ICSI’s SRE05 System

Nikki Mirghafori, Andy Hatch, Steve Stafford, Kofi Boakye,
Dan Gillick, Shawn Cheng, and Barbara Peskin

With special thanks to:
our collaborators at SRI
&
our advisor George Doddington

Overview

- SRI shared resources
  - ASR
  - Development data
  - Cepstral GMM
- ICSI’s individual sub-systems
  - Keyword conditional HMM (WordHMM)
  - Phone n-grams
  - Sequential Non-Parametric (SNP)
- System combination
  - LNKnet combination of the sub-systems
  - Combining English & nonEnglish scores
- Ongoing/future work

Shared Resources Acknowledgment

- ASR:
  - Our three systems relied on word or phone recognition from SRI
- Background data:
  - Used subset of SWBII and Fisher, as defined by SRI in the previous talk
- Cepstral GMM system:
  - We're grateful to SRI for sharing their cepstral GMM system with us

… and, of course, many thanks for ongoing advice and support!

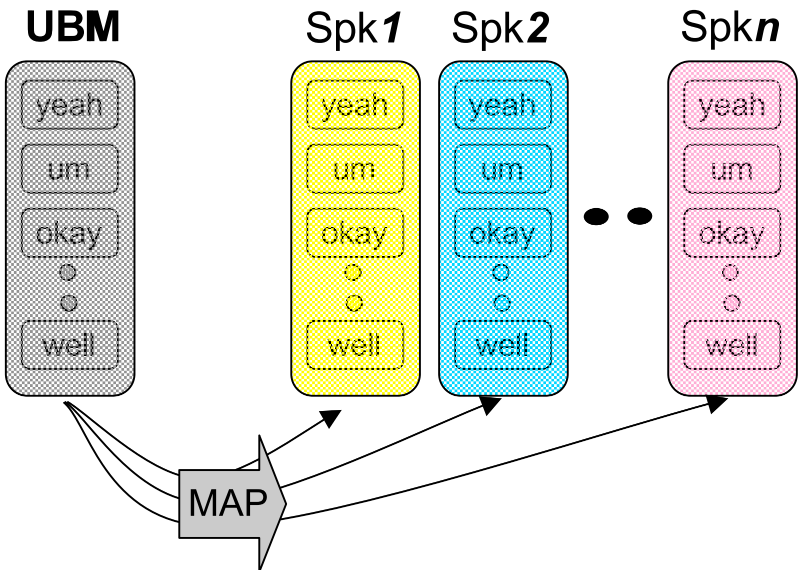
Keyword Conditional HMM (WordHMM) [1/3]

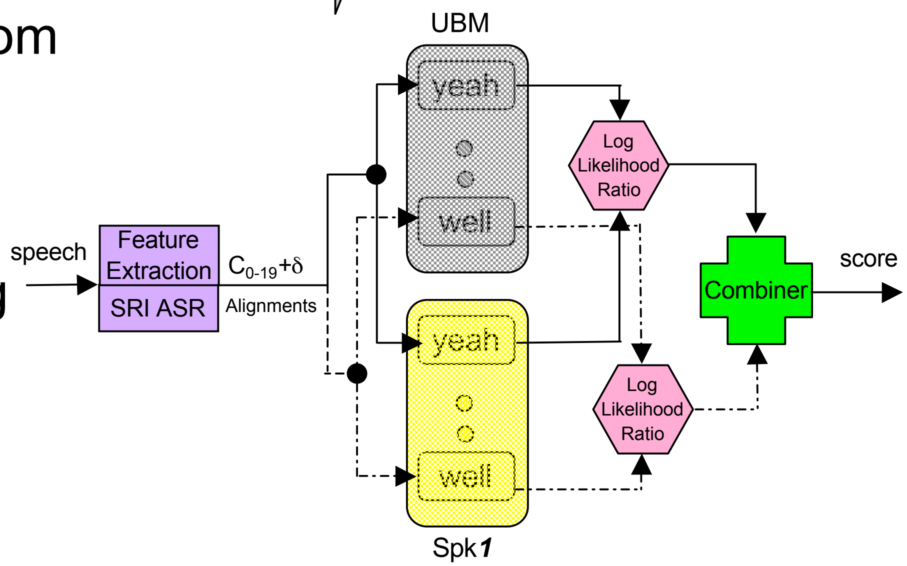
- **Main idea:**
  - Capitalize on advantages of text-dependent systems in a text-independent domain
  - Use frequent keywords that are rich with speaker characteristic cues (total of 19):
    - Discourse markers: {actually, anyway, like, see, well, now, you_know, you_see, i_think, i_mean}
    - Filled pauses: {um, uh}
    - Backchannels: {yeah, yep, okay, uhhuh, right, i_see, i_know }
  - Use whole-word HMMs, instead of GMMs, to model the evolution of speech in time

