The SRI NIST SRE10 Speaker Verification System

L. Ferrer, M. Graciarena, S. Kajarekar,

N. Scheffer, E. Shriberg, A. Stolcke

Acknowledgment: H. Bratt

SRI International

Menlo Park, California, USA

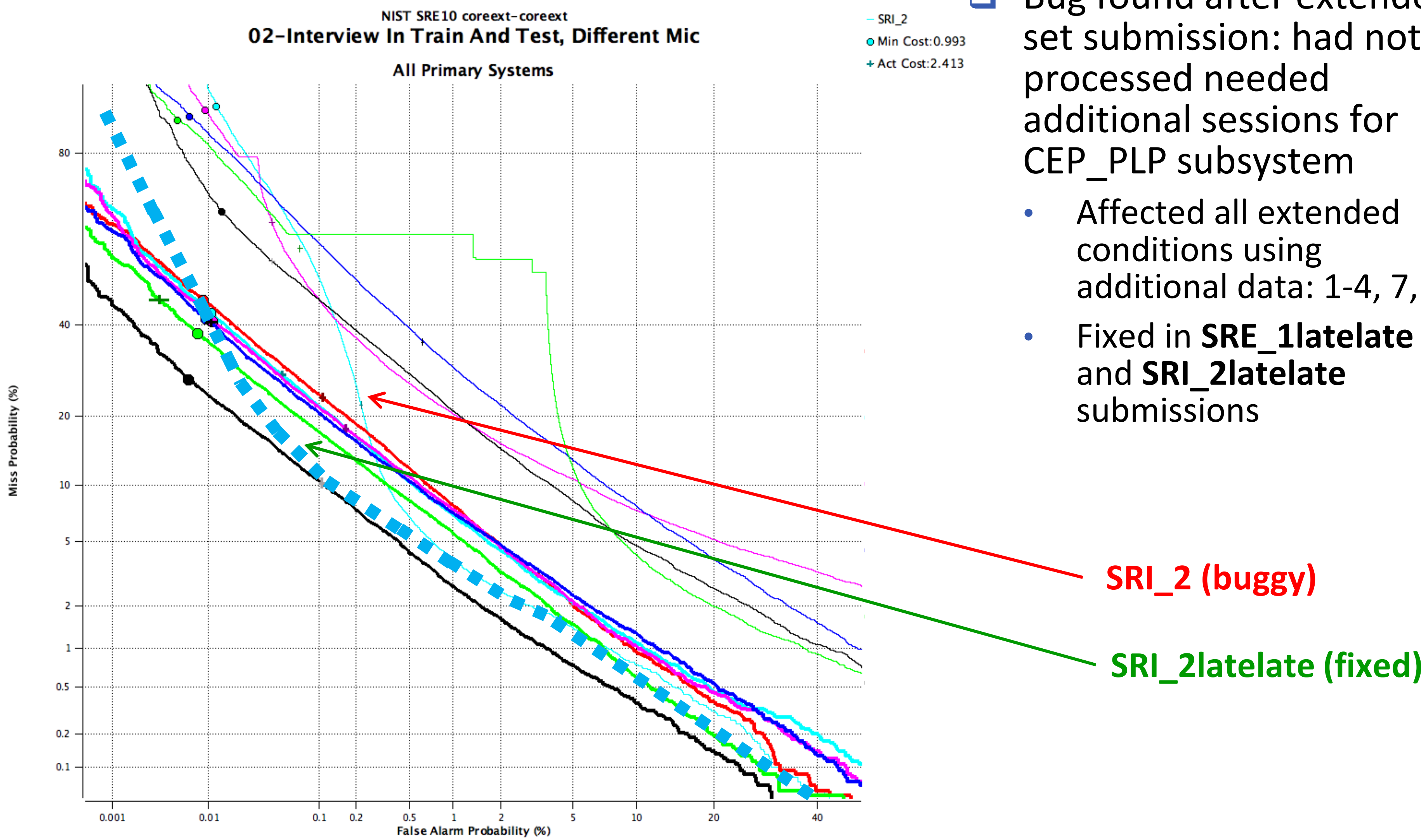
Talk Outline

- Introduction
  - SRI approach to SRE10
  - System overview
  - Development data design
- System description
  - Individual subsystems
  - VAD for microphone data
  - System combination
- SRE results and analysis
  - Results by condition
  - N-best system combinations
  - Errors and trial (in)dependence
  - Effect of bandwidth and coding
  - Effect of ASR quality
- Summary

Introduction: SRI Approach

- Historical focus
  - Higher-level speaker modeling using ASR
  - Modeling many aspects of speaker acoustics & style
- For SRE10: Two systems, multiple submissions
    - **SRI_1**: 6 subsystems, plain combination, ASR buggy on some data (Slide 35)
    - **SRI_2**: 7 subsystems, side-info for combination
    - **SRI_1fix**: same as *SRI_1* with completed ASR bug fix
  - Some additional systems were discarded for not contributing in combination
  - Submission was simplified by the fact that eval data was all English
- Excellent results on the traditional tel-tel condition
- Good results elsewhere, modulo bug in extended trial processing
- Results reported here are after all bug fixes, on the *extended core* set (unless stated otherwise)

Extended Trial Processing Bug



- Bug found after extended set submission: had not processed needed additional sessions for CEP_PLP subsystem
  - Affected all extended conditions using additional data: 1-4, 7, 9.
  - Fixed in **SRE_1latelate** and **SRI_2latelate** submissions

