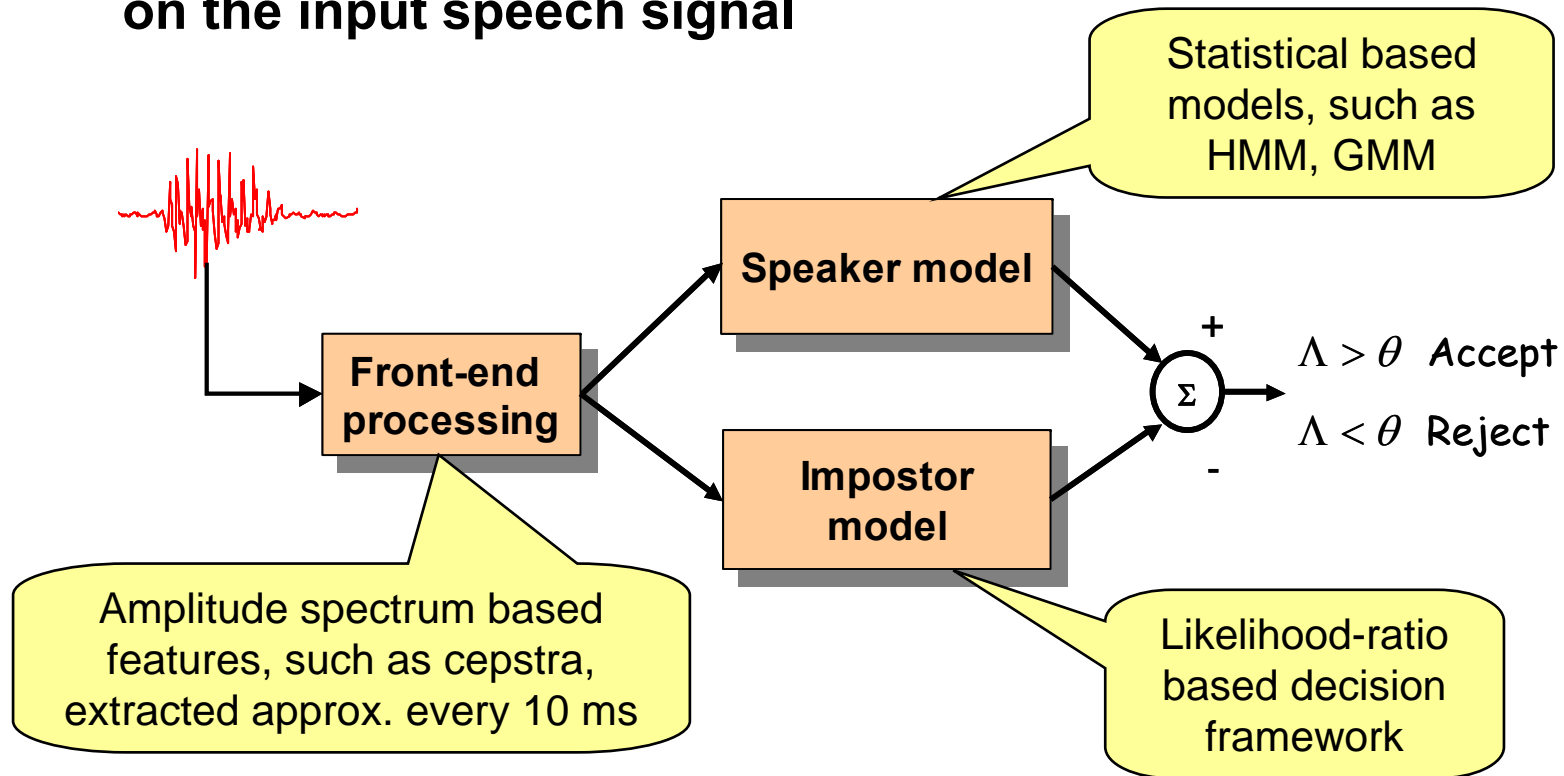
- Conversational patterns
- Fusion
- Conclusions



# Spectral Features

## GMM with Cepstra

- **State-of-the-art speaker recognition algorithms are based on statistical models of short-term acoustic measurements on the input speech signal**



- **Acoustic information is very powerful for speaker recognition, but it can be sensitive to channel distortion**

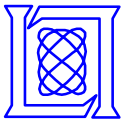


# Prosodic Features

## Pitch and Energy Distributions

---

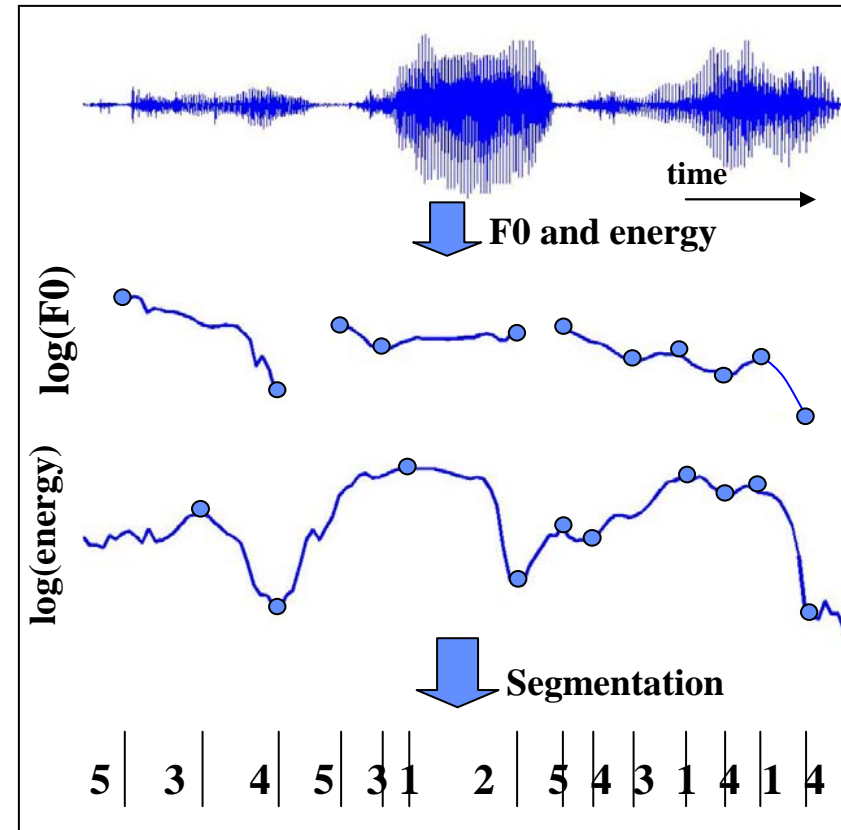
- **Prosodic features are measures of stress, accent, and intonation**
  - Pitch
  - Energy
  - Duration
- **Traditionally, feature vectors of energy and pitch are used**
  - GMM-UBM classifier
  - Feature vector is concatenation of log pitch, log energy and their deltas (+-5 frames)
  - 512 mixture UBM trained on held-out splits
- **Newer methods are looking for long-term characteristics**