- This system was our only entry in SRE04
- For more details, see: *K. Boakye & B. Peskin, "Text-Constrained Speaker Recognition on a Text-Independent Task", Odyssey 2004*

Keyword Conditional HMM (WordHMM) [2/3]

- Models:
  - HMMs with self loops, no skips
  - 8 Gaussians/state
  - #states/word = min(#phones*3, median #frames/4)
  - C_0-C_{19} plus deltas
- UBM trained on 1,128 Fisher and 425 SWBII conversation sides
- Speaker models MAP adapted from UBM
- SRI's ASR used for finding word alignments
- HTK used for training and scoring HMMs



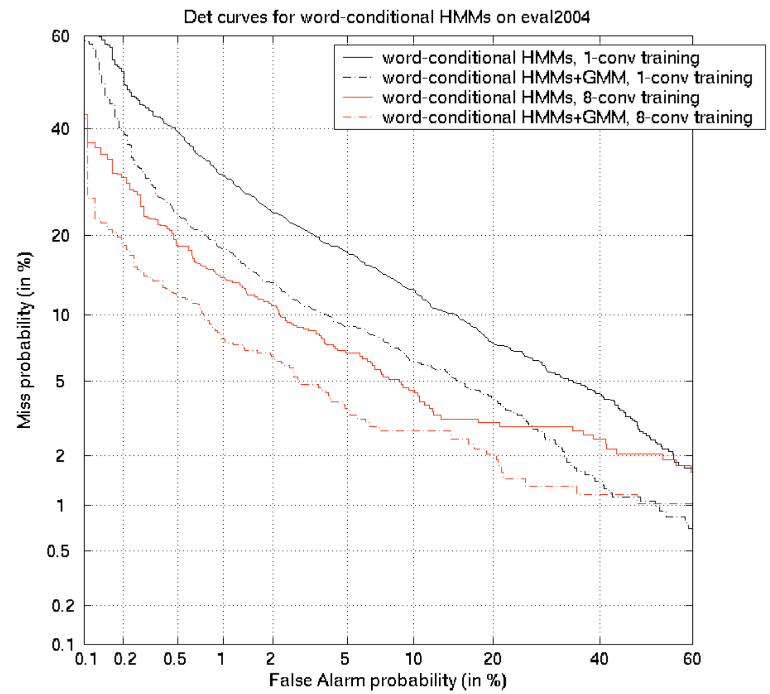


Keyword Conditional HMM (WordHMM) [3/3]

- New this year:
  - Major infrastructure changes, resulting in better alignments
  - Use of improved SRI ASR
  - Speed enhancements
  - Addition of TNORM
  - 8, instead of 4, Gaussians/state
  - Fisher, in addition to SWBII data, for UBM training

WordHMM on all English trials of Eval04	1-side training		8-side training	
	EER	DCF	EER	DCF
SRE04 system	13.06%	0.526	8.85%	0.382
SRE04 post-eval	12.98%	0.445	7.06%	0.306
SRE05 system	11.38%	0.399	6.27%	0.224

SRE04 UBM was trained entirely on SWBII, whereas SRE04 post-eval was trained entirely on Fisher. SRE05 UBM was trained on subsets of both.

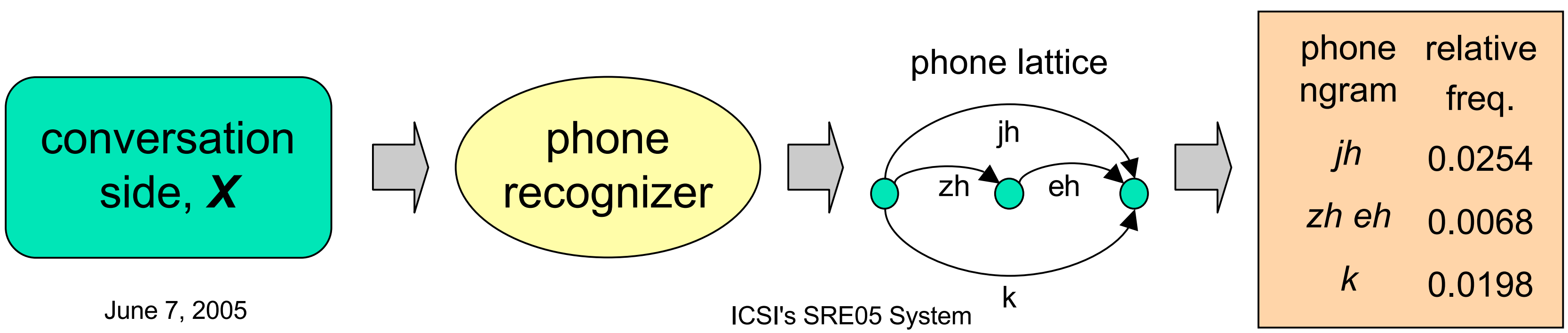


All English trials of Eval04	1-side training		8-side training	
	EER	DCF	EER	DCF
WordHMM	11.38%	0.3990	6.27%	0.2244
GMM	7.73%	0.3113	4.96%	0.2115
WordHMM+GMM	7.59% (2%)	0.2721 (13%)	4.08% (18%)	0.1672 (21%)