Overview of Systems

Feature	ASR-independent	ASR-dependent
Cepstral	**MFCC GMM-SV** **Focused MFCC GMM-SV**	**Constrained MFCC GMM-SV** **PLP GMM-SV**
MLLR		MLLR
Prosodic	**Energy-valley regions GMM-SV** **Uniform regions GMM-SV**	**Syllable regions GMM-SV**
Lexical		Word N-gram SVM

- Systems in **red** have improved features
- Note: prosodic systems are precombined with fixed weights
  - We treat them as a single system

Development Data - Design

- **Trials:** Designed an extended development set from 2008 original and follow up SRE data
  - Held out 82 interview speakers
  - Models and tests are the same as in SRE08
  - Paired every model with every test from a different session (exception: target trials for tel-tel.phn-phn condition were kept as the original ones)
  - Created a new shrt-long condition
  - Corrected labeling errors as they were discovered and confirmed by LDC
- **Splits:**
  - Split speakers into two disjoint sets
  - Split trials to contain only speakers for each of these sets
  - Lost half of the impostor trials, but no target trials
  - Use these splits to estimate combination and calibration performance by cross-validation
- For BKG, JFA and ZTnorm, different systems use different data, but most use sessions from SRE04-06 and SWBD, plus SRE08 interviews not used in devset.

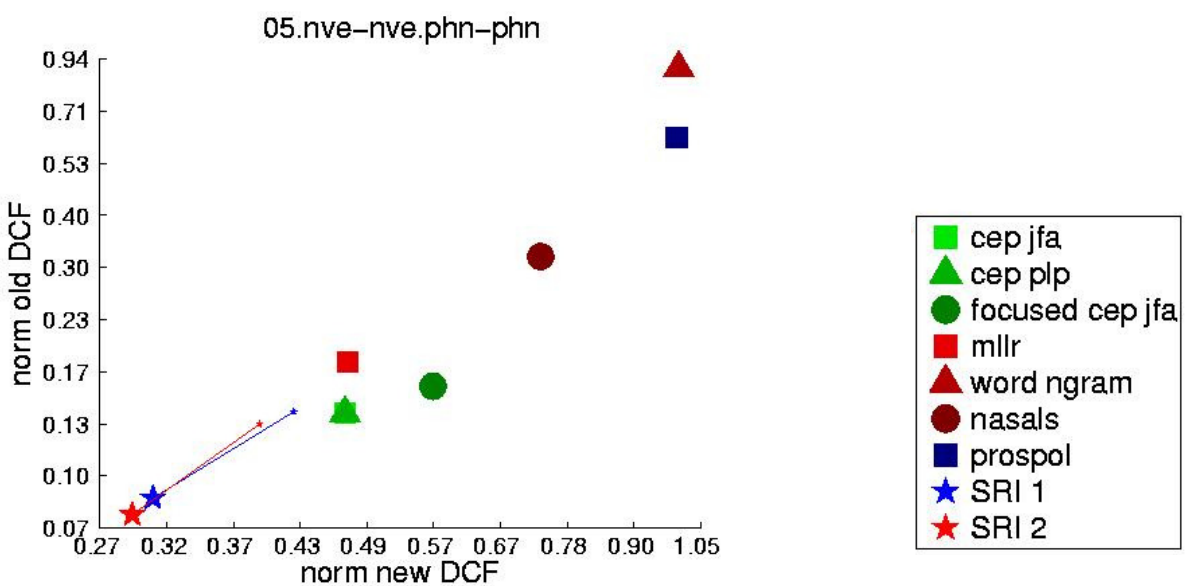
Development Data – Mapping to SRE

- Dev trials used for combination and calibration chosen to match as well as possible the conditions in the SRE data
  - Duration and microphone conditions of train and test matched pretty well
    - We cut the 24 and 12 min interviews into 8 minutes
  - When necessary, the style constraint is relaxed (interview data is used for telephone convs)

TRAIN-TEST Duration.Style.Channel	#trials	%target	Used for SRE trials
long-long.int-int.mic-mic	330K	3.0	long-long.int-int.mic-mic (1, 2)
shrt-long.int-int.mic-mic	347K	3.0	shrt-long.int-int.mic-mic (1, 2)
long-shrt.int-int.mic-mic	1087K	3.0	long-shrt.int-***.mic-mic (1, 2, 4)
shrt-shrt.int-int.mic-mic	1143K	3.0	shrt-shrt.***-***.mic-mic (1, 2, 4, 7, 9)
long-shrt.int-tel.mic-phn	777K	0.2	long-shrt.int-tel.mic-phn (3)
shrt-shrt.int-tel.mic-phn	822K	0.2	shrt-shrt.int-tel.mic-phn (3)
shrt-shrt.tel-tel.phn-phn	1518K	0.1	shrt-shrt.tel-tel.phn-phn (5,6,8)

Format of Results

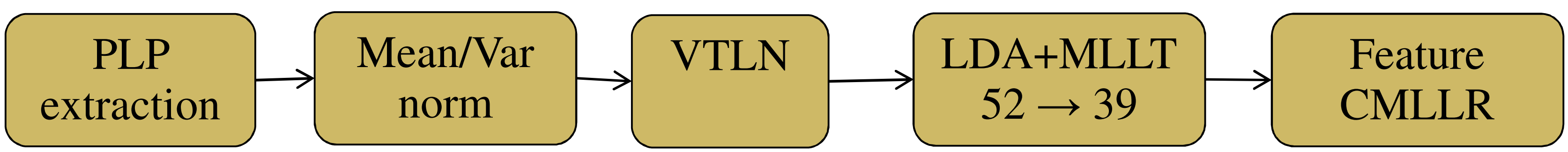
- We show results on the extended trial set
- Scatter plot of cost1 (normalized min new DCF, in most cases) versus cost2 (normalized min old DCF, in most cases)
- In some plots, for combined systems we also show actual DCFs (linked to min DCFs by a line)
- Axes are in log-scale



