# ICSI's SRE06 System

Nikki Mirghafori, Lara Stoll, Andy Hatch, Howard Lei

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

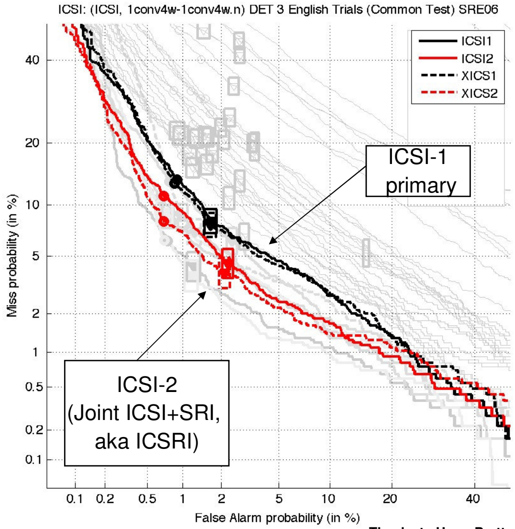
Updated presentation can be downloaded from:
www.icsi.berkeley.edu/~nikki/ICSI.pdf.gz



## Summary: The Tale of Two Evaluations

- The Good…
  - 5 ICSI sub-systems this year
    - 4/5 new or improved
    - 3/5 use word-conditioning
    - All "high-level": rely on phone/word recognition
  - 1 new normalization method developed
    - WCCN: Within-Class Covariance Normalization
  - 2006 system improved by ~10% on 1-side and ~25% on 8-side compared to 2005 (on SRE05)
- The Bad…
  - Problems with the evaluation caused submission results and, especially, the researchers who re-ran them, to suffer…
- No Ugly!

# Results: ICSI and XICSI (resubmission)




**Thanks to Harry Bratt of SRI for creating the overlay DETs.**



# Submission Overview

- Submitted conditions:
  - {1,8}conv4w-1conv4w
  - {1,8}conv4w-1convmic
    (presented tomorrow by Lara Stoll)
- Submitted Systems:

ICSI-1 (primary) ← SVM Fusion (6 Sub-Sys)

ICSI-2 (joint ICSI+SRI, aka ICSRI-2)

**ICSI Sub-systems**
1. Word-Conditioned (WC) HMM
2. WC Phone Ngram SVM + WCCN
3. WC POS Ngram SVM
4. Phone Ngram SVM + WCCN
5. Lexical Statistics SVM

**SRI Sub-systems**
1. Baseline Cepstral GMM
2. Cepstral SVM
3. Word-Ngram SVM
4. State Duration GMM
5. Word Duration GMM
6. Grammar + Syllable NERF
7. MLLR SVM

**Jointly developed**
1. MLLR SVM + WCCN

# Submission Overview

- Submitted conditions:
  - {1,8}conv4w-1conv4w
  - {1,8}conv4w-1convmic (presented tomorrow by Lara Stoll)
- Submitted Systems:

ICSI-1 (primary)

ICSI-2 (joint ICSI+SRI, aka ICSRI)

SVM Fusion (13 Sub-Sys)

ICSI Sub-systems
1. Word-Conditioned (WC) HMM
2. WC Phone Ngram SVM + WCCN
3. WC POS Ngram SVM
4. Phone Ngram SVM + WCCN
5. Lexical Statistics SVM

SRI Sub-systems
1. Baseline Cepstral GMM
2. Cepstral SVM
3. Word-Ngram SVM
4. State Duration GMM
5. Word Duration GMM
6. Grammar + Syllable NERF
7. MLLR SVM

Jointly developed
1. MLLR SVM + WCCN

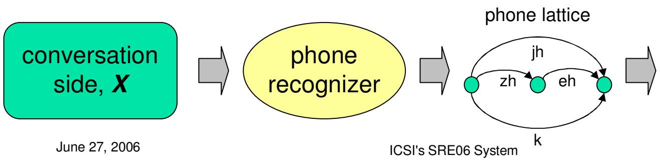


# Word-Conditioned HMM

- **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
- Same system used in SRE05
- For more details, see: *K. Boakye & B. Peskin, "Text-Constrained Speaker Recognition on a Text-Independent Task", Odyssey 2004*