Values in () are % improvements relative to GMM sys alone.
“DCF” is short for “Min DCF” in tables throughout.

SVM-based Phone N-gram System [1/2]

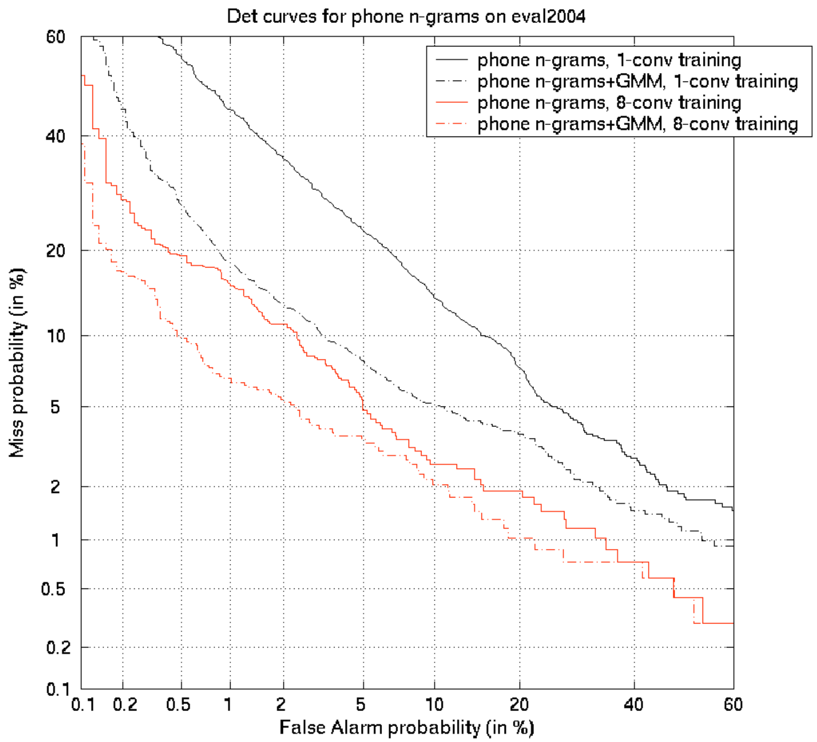
- **Main idea:**
  - To compute relative frequency of phone n-grams, use lattice open-loop phone decoding, instead of 1-best
  - Utilize SVMs for modeling
    - Relative frequencies of phone n-grams used as feature vectors
    - One feature vector for every conversation side
    - Target model's conversation(s): positive example(s)
    - Background model's conversations: negative examples
    - Use kernelized form of LLR [Campbell et al., NIPS 2003]
- **The System:**
  - Used a vocabulary of 46 phone units
  - Used only phone bigrams and the top 8500 phone trigrams



SVM-based Phone N-gram System [2/2]

- For more information, see: *A. O. Hatch, B. Peskin, A. Stolcke, "Improved Phonetic Speaker Recognition Using Lattice Decoding", ICASSP 2005*

All English trials of Eval04	1-side training		8-side training	
	EER	DCF	EER	DCF
Phone N-grams	12.09%	0.5408	4.96%	0.2358
GMM	7.73%	0.3113	4.96%	0.2115
PhoneNg+GMM	6.47% (16%)	0.2767 (11%)	3.64% (27%)	0.1443 (32%)