System Description

Cepstral Systems Overview

- All cepstral systems use the Joint Factor Analysis paradigm
  - MFCC System
    - 19 cepstrum + energy + Δ + ΔΔ
    - Global CMS and variance normalization, no gaussianization
  - PLP System:
    - Frontend optimized for telephone ASR
    - 12 cepstrum + energy + Δ + ΔΔ + ΔΔΔ, VTLN + LDA + MLLT transform
    - Session-level mean/var norm



    - CMLLR feature transform estimated using ASR hypotheses

- 3 cepstral systems submitted, others in stock
  - 2 MFCC systems: 1 GLOBAL, 1 FOCUSED
  - 1 PLP system: 1 FOCUSED

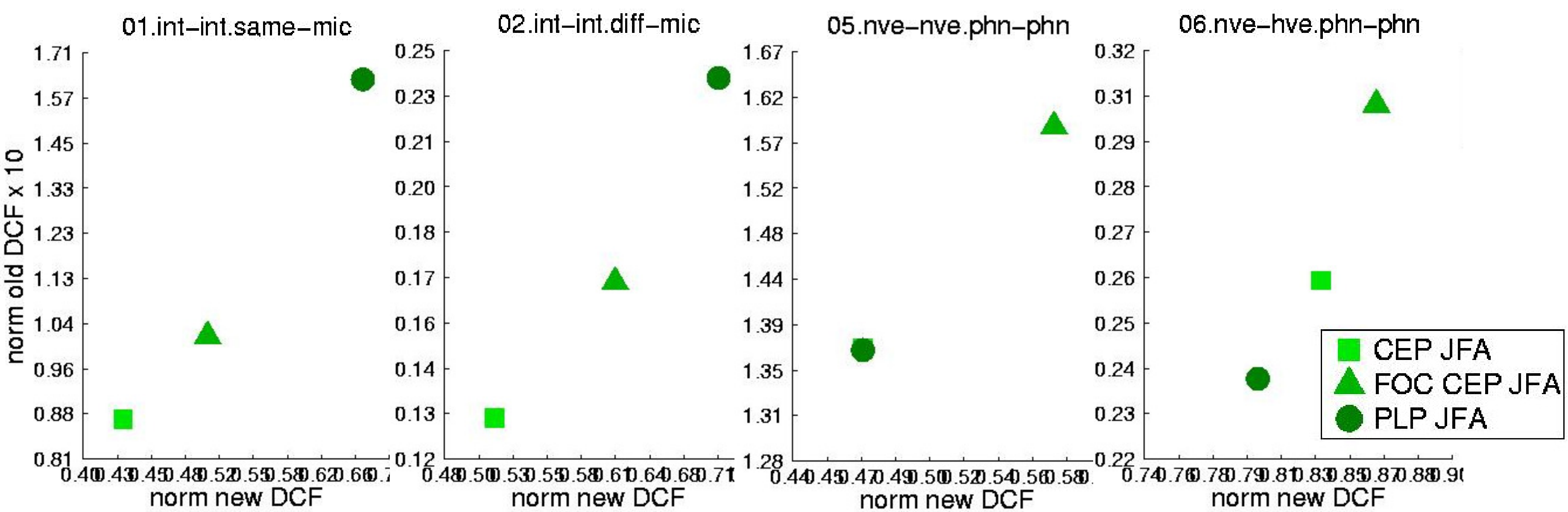
Cepstral Systems: Global vs. Focused

- Promoting system diversity
  - Two configurations: *global* versus *focused*
  - Global does not take any class or condition into account *(except gender-dependent ZTnorm)*

	Global	*data used*	Focused	*data used*
UBM	1024		512	
Gender	No		Yes	
E-voices	600	*SRE+SWB*	400 (500)	*SRE+SWB*
E-channels	500 *300 tel 200 int*	*SRE04,05,06 Dev08, SRE08 HO*	455 (300*3) *150 tel, 150 mic, 150 int, 5 voc*	*SRE04,05,06,08HO Dev08, dev10*
Diagonal	Yes	*04,05,08HO*	No	
ZTnorm	Global	*SRE04,05,06*	Condition-dependent	*SRE04,05,06,08HO*

Cepstral Systems: Performance

- Eval results for SRI's 3 cepstral systems
  - CEP_JFA is the best performing system overall
  - CEP_PLP has great performance on telephone
    - System performs worse on interview data
    - Due to poorer ASR and/or mismatch with tel-trained CMLLR models