## Phone Lattice Ngram SVM System

- **Main idea:**
  - Compute relative frequency of phone Ngrams using <u>lattice</u> open-loop phone decoding
- **Modeling with SVM:**
  - Relative frequencies of phone Ngrams 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
- For more information, see: *A. O. Hatch, B. Peskin, A. Stolcke, "Improved Phonetic Speaker Recognition Using Lattice Decoding", ICASSP 2005*




| phone ngram | relative freq. |
|---|---|
| *jh* | 0.0254 |
| *zh eh* | 0.0068 |
| *k* | 0.0198 |



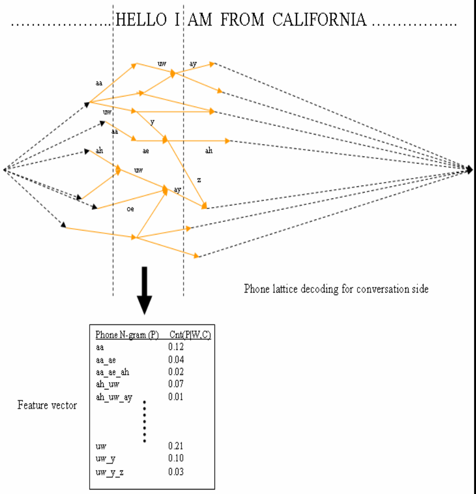
## Word-Conditioned Phone Lattice N-gram SVM System

**Main Idea**
- Similar to previous system, except word conditioned

**Features**
- Relative frequencies of phone Ngrams from word-conditioned phone lattice segments
- Concatenate phone N-grams from different words for each conv. side

**Modeling**
- SVM with kernelized form of LLR

**Vocabulary**
- 52 word unigrams with highest frequency from background data
- Uni,bi,tri-phones from 46 different phones
- Use top 27,410 phone N-grams




| Phone N-gram (P) | Cnt(P\|W,C) |
|---|---|
| aa | 0.12 |
| aa_ae | 0.04 |
| aa_ae_ah | 0.02 |
| ah_uw | 0.07 |
| ah_uw_ay | 0.01 |
| ⋮ | |
| uw | 0.21 |
| uw_y | 0.10 |
| uw_y_z | 0.03 |

# Word-Conditioned POS SVM System

- **Main idea**:
  - Similar to a word Ngram system [Doddington '01], except relative frequencies of the joint POS/words used
- **Features**:
  - Part of Speech (POS) tags generated using Brill's Supervised Tagger
  - Example: but/CC i/NN see/VB
    - (CC: Coordinating conjunction, NN: Noun, VB: Verb)
  - A total of 125,700 uni-, bi-, and tri-grams used
- **Model**:
  - SVM with a linear kernel trained using SVMLite



# Lexical Statistics SVM System

- **Main Idea**:
  - Capture sentence and conversation level information
- **Features**:
  - Eight features per conv side, such as:
    - Number of conversation turns
    - Number of words (per conversation, per turn)
    - Number of characters (per conversation, per turn)
    - Speaking rate (words per second)
- **Model**:
  - SVM with a linear kernel trained using SVMLite

## Within-Class Covariance Normalization (WCCN) [1/2]

- **Intuition:**
  - WCCN emphasizes "directions" in feature space that are informative while attenuating directions that are noisy.
  - Session variability modeling ala NAP (Solomonoff 04), Factor Analysis (Kenny 04), Modeling Session Variability (Vogt 05)
  - Can show that WCCN minimizes an upper bound on classification error in SVMs.
  - Weighs the directions in feature space that are retained.

- **Main idea:**
  - Given a set of input feature vectors, normalize the expected *within-speaker covariance matrix* to equal the identity matrix over some training set.
  - **Implementation:** A linear feature transformation, $\Phi$ , is used which is defined as:

$$\Phi(x) \equiv A^T x, \qquad AA^T \equiv W^{-1}, \qquad W \equiv \frac{1}{N}\sum_{i=1}^{N} C_i.$$

cov. matrix for speaker *i*.



## Within-Class Covariance Normalization (WCCN) [2/2]

How can we perform WCCN on large feature sets?