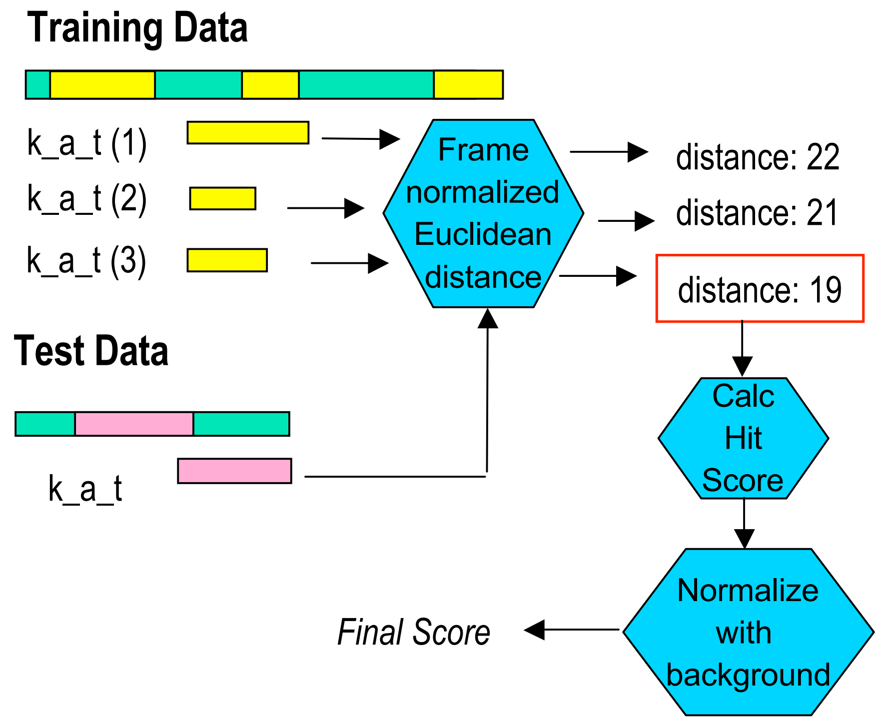


Sequential Non-Parametric (SNP) System [1/2]

- **Main Idea**:
  - Compare a test segment directly to similar segments in training data
  - Non-parametric -- no explicit models built
  - New scoring method -- capture primarily positive evidence ("hit score")
- **The system:**
  - C_0-C_{19} plus deltas
  - 60 SWBII and 40 Fisher conversation sides for background
  - Phone trigram sequences
  - DTW to align frames
  - Euclidean distance between aligned frames
  - Calculate the *best* "Hit Score"

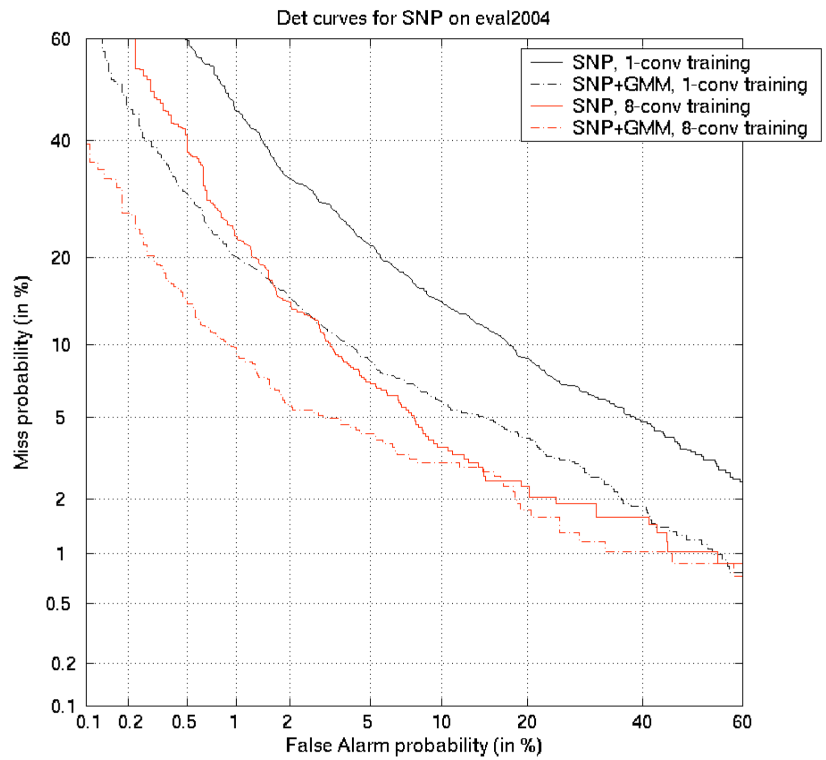
$$HS = \sum_{i \in \text{test tokens}} \frac{\text{number of matched frames in } i}{k^{\text{dist}[i]}}$$

  - Divide HS by background HS



Sequential Non-Parametric (SNP) System [2/2]

- Includes Znorm
- But no TNORM, for lack of computational resources
- For more information, see: *D. Gillick, S. Stafford, B. Peskin, "Speaker Detection Without Models", ICASSP 2005*
- This system was inspired by Dragon's SRE98 submission



All English trials of Eval04	1-side training		8-side training	
	EER	DCF	EER	DCF
SNP	12.65%	0.5177	6.12%	0.3169
GMM	7.73%	0.3113	4.96%	0.2115
SNP+GMM	7.10% (8%)	0.2943 (6%)	4.37% (12%)	0.1777 (16%)