# Prosodic Features

## Pitch and Energy Gestures

- Approach: Model pitch and energy track dynamics\*
- Create a sequence of symbols describing the slope of the pitch and energy tracks
  - Segment at inflection points in each sequence
- Tag with quantized duration
  - Voiced Segments:
    - Short  $\leq 80\text{ms}$
    - Long  $> 80\text{ms}$
  - Unvoiced Segments
    - Short  $\leq 140\text{ms}$
    - Long  $> 140\text{ms}$
- Can also tag with the phone or word context occurred
- Apply n-gram modeling to sequence



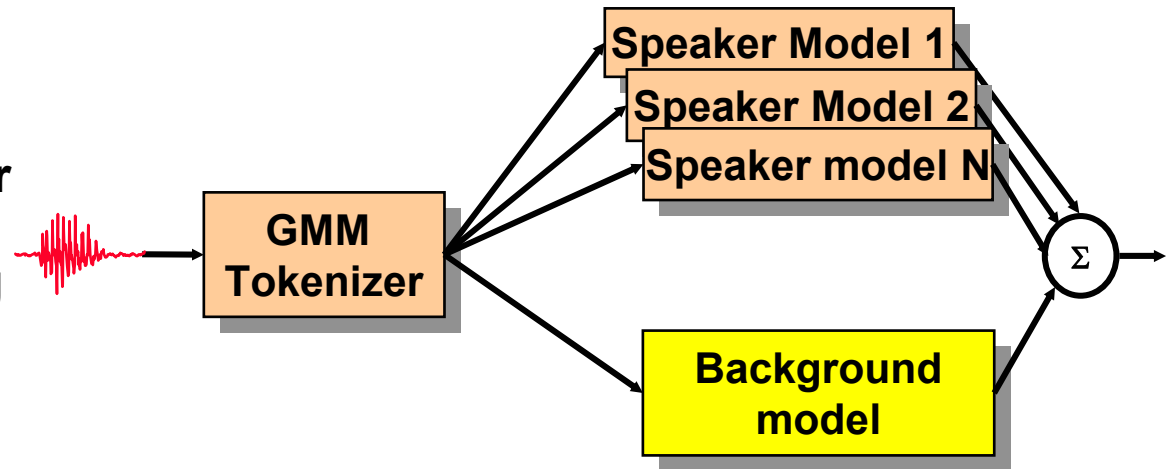
Class	Temporal Trajectory
1	rising f0 and rising energy
2	rising f0 and falling energy
3	falling f0 and rising energy
4	falling f0 and falling energy
5	unvoiced segment



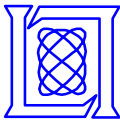
# Sub-Phonetic Features

## GMM-tokenizer

- Based on previous work on LID
  - 2048 model for tokenizer
  - Bigram model based on sequence of top scoring mixture component at each frame