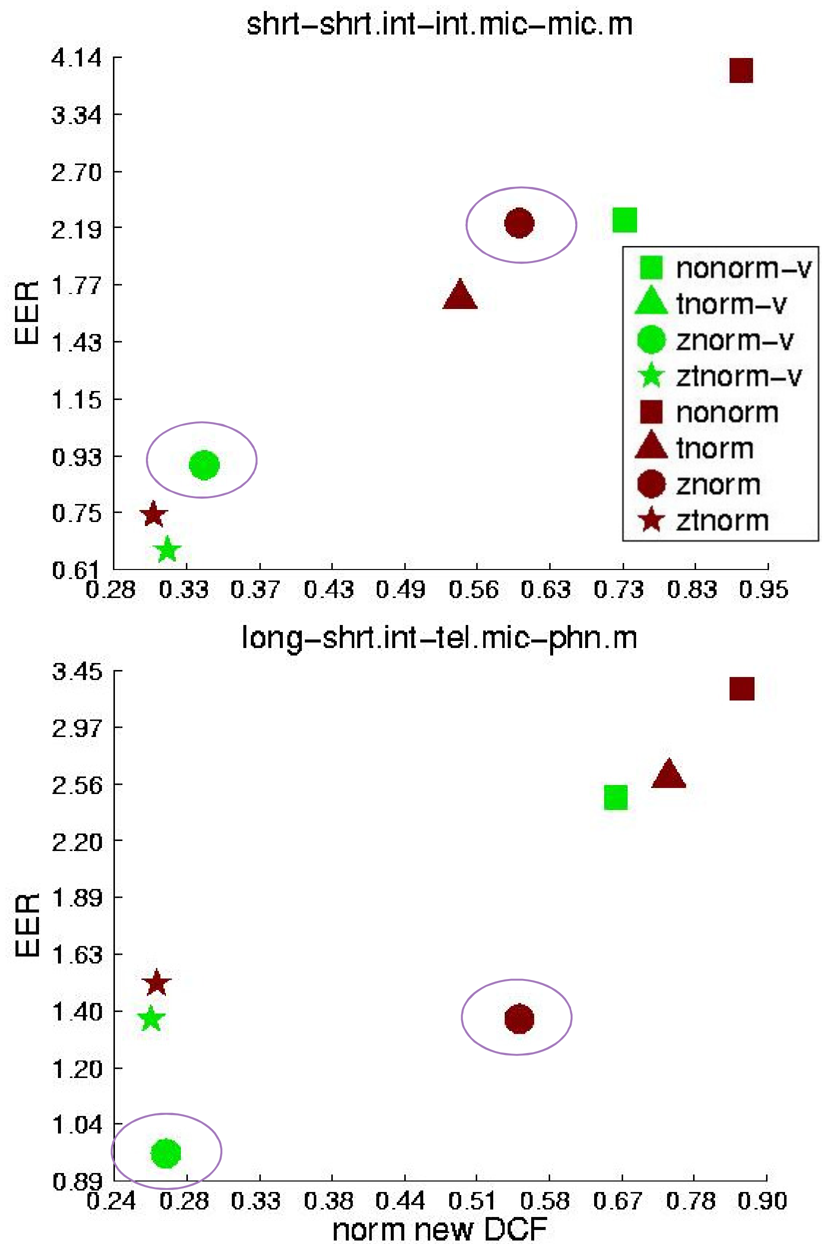


Tnorm for Focused Systems

- Speaker models are distributed among N(0,I) (speaker factors)
  - Synthetic Tnorm uses sampling to estimate the parameters
  - Veneer Tnorm computes the expected mean/var

$$y V \hat{F}, y \sim N(0, I)$$

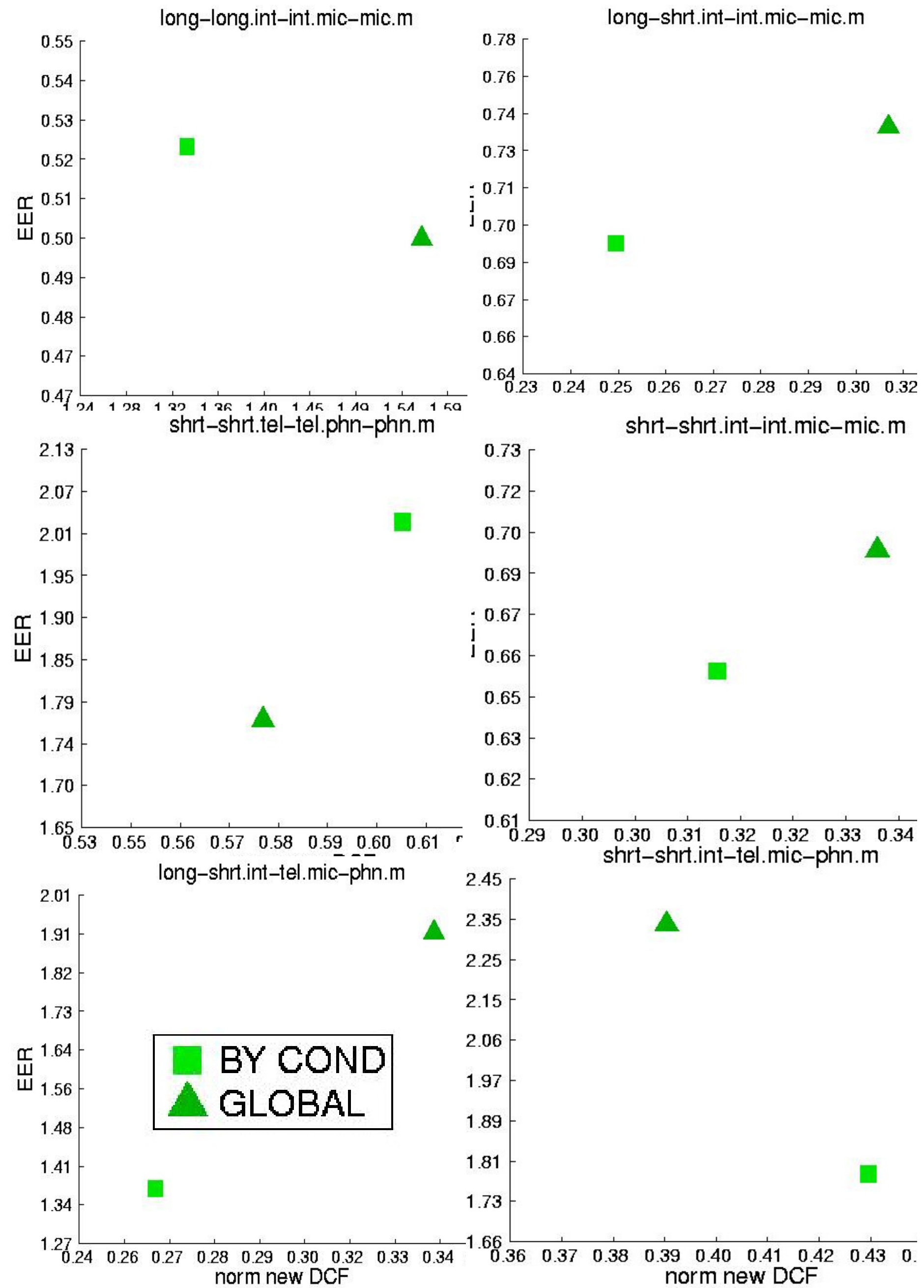
  - Impostor mean is 0
  - Impostor variance is the norm of $V\hat{F}$
  - Can replace/be used on top of Tnorm
  - ***Large effect after Znorm***
- Justification for the cosine kernel in i-vector systems?