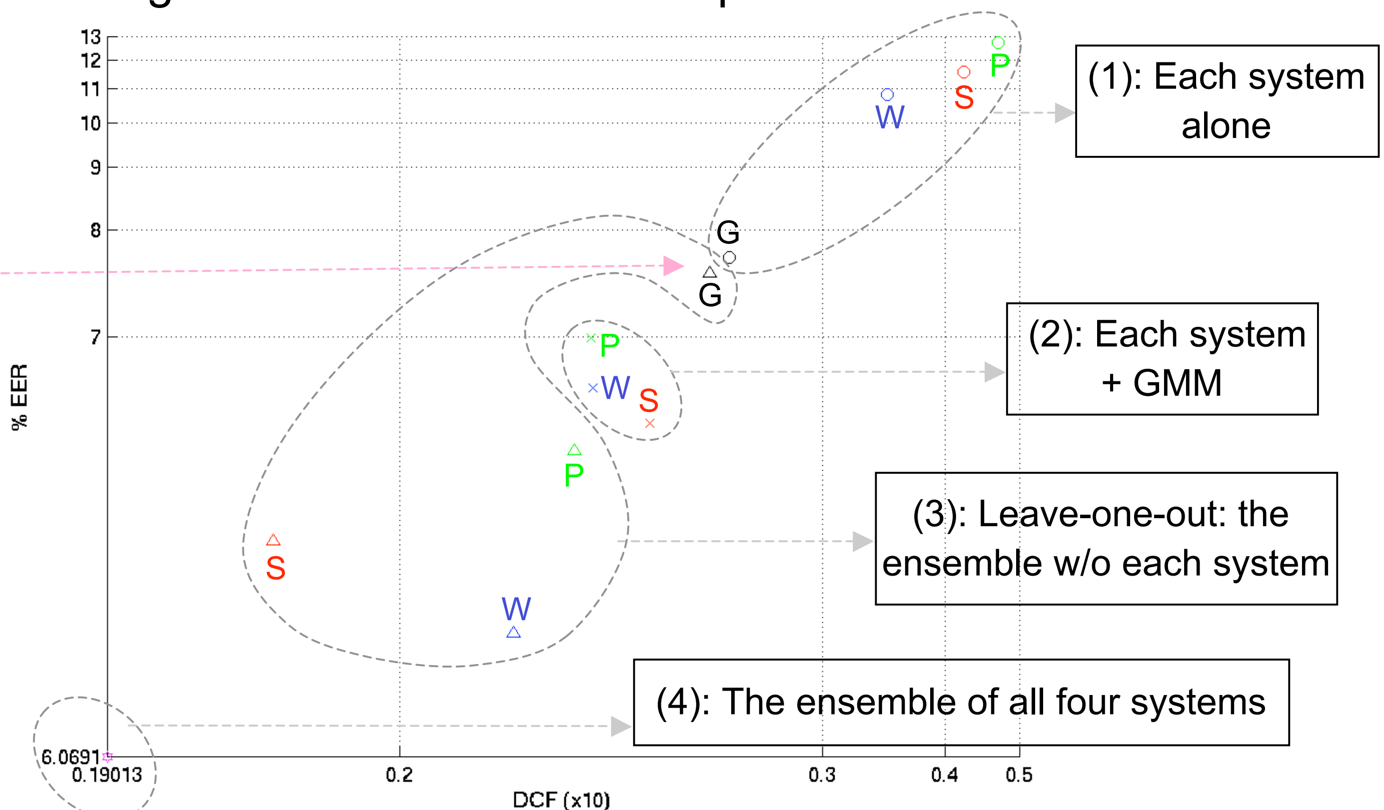
Combination of Systems -- 1-side

- Used LNKnet neural network package
- No hidden layer
- Sigmoid output nonlinearity
- Combination weights trained on Eval04 to optimize DCF

Observations:

- All systems contributed
- Excluding GMM hurt most in 1-side case

Color legend:
W WordHMM
P Phone N-gram
S SNP
G GMM

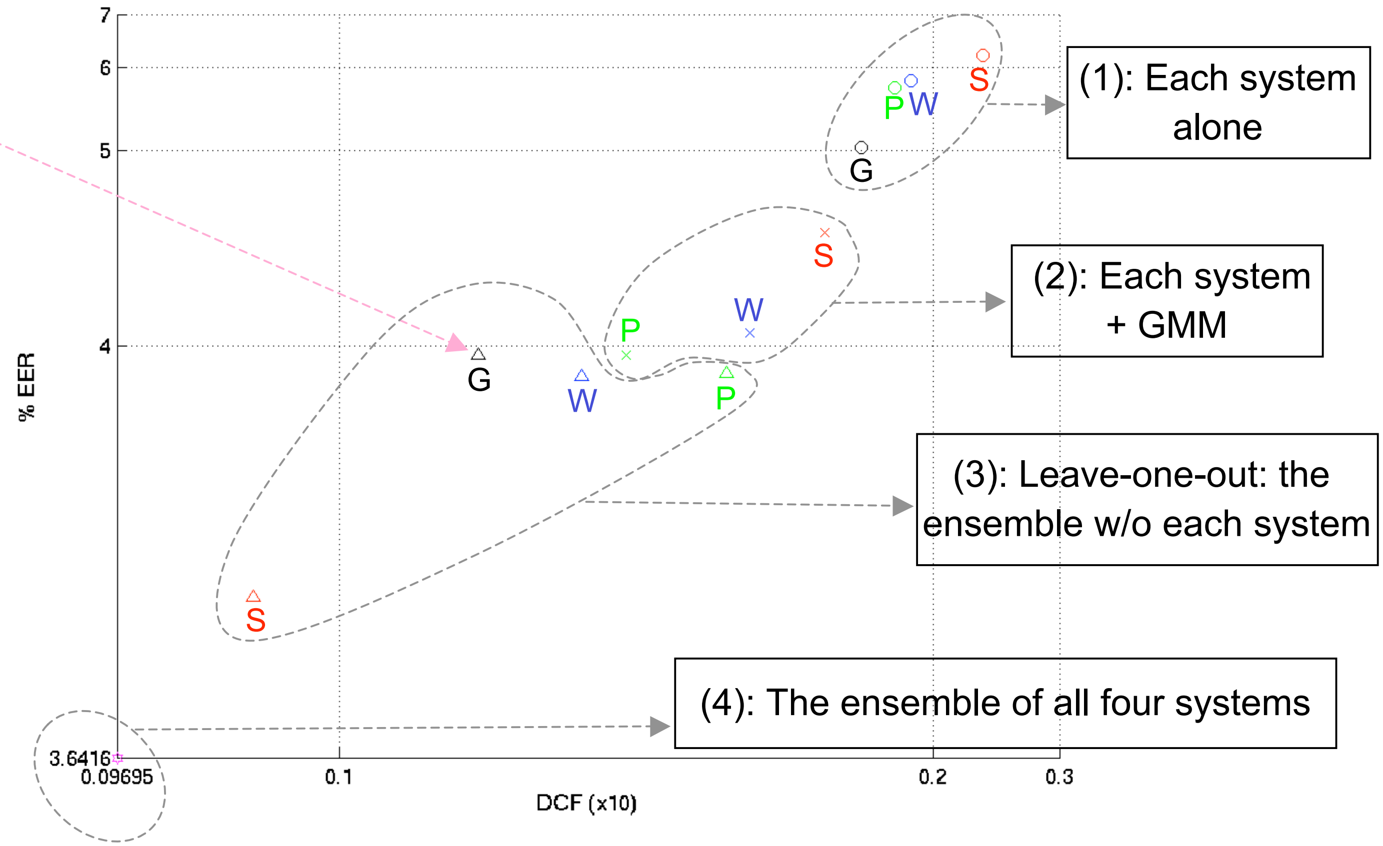


1-side training results on **all English** trials of Eval05

Combination of Systems -- 8-side

Observations:

- All systems contributed in 8-side training condition, as well
- Excluding GMM did not hurt as much, relatively, as in the 1-side training condition