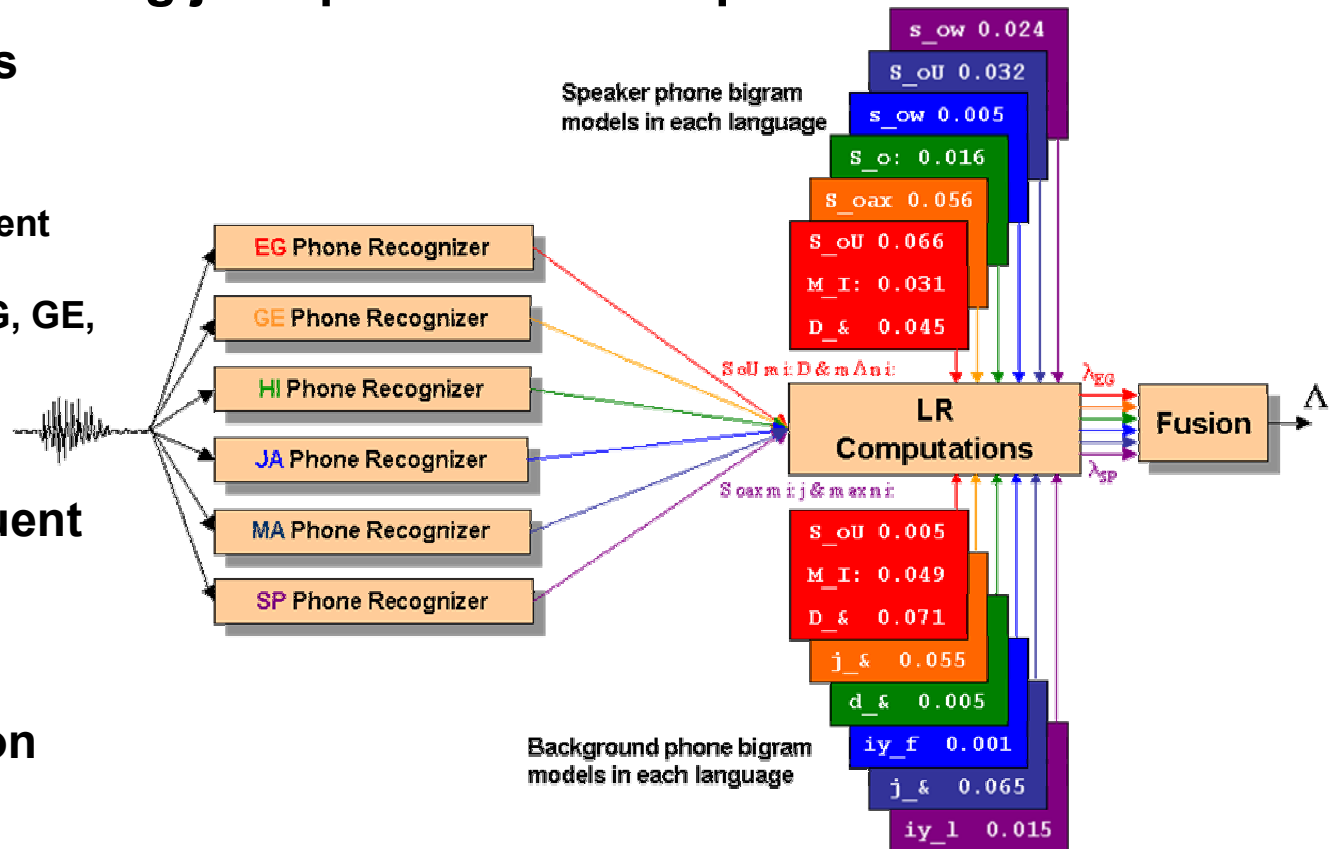
- Tokenizer and background model trained using splits 1-5 for testing utterances in splits 6-10 and vice versa
- Testing uses a minimum number of observations (background + speaker) criteria for each number of conversations condition (c\_min 249-2999 for 1-16 conv)

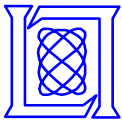


# Phonetic Features

## Phone N-grams

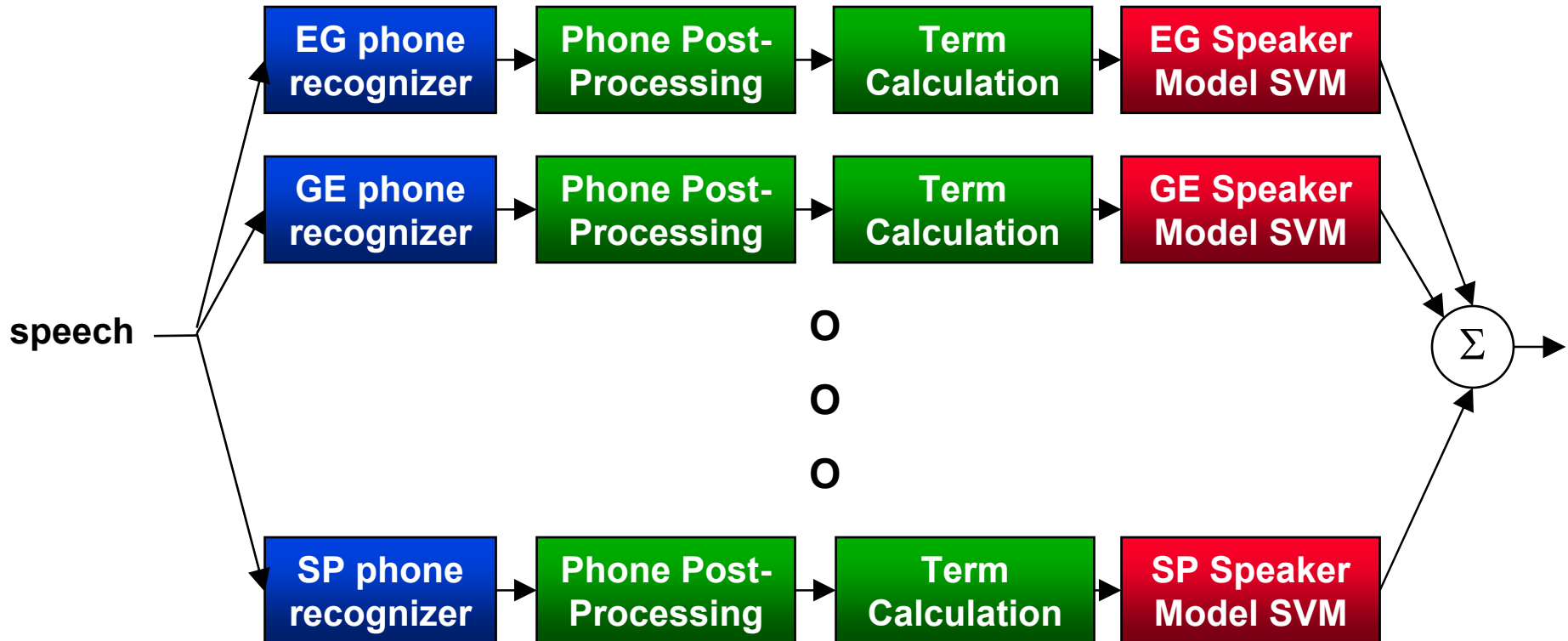
- Aim: Learn speaker-dependent pronunciations using phone streams (sequences of sound units)
- Approach: Apply n-gram modeling to multiple open-loop phone streams using joint probabilities of phones
- Most experiments conducted with PPRLM phones
  - Gender-dependent phone models
  - 5 languages (EG, GE, JA, MA, SP)
- Prune out infrequent n-grams
- Language phone scores fused via linear combination