Condition-Dependent ZTnorm

- Match Znorm/Tnorm data sources to the targeted test/train condition
  - Significant gain or no loss in most conditions
  - Only loss in ***tel-tel*** condition *(global ztnorm uses 3 times more data)*

Trial	Matched Impostors
TRAINING (eg: short, tel)	TNORM short, tel
TEST (eg: long, mic)	ZNORM long, mic



On the Cutting Room Floor …

- i-Vector
  - 400 dimensional i-vector followed by LDA+WCCN. Generated by a 2048 UBM trained with massive amount of data.
  - Results comparable to baseline, brought nothing to combination
- i-Vector complement
  - Use the total variability matrix as a nuisance matrix
  - Great combination w/system above, no gain in overall combination
- Superfactors
  - Gaussian-based expansion of the speaker factors, symmetric scoring
  - No gain in combination
- Full-covariance UBM model
  - Small number of Gaussians (256), complexity in the variances
  - Error rate too high, needs work on regularization and optimization

Improved VAD for Mic/Interview Data

- Evaluated use of distant-mic speech/nonspeech models (trained on meetings)
- Explored use of NIST-provided ASR as a low-false-alarm VAD method
- Back-off strategy (from ASR to frame-based VAD) depending on ratio of detected speech to total duration (as in Loquendo SRE08 system)
- Evaluated oDCF/EER on SRE08 short mic sessions, using old cepstral system

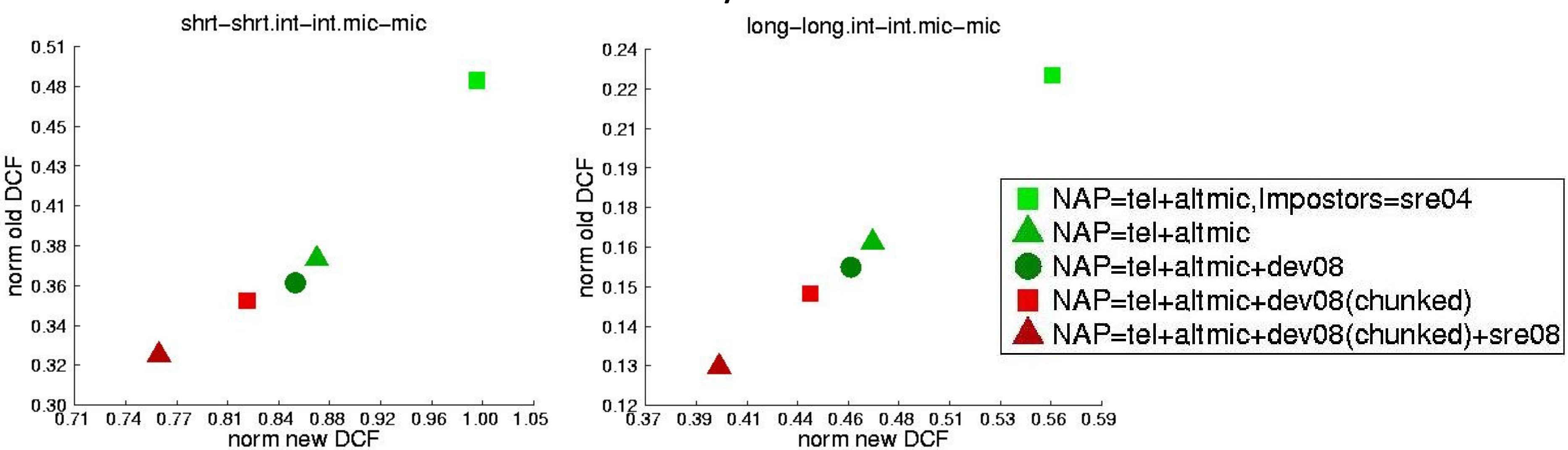
VAD Method	Interview	Ph. Convs.	
NIST VAD (SRI SRE08 method)	.173 / 3.8		
Combine NIST ASR and NIST VAD with backoff	.160 / 3.0		
Telephone VAD (no crosstalk removal)	.210 / 4.1	**.188 / 5.2**	
Distant-mic VAD (no crosstalk removal)	.202 / 4.0	.302 / 8.0	
Telephone VAD, remove crosstalk w/ NIST ASR	.170 / 3.3		
Distant-mic VAD, remove crosstalk w/ NIST ASR	**.160 / 3.1**		← "Fair"
Combine NIST ASR and dist-mic VAD w/ backoff	**.157 / 3.0**		← used for SRE10