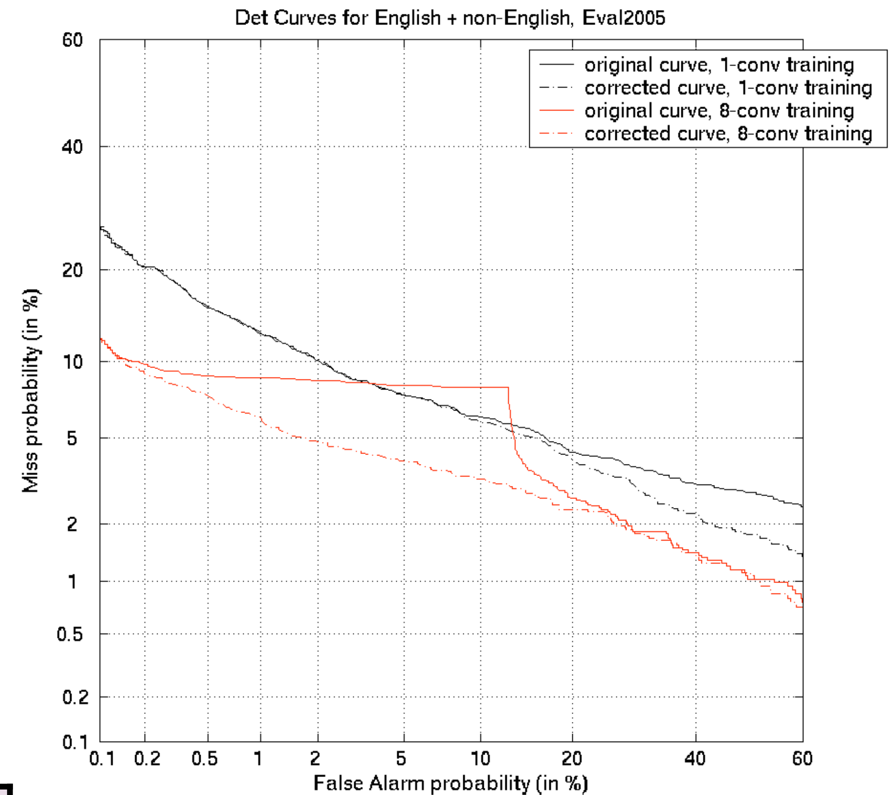


8-side training results on **all English** trials of Eval05

Color legend:
W WordHMM
P Phone N-gram
S SNP
G GMM

Appending NonEnglish and English Scores

- English scores calculated on combination of all four systems
- NonEnglish scores calculated on combination of GMM and phone-Ngram systems only
- For each set of scores (English and nonEnglish) independently:
  1. Optimize LNKnet weights using Eval04
  2. Remove sigmoidal non-linearity
  3. Z-normalize scores using Eval04 stats
  4. Calculate score threshold for min DCF
  5. Subtract threshold from scores
- Append two sets of scores from step 5

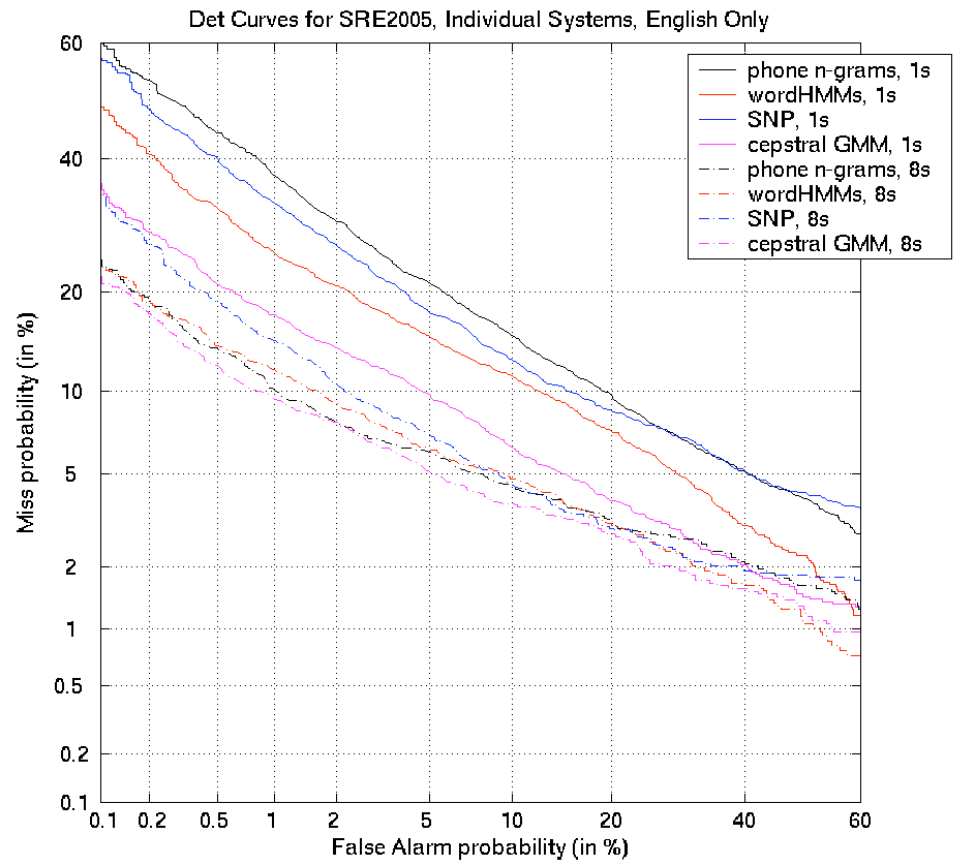


All trials of Eval05	1-side training		8-side training	
	EER	DCF	EER	DCF
Original submission	6.85%	0.2009	8.02%	0.1163
Corrected	6.83%	0.2028	4.10%	0.1099