- **Experimental Procedure:**
  1. Do per feature within class variance normalization.
  2. Use PCA to reduce dimensionality of input feature vectors.
     1. PCA was trained on ~3600 files of SRE-2003.
  3. Perform WCCN on reduced-dimensionality feature vectors.
     1. WCCN was trained on ~7200 files of SRE-2003.
     2. Linear smoothing was applied to the final estimate of the expected WCC matrix, W.
  4. Concatenate each resulting feature vector with scaled version of its "PCA-complement" (i.e. the portion of the original feature vector that was filtered out by performing PCA).

WCCN was applied to the following systems: Phone Ngrams, WC-Phone Ngrams, and MLLR.

For more information, see: *A. Hatch, et al., "Within-class covariance normalization for SVM-based speaker recognition," to appear at ICSLP 2006.*

## Fusion Strategy

- SVMLite with a linear kernel in classification mode
- English and non-English scores combined separately
- For non-English:
  - GMM and PhoneNgram+WCCN were combined
- For English:
  - All systems were combined
- SRE05 used to train combiner and optimize DCF threshold
  - post-eval analysis showed using SRE04 would have similar results

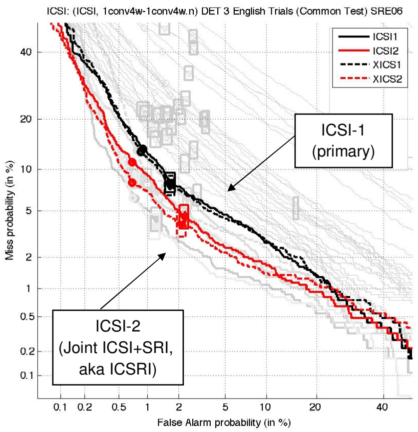


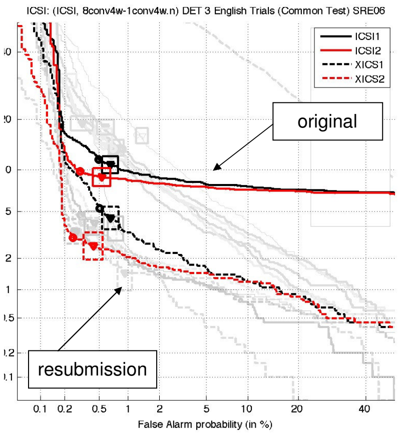
## Resources common to All Sub-systems

- Background data
  - Used subset of SWBII and Fisher
- TNORM
  - 249 gender-balanced utterances from Fisher
- ASR
  - All our systems used word or phone recognition (from SRI)

# Our (Re)submission (X)ICSI

- Problems with initial submission:
  - Difficulty determining English trials list (both 1s and 8s)
  - Assigned dummy scores for "presumed" empty files/models (8s)




**Thanks to Harry Bratt of SRI for creating the overlay DETs.**



# Performance on "English" trials

- As reported by Lucianna Ferrer to the SRE06 list, non-English trials were found in the core condition list
- Results on "cleaned v2" English lists are shown in the rest of the presentation (helps clarity of analysis)

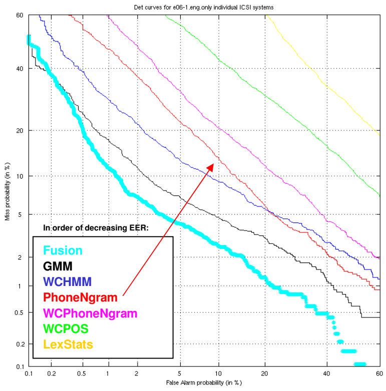
| ICSI-1 primary submission Effect of non-English trials | 1-side training | | | | 8-side training | | | |
|---|---|---|---|---|---|---|---|---|
| | EER | mDCF | aDCF | # Trials | EER | mDCF | aDCF | # Trials |
| NIST Eng v2 list | 5.34% | 0.271 | 0.301 | 22,433 | 2.34% | 0.115 | 0.131 | 17,387 |
| **Cleaned** Eng v2 list | **4.35%** | **0.209** | **0.226** | 18,926 | **2.28%** | **0.095** | **0.103** | 15,887 |
| Percent improvement | (19%) | (23%) | (25%) | N/A | (3%) | (17%) | (21%) | N/A |