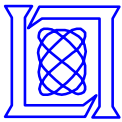




# Phonetic Features

## Phone Support Vector Machine





# Phonetic Features

## Conditional Phone Probabilities

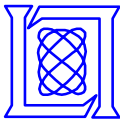
- Aim: Learn speaker-dependent pronunciations by aligning constrained ASR phonemes with open-loop phones
- Approach: Align ASR word phonemes with open loop (OL) phones at frame level and compute conditional probabilities

ASR and open loop phone alignment

WORD	TIME	ASR	EG	GE	SP	JA	MA
	24964	t	n	n	n	sh	N
	24965	t	s	h	s	sh	N
	24966	t	s	h	s	sh	N
	24967	t	s	h	s	sh	S
TO	24968	t	s	h	s	sh	S
	24969	t	s	h	s	sh	S
	24970	t	s	h	s	rx	S
	24971	ax	l	h	s	rx	i:
	24972	ax	l	h	iy	rx	i:
	24973	ax	l	h	iy	y	i:

$$\Pr(\text{OL\_phone} \mid \text{ASR\_phoneme}, \text{speaker}) = \frac{\#(\text{OL\_phone}, \text{ASR\_phoneme})}{\#(\text{ASR\_phoneme})}$$

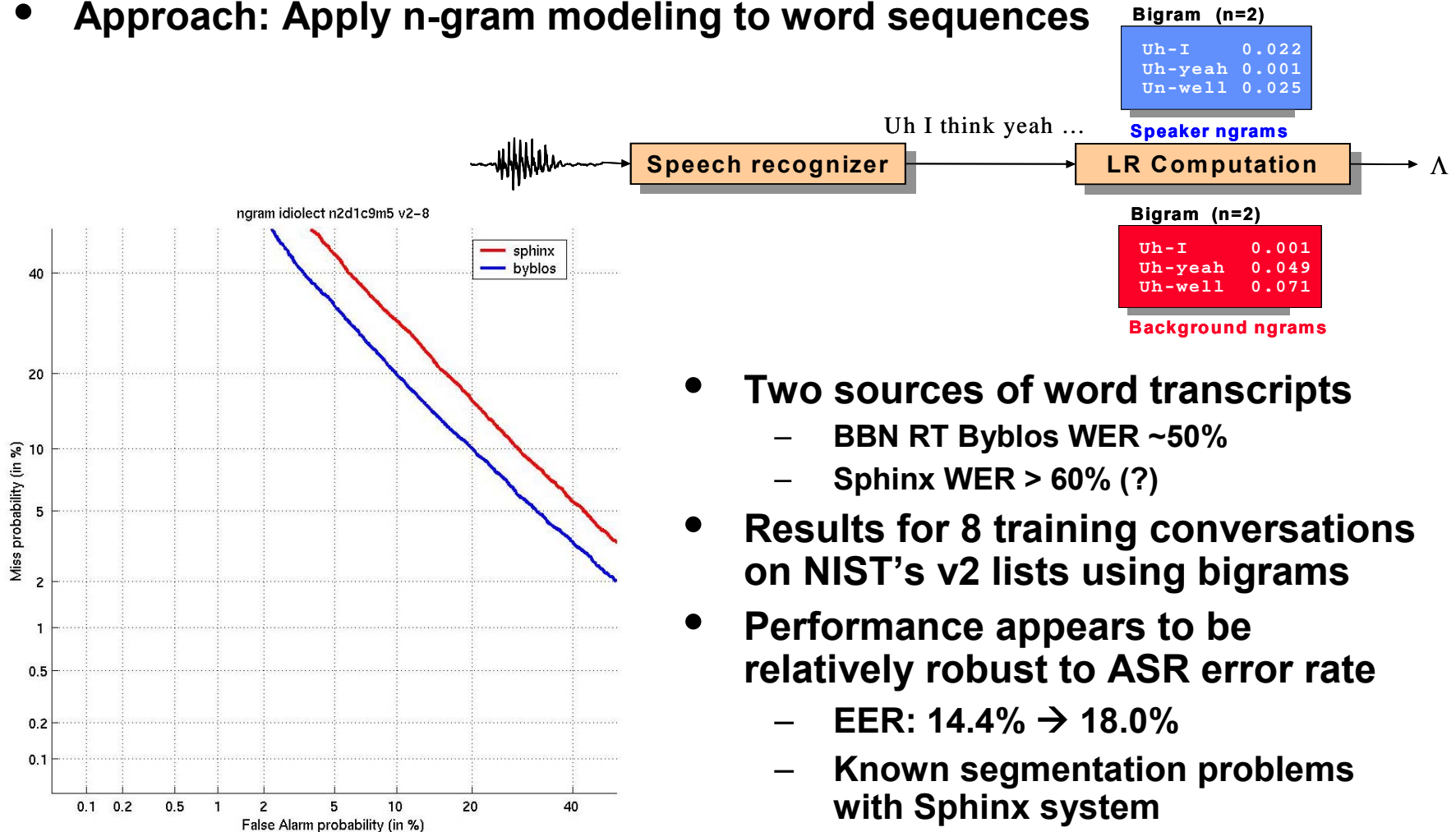
- During scoring compute likelihood of observed (OL\_phone, ASR\_phoneme) sequence against speaker and background models
- Scores from five OL phone streams linearly combined



# Lexical Features

## Word N-grams

- **Aim: Learn speaker-dependent word usage (idiolect)**
- **Approach: Apply n-gram modeling to word sequences**