VAD Result Summary

- Conclusions so far:
  - Using ASR information from the interviewer channel is critical for good results
  - For interviews, it is slightly better to use VAD models trained for distant microphones (from 8kHz-downsampled meeting data)
  - But for phonecalls, the telephone-trained VAD models work better, in spite of capturing 53% more speech. It could be that models work better if only high-SNR speech portions are used.
- Interviewer ASR with distant-mic VAD is a winner because
  - It is "fair": close mic for the interviewer, but distant mic for the speaker of interest
  - Works much better than distant-mic VAD by itself
  - Gives results close to those obtained with “cheating” close-mic ASR on interviewee

MLLR SVM

- Raw features same as in SRI's English-only MLLR system in SRE08
  - PLP-based, LDA & MLLT & CMLLR for normalization
  - (8 phone classes) x ("male", "female") transforms
  - 24,960 feature dimensions, rank-normalized
- Impostor data updated with SRE06 tel+altmic and SRE08 interviews
  - Previously used SRE04 only
- NAP data augmented with interviews for SRE10
  - "chunked" dev08 interviews into 3-minute pseudo-sessions
  - 48 nuisance dimensions
- Added ZT-normalization – actually hurt on SRE10 data!

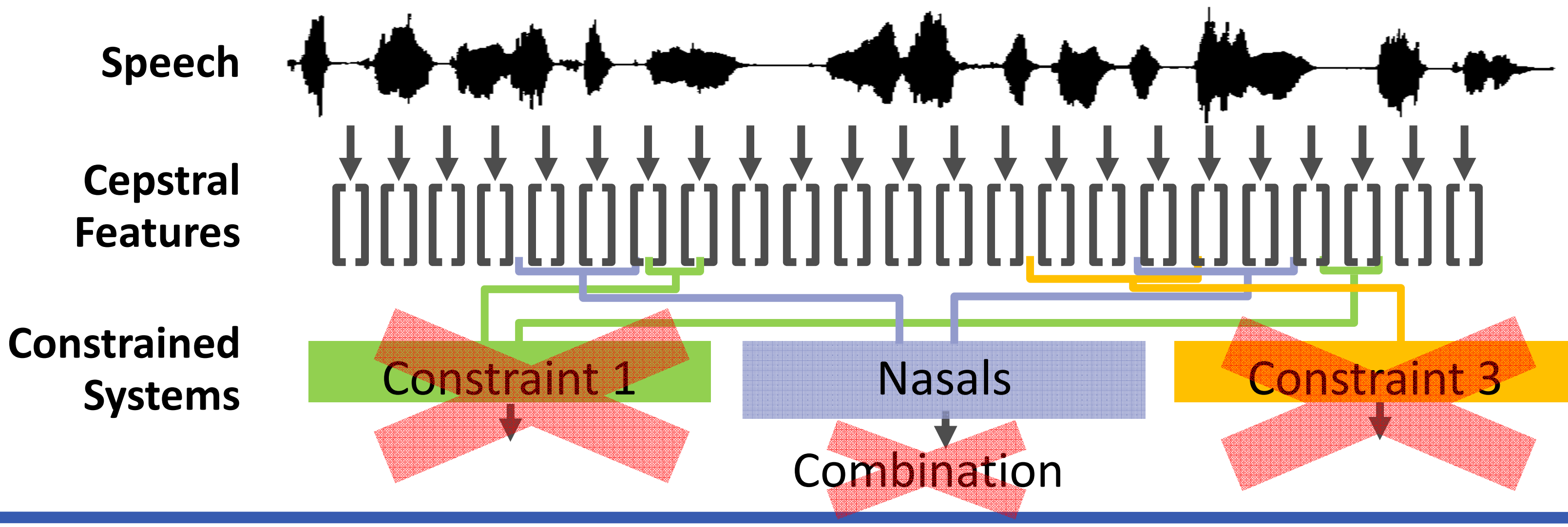


Word N-gram SVM

- Based on English ASR, which was unchanged from SRE08
  - But benefits from better interview VAD
- 9000 most frequent bigrams and trigrams in impostor data, features are rank-normalized frequencies
- Added held-out SRE08 interviews to SRE04 + SRE05 impostors
  - Minimal gains
- Score normalization didn’t help, was not used
- Word N-gram in combination helps mainly for telephone-telephone condition
  - But that could change if better ASR for interviews is used
  - See analysis reported later

Constrained Cepstral GMM (Nasals System)

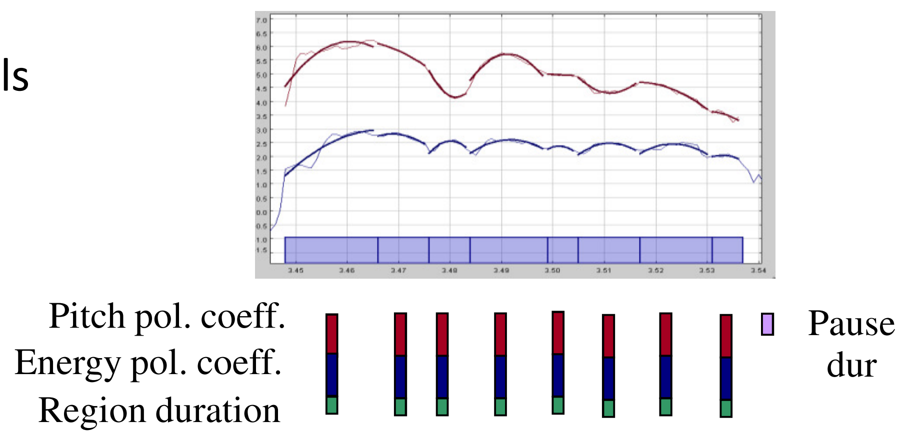
- Idea: use same cepstral features, but filter and match frames in train/test
- Linguistically motivated regions; combine multiple regions since each is sparse
- But: our constrained system was itself "constrained" due to lack of time and lack of training data for reliable constraint combination . . . .
- So only a **single constraint** was used in SRE10: **syllables with nasal phones**
  - Constraint captures 12% of frames (after speech/nonspeech segmentation)
  - UBM = 1024 Gaussians (from unconstrained CEP_JFA)
  - JFA = 300 eigenchannels, 600 eigenvoices, diagonal term (from CEP_JFA)