- Because of the strong sigmoidal non linearity in LNKnet, ignoring steps 2 & 3 can result in DET displaying flat regions (as in our official submission for 8-side)

Comparing Individual Systems: 1-side vs. 8-side Training

- Phone N-gram system (black DET) improves the most with increase of training data
- Other systems preserve their relative order
- GMM remains the best in both training conditions
- But, the gap is closing for 8-side training



Ongoing/Future Work

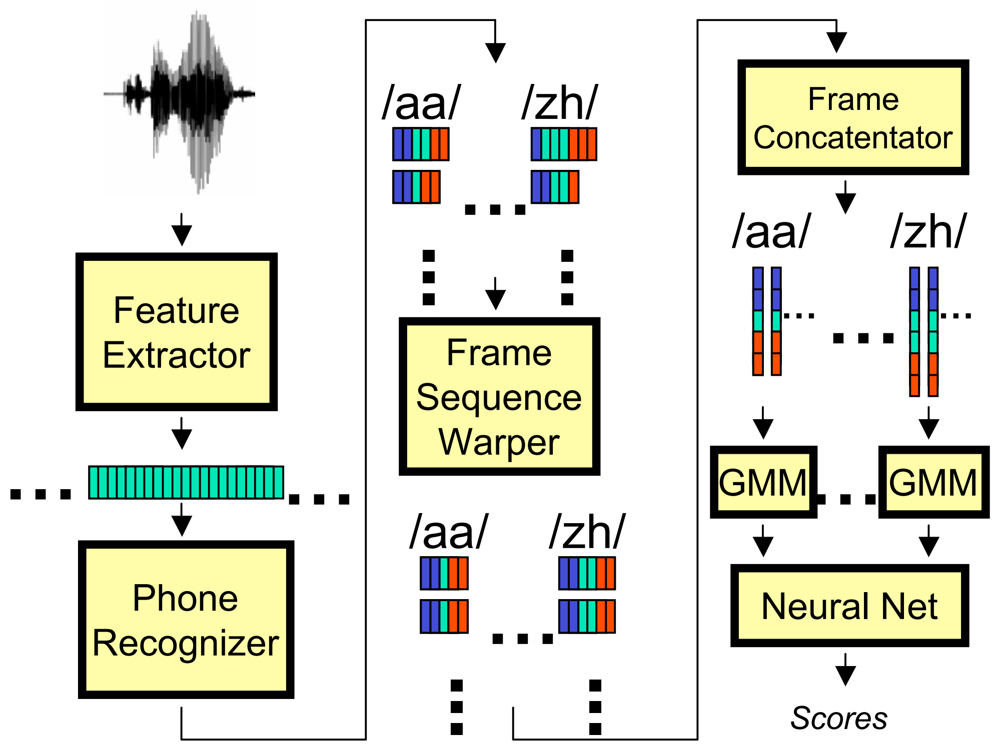
- Addition of prosodic features to WordHMM system
- Development of inhouse GMMs using Torch toolkit
- Use of discriminant long-term (calculated over 500 ms) features in GMMs
- Study and experimentation with cross-channel data for robustness to channel variation
- Sequential GMM
- Assignment of optimal weights to feature sets combined via SVMs

Sequential GMM (SGMM) System

- **Main Idea:**
  - Use concatenated phoneme-length feature vectors, one “stacked frame” for each phone token
  - Build a separate GMM system for each phone (46)
  - Combine resulting scores using a neural net

- **Results**:
  - Combines well with GMM
  - Can take advantage of the ubiquity of GMM

SWB I	EER	DCF
SGMM	1.14%	0.0575
GMM	0.90%	0.0509
SGMM+GMM	0.57%	0.0180



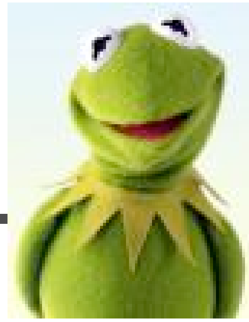
See: *S. Stafford, “The Sequential GMM…” , Masters thesis, UC Berkeley, May 2005*.

Optimal Weights for SVM Features

- **Main idea**:
  - When combining different feature sets with SVMs, automatically learn optimal weights to minimize the EER for a given set of SVM-based speaker models
- Optimize:

$$K(A,B) = \sum_i \mu_i K_i(A,B),$$

  - where A and B are conversation sides, μ_i are a set of positive weights, and $K_i(A,B)$ represents a kernel for a particular set of features (e.g. phone n-grams).
- Preliminary results:
  - Trained relative weights for the 8 feature sets in SRI's MLLR-SVM system
  - Relative improvements:
    - 6.8% on SWBII
    - 4.2% on Eval04

As  the frog said,