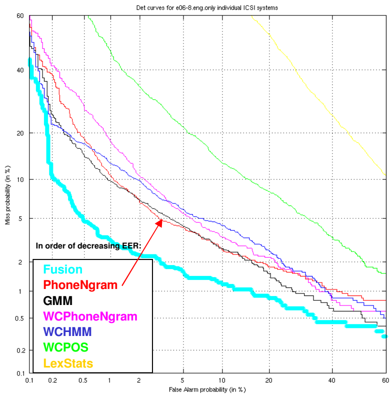
| ICSI-2 submission (ICSI+SRI) | 1-side training | | | | 8-side training | | | |
|---|---|---|---|---|---|---|---|---|
| | EER | mDCF | aDCF | # Trials | EER | mDCF | aDCF | # Trials |
| NIST Eng v2 list | 3.60% | 0.198 | 0.299 | 22,433 | 1.73% | 0.060 | 0.0785 | 17,387 |
| **Cleaned** Eng v2 list | **2.84%** | **0.146** | **0.228** | 18,926 | **1.69%** | **0.050** | **0.0606** | 15,887 |
| Percent improvement | (21%) | (26%) | (24%) | N/A | (2%) | (17%) | (23%) | N/A |

- Note: DCF refers to min DCF in all tables unless specified.

# System for **English**: 1-side vs. 8-side Training

- Phone Ngram system (red DET) improved most with increased training data
- GMM remained the best in both training conditions
- But, the gap was close for 8-side training

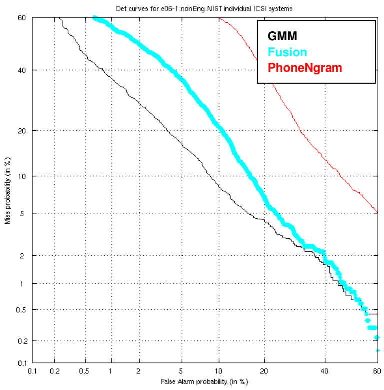


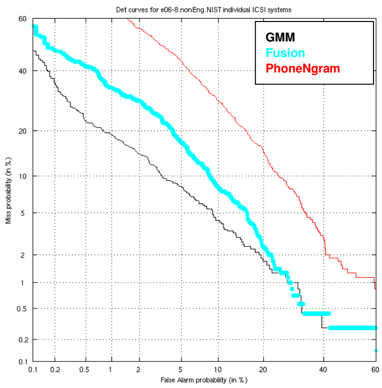






# System for **nonEnglish**

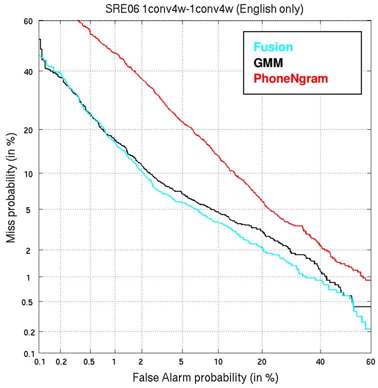
- For nonEnglish, Phone Ngram system (red) **degraded** fusion results (thick blue)
- However, if the fusion weights are ideally trained (cheating experiment), addition of PhoneNgram improves results
- Explained if nonEnglish SRE06 data has different statistics compared to SRE05

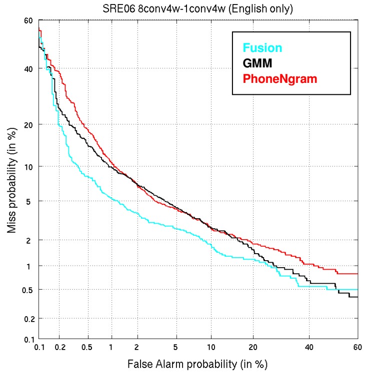


## nonEnglish Sub-Systems for English Trials

- From last slide, does addition of PhoneNgram to GMM degrade results also for English trials?
  - Addition of Phone N-gram **improves** fusion