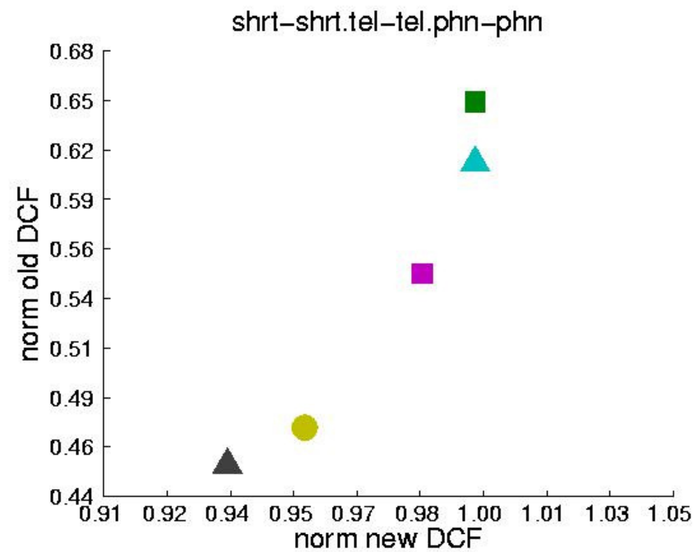
SRI International

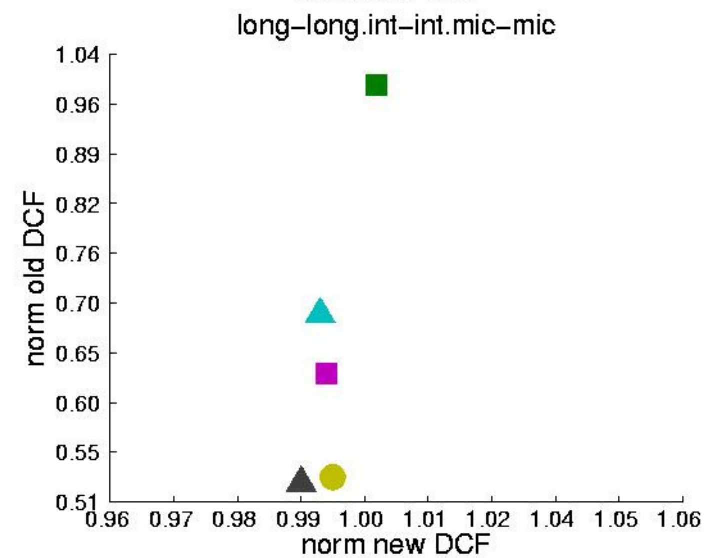
Prosodic System

- Pitch and energy signals obtained with get_f0
  - Waveforms preprocessed with a bandpass filter (250-3500)
  - No Wiener filtering used (did not result in any gains)
- Features: Order 5 polynomial coefficients of energy and pitch, plus length of region (Dehak'07)
- Regions: Energy valley, uniform regions and syllable regions (New) (Kockmann '10)
- GMM supervector modeling:
  - JFA on gender-dependent GMM models
  - 100 eigenvoices, 50 eigenchannels (963 females, 752 males)
  - **New:** modeling of bigrams (Ferrer '10)



Prosodic Systems - Results





Results on development data

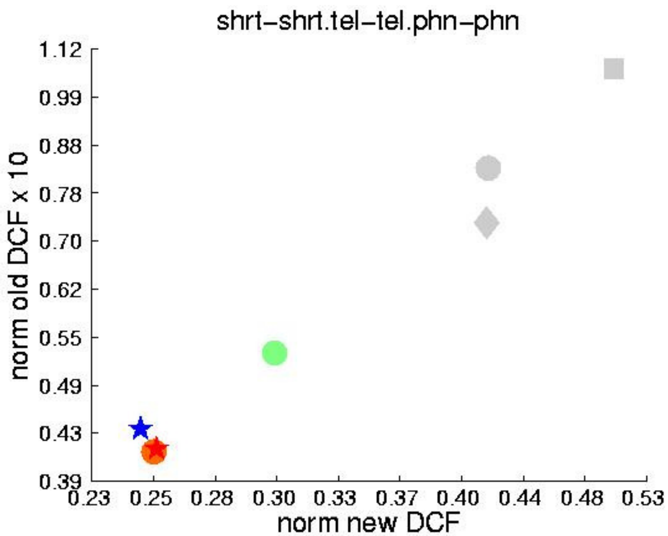
- Showing two conditions with different behavior
  - Others are very similar to long-long.int-int.mic-mic
- Regions:
  - ev (energy valley)
  - un (uniform, 30ms with a shift of 10 ms)
  - syl (syllables)
- Very small gains in new DCF, but in old DCF:
  - Big gain from sre08 system due to addition of SWBD and held-out interview data
  - Additional gains from adding bigrams (2g) and uniform regions
  - Smaller gains from adding syllable regions