- **Two sources of word transcripts**
  - BBN RT Byblos WER ~50%
  - Sphinx WER > 60% (?)
- **Results for 8 training conversations on NIST's v2 lists using bigrams**
- **Performance appears to be relatively robust to ASR error rate**
  - **EER: 14.4% → 18.0%**
  - **Known segmentation problems with Sphinx system**

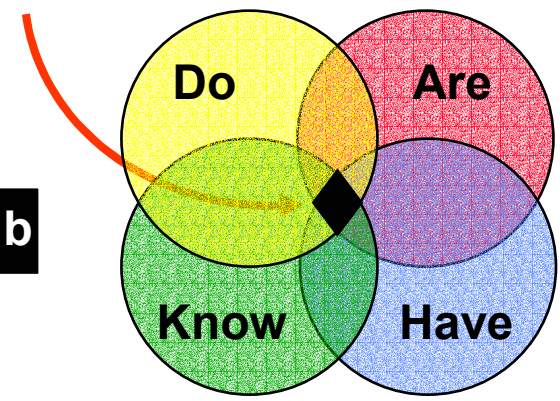
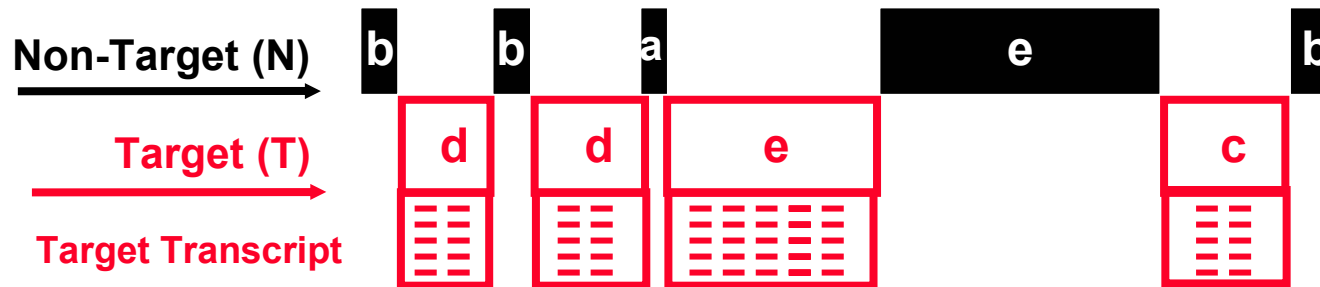


# Dialogic Features

## Conversation Pattern N-grams

Strongest authentication

### Turn Durations



Encoding:

Nb **Tdhf** Nb Tdhf Na Teig Ne Tcge Nb

# word label	a	b	c	d	e	f	...
# char label	1	2	4	8	16	32	
duration label							
Target/Non-Target							



# Outline

---

spectral

- Gaussian mixture model: GMM-UBM
- Support vector machine (SVM)
- Text Constrained GMM-UBM