## Combination of Systems -- English 1-side

**Observations**:

- All, but LexStat, contributed to Min DCF
- WCHMM best system after GMM with most significant contribution

| Best | GMM | WCHMM | PhoneNgram | WCPOS | WC Phone Ngram | LexStats | Min DCF | Pct. Chg |
|---|---|---|---|---|---|---|---|---|
| 1 sys | X | | | | | | 0.2639 | N/A |
| 2 sys | X | X | | | | | 0.2279 | 13.7% |
| 3 sys | X | X | X | | | | 0.2111 | 7.3% |
| 4 sys | X | X | X | X | | | 0.2095 | 0.7% |
| 5 sys | X | X | X | X | X | | 0.2077 | 0.9% |
| 6 sys | X | X | X | X | X | X | 0.2087 | -0.5% |

## Combination of Systems -- English 8-side

**Observations**:

- WCPhoneNgram best system after GMM
- As amount of data increases, word-conditioned systems become more powerful

| Best | GMM | WCPhone Ngram | WCHMM | WCPOS | Phone Ngram | LexStats | Min DCF | Pct. Chg |
|---|---|---|---|---|---|---|---|---|
| 1 sys | X | | | | | | 0.1865 | N/A |
| 2 sys | X | X | | | | | 0.1169 | 37.3% |
| 3 sys | X | X | X | | | | 0.1023 | 12.5% |
| 4 sys | X | X | X | X | | | 0.0950 | 7.1% |
| 5 sys | X | X | X | X | X | | 0.0946 | 0.4% |
| 6 sys | X | X | X | X | X | X | 0.0955 | -0.9% |



## The contribution of WCCN

- On average, WCCN helps improve individual systems
- Especially for non-stylistic ones (i.e., MLLR)

| MLLR | 1-side training | | 8-side training | |
|---|---|---|---|---|
| | EER | DCF | EER | DCF |
| No WCCN | 4.51% | 0.2076 | 2.28% | 0.0872 |
| With WCCN | 4.24% | 0.1846 | 2.23% | 0.0737 |
| Percent imprvmnt | (+6.0%) | (+11%) | (+2.1%) | (+15.4%) |

| Phone Ngram | 1-side training | | 8-side training | |
|---|---|---|---|---|
| | EER | DCF | EER | DCF |
| No WCCN | 13.25% | 0.6174 | 5.28% | 0.2638 |
| With WCCN | 12.69% | 0.63528 | 4.63% | 0.2457 |
| Percent imprvmnt | (+4.2%) | (-2.9%) | (+12.4%) | (+6.8%) |

| WCPhone Ngram | 1-side training | | 8-side training | |
|---|---|---|---|---|
| | EER | DCF | EER | DCF |
| No WCCN | 15.83% | 0.6672 | 4.81% | 0.2558 |
| With WCCN | 15.45% | 0.6698 | 5.26% | 0.2720 |
| Percent imprvmnt | (+2%) | (-0.4%) | (-9.3%) | (-6.3%) |

## SRE05 vs. SRE06

- ICSI's system has improved by ~10% on 1-side and ~25% on 8-side compared to 2005

| ICSI's 2005 vs. 2006 system | 1-side training | | 8-side training | |
|---|---|---|---|---|
| | EER | DCF | EER | DCF |
| 2005 system on SRE05 | 6.08% | 0.19049 | 3.64% | 0.09695 |
| 2006 system on SRE05 | 5.34% | 0.17298 | 2.65% | 0.07669 |
| Percent imprvmnt | **(+12%)** | **(+9%)** | **(+27%)** | **(+21%)** |

- SRE06 data produces lower EER and higher DCF compared to SRE05 data

| ICSI's 2005 vs. 2006 system | 1-side training | | 8-side training | |
|---|---|---|---|---|
| | EER | DCF | EER | DCF |
| 2006 system on SRE05 | 5.34% | 0.17298 | 2.65% | 0.07669 |
| 2006 system on SRE06 | 4.35% | 0.20874 | 2.28% | 0.09546 |
| Percent imprvmnt | **(+19%)** | **(-21%)** | **(+14%)** | **(-25%)** |



## Summary

- 5 ICSI sub-systems this year
  - All "high-level": rely on phone/word recognition
  - 3/5 use word-conditioning
  - 4/5 new or improved
- 1 new normalization method developed
  - WCCN: Within-Class Covariance Normalization
- 2006 system improved by ~10% on 1-side and ~25% on 8-side compared to 2005 (on SRE05)
- Problems with the evaluation caused difficulty
  - Request for English and Common Condition trial lists to be provided