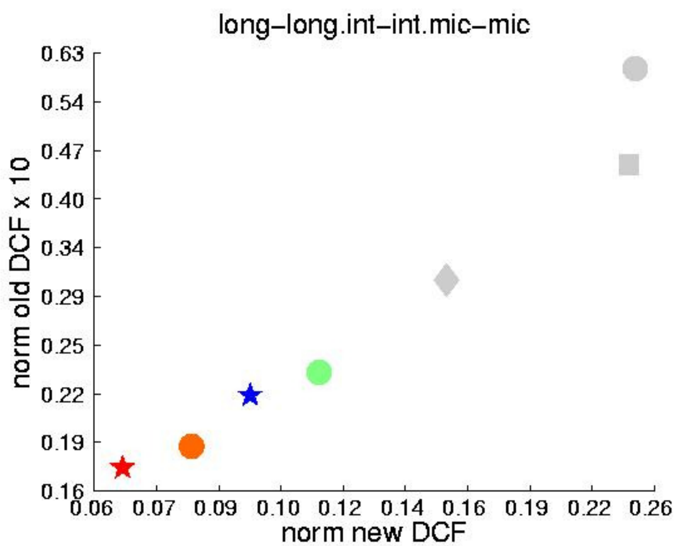
→ System used in submissions (prospol)

Combination Procedure

- Linear logistic regression with metadata (ICASSP'08)
  - Metadata used to condition weights applied to each system
- SRI_1 uses no metadata
- SRI_2 uses:
  - Number of words detected by ASR (<200, >200)
  - SNR (<15, >15)
  - Also tried RMS, nativeness, gender, but they did not give gains
- In both cases, the combiner is trained by condition (duration, speech style and channel type) as indicated in earlier slide
- *Apropos nativeness:* it used to help us a lot, but not on new dev set and with new systems, so was not used
  - Current lack of gain probably due to improvements in our systems that made them more immune to nonnative accents
  - Also: classifier scores on SRE10 data show almost no nonnatives

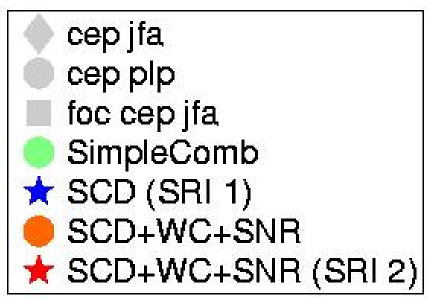
Combination Results





Results on development data

- Showing two conditions with different behavior
  - Others are somewhat similar to one or the other
- SimpleComb: single set of weights for all trials
- SCD: separate combiner trained for each combination of **S**peech, **C**hannel and **D**uration conditions
- SCD+WC+SNR: using metadata within each condition

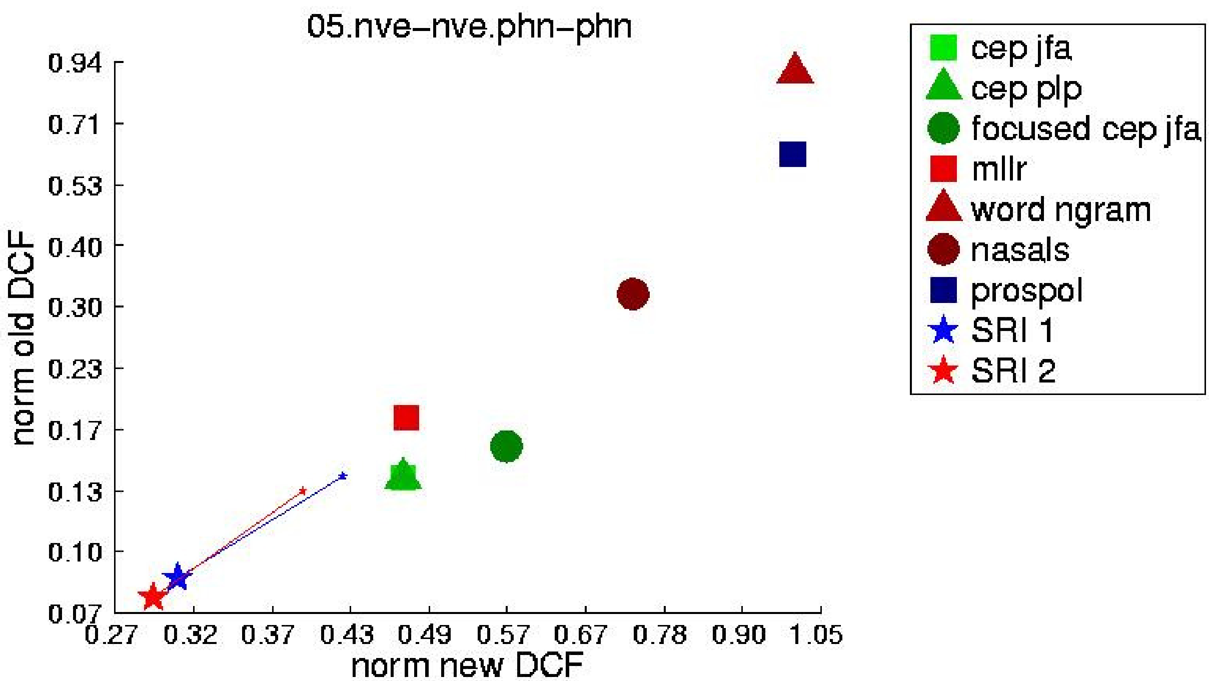


} Using 6 systems

→ Using 7 systems (6 above + nasals)