prosodic

- Pitch and energy distributions
- Pitch and energy gestures
  
- GMM mixture tokens

phonetic

- Phone n-gram
- Phone SVM
- Conditional pronunciations

idiolectal

- Word n-gram

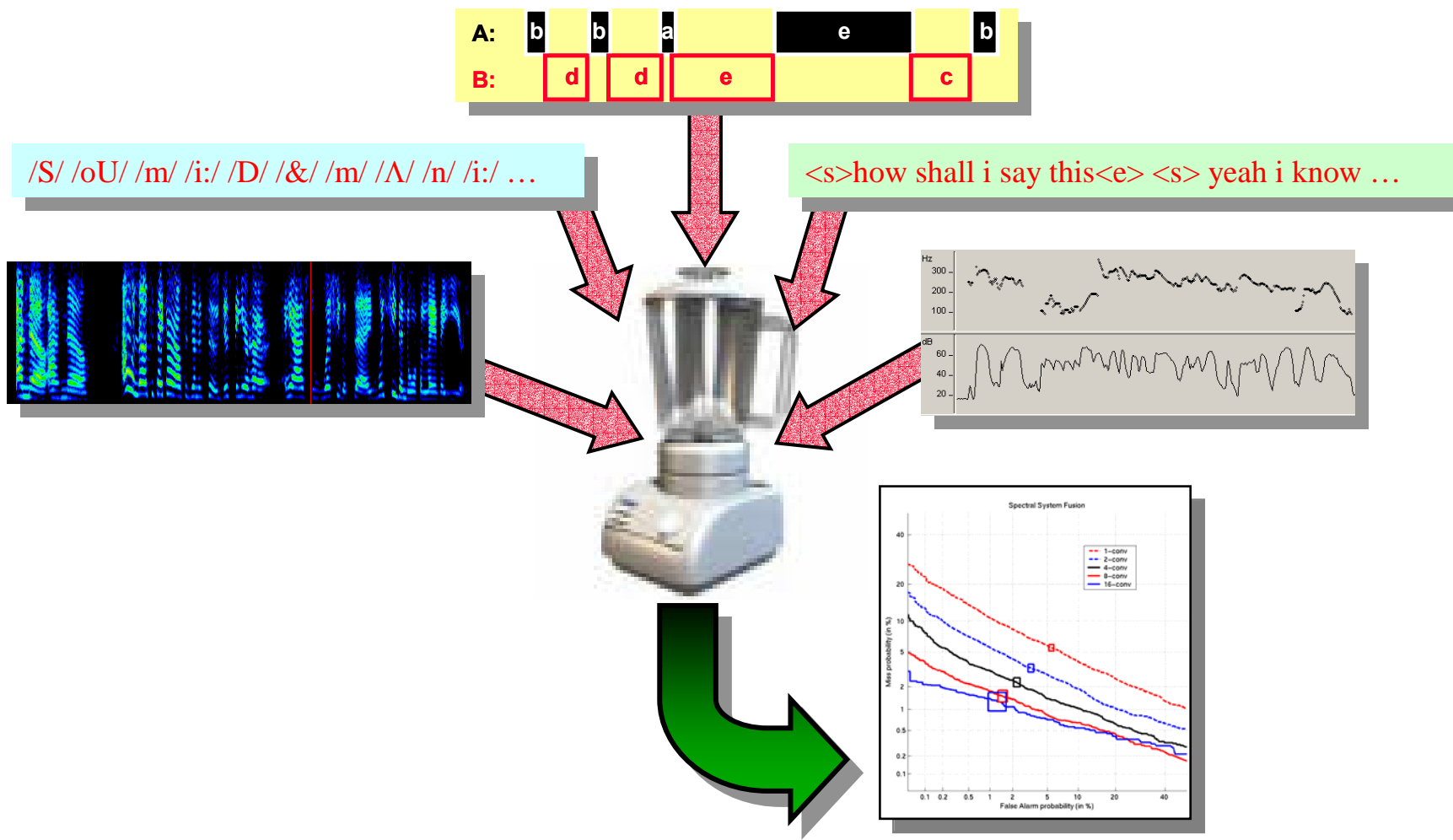
dialogic

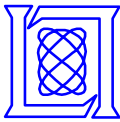
- Conversational patterns
  
- **Fusion**
- **Conclusions**



# Information Fusion

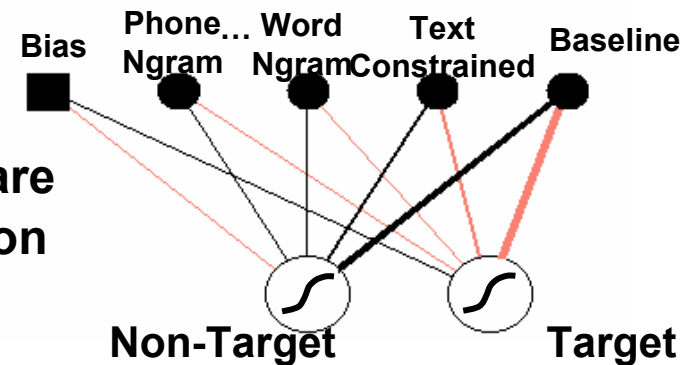
- Equally important to extracting the different levels of information is effectively combining them



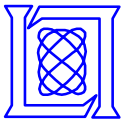


# Fusion System Design

- Extension of SRE '02 fusion approach
- Used LNKnet\* pattern classification software
- Selected fusion classifiers run on evaluation data using 10-fold cross-validation
  - Classifiers for split 1 trained on splits 2-10
  - Classifiers for split 2 trained on splits 1, 3-10
- Oracle fusion (MITL3) tried all system combinations = 501 fusers
- Separate classifiers were trained and used for each of 5 training conditions: 1, 2, 4, 8, and 16 training conversations
- Separate classifiers for each of the 10 jackknifed splits
- Perceptron chosen for accuracy & robust decision regions
- Perceptron used with no hidden layer (similar to linear discriminant but with sigmoid on output)
- LNKnet's priors set to minimize errors at DCF operating point
- LNKnet's decisions & scores used; score =  $[S_{tar} + (1.0 - S_{non})]/2$



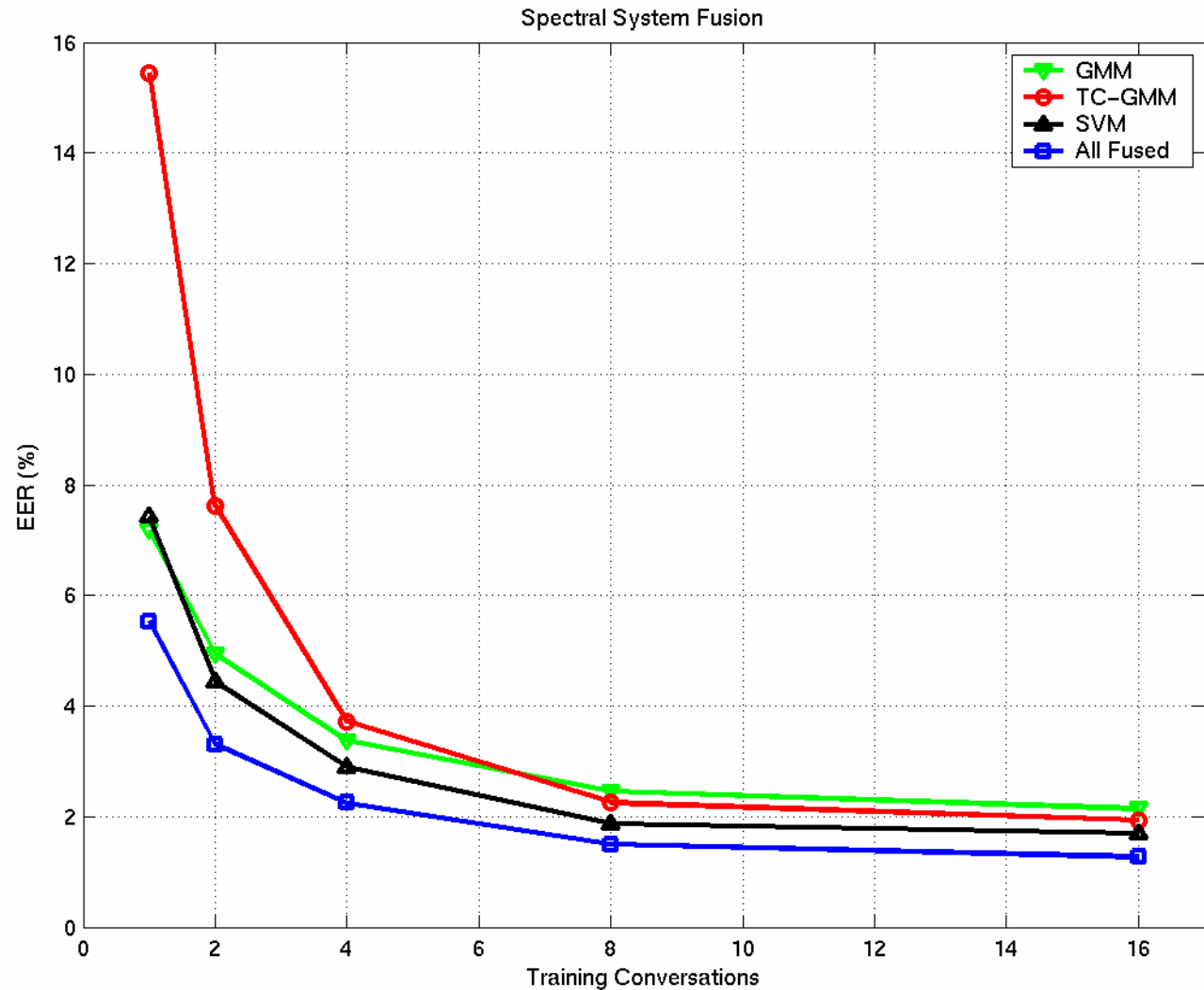
\*LNKnet by R. Lippman et. al., free at <http://www.ll.mit.edu/IST/lnknet>



# Spectral System Components

## EER vs No. Training Conversations

- GMM & SVM similar EERs with different features and classifiers
- TC-GMM improves greatly with training
- MITL4 spectral fusion gain
- Complementary features

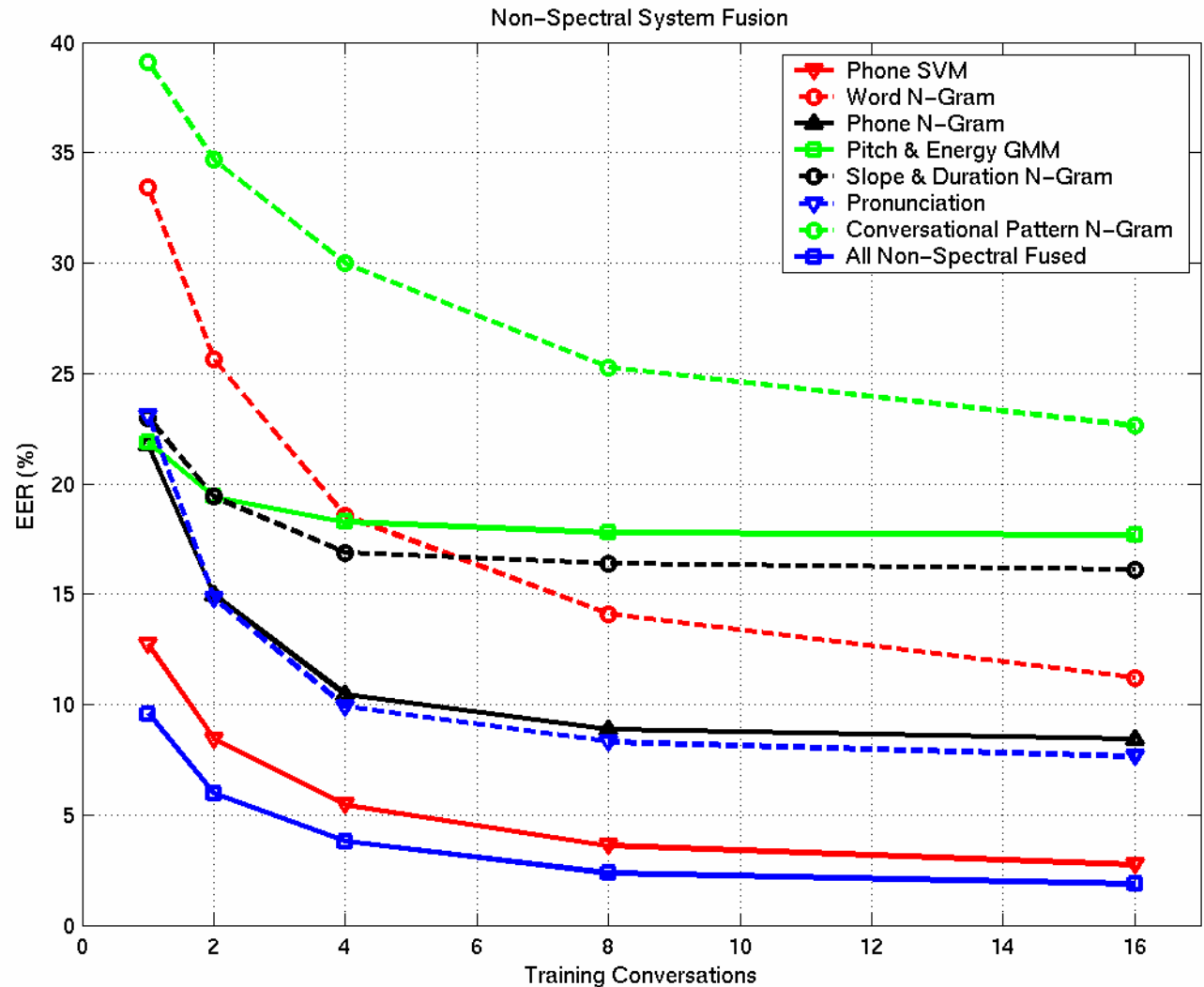


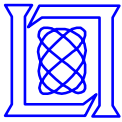


# Nonspectral System Components

## EER vs No. Training Conversations

- Wide range of EERs
- Phone n-gram and Pronunciation nearly identical
- Phone SVM is efficient
- Word n-gram improves greatly with training
- Nonspectral fusion gain
- Complementary features

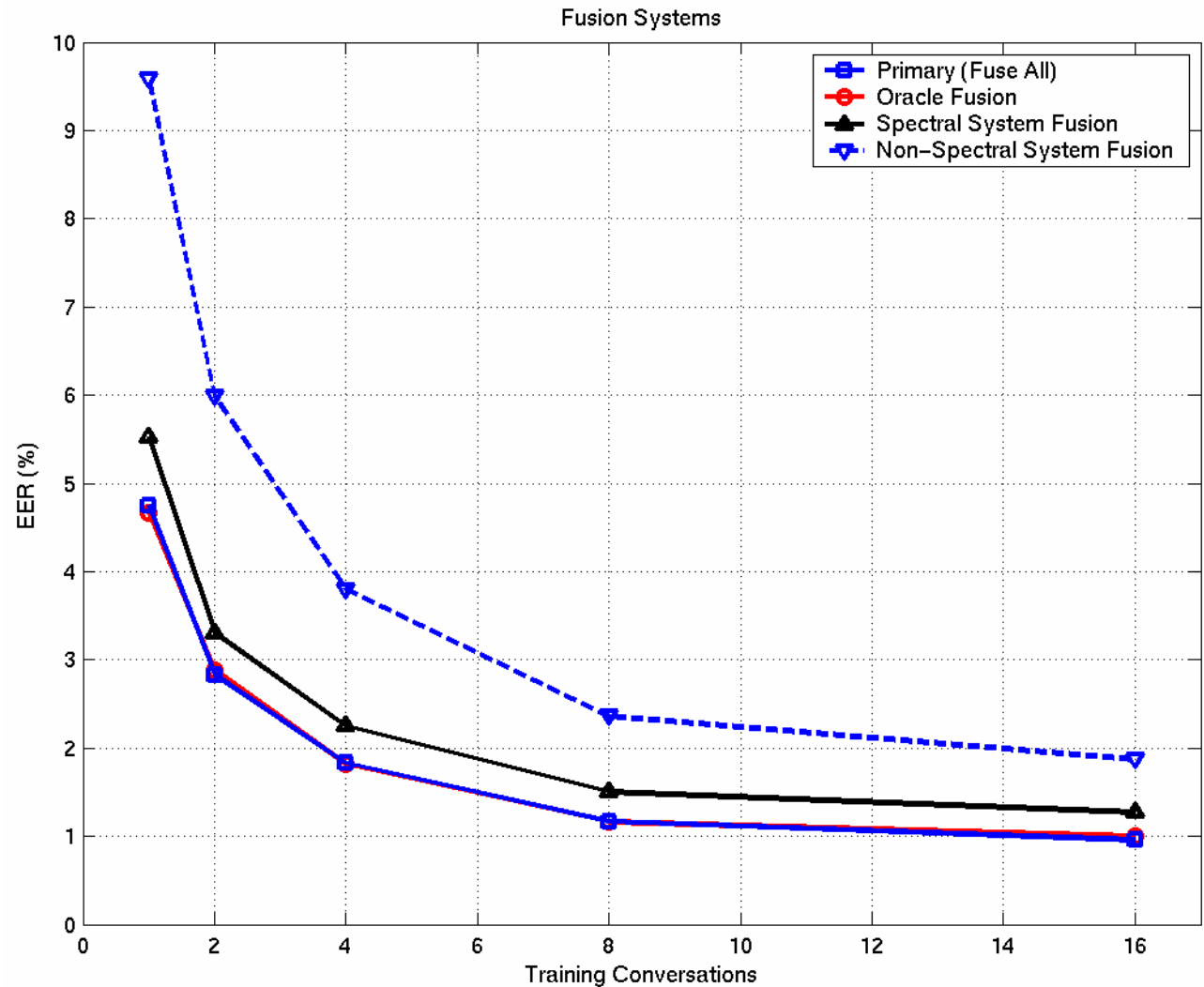


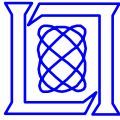


# System Fusion

## EER vs No. Training Conversations

- **Primary Fuses All**
- **Oracle selects best**
  - Min DCF
- **Primary & Oracle have nearly identical EER**
  - Robust fusion wrt EER & DCF and extra systems
- **Spectral Fusion**
  - Source of greatest accuracy
- **NonSpectral Fusion**
  - Nonspectral systems improve greatly with training
  - Complementary; halve EER @ 16





# Conclusions

---

- **Higher-level features provide a wealth of new information for high-accuracy speaker recognition**
- **Achieved astonishingly low error rates - fusion of high-level features and low-level features yielded new record on this task**
- **Even at extremely low error rates, there is still significant benefit in combining complementary types of information**
- **Speaker recognition accuracy breakthrough**
- **Future**
  - Need for continued research
  - Scientific evaluations
  - Publicly available corpora
  - Interaction with users about desired applications
- **Thanks to**
  - Douglas Reynolds, Walter Andrews, Jiří Navrátil, Barbara Peskin, Andre Adami, Qin Jin, David Klusáček, Joy Abramson, Radu Mihaescu, John Godfrey, Douglas Jones, Bing Xiang
  - Bob Dunn, Richard Lippmann, Pedro Torres-Carrasquillo, George Doddington, Larry Heck
  - CLSP Group, Johns Hopkins U. for hosting WS'02
- **More: ICASSP '03 & [www.clsp.jhu.edu/ws2002/groups/supersid/](http://www.clsp.jhu.edu/ws2002/groups/supersid/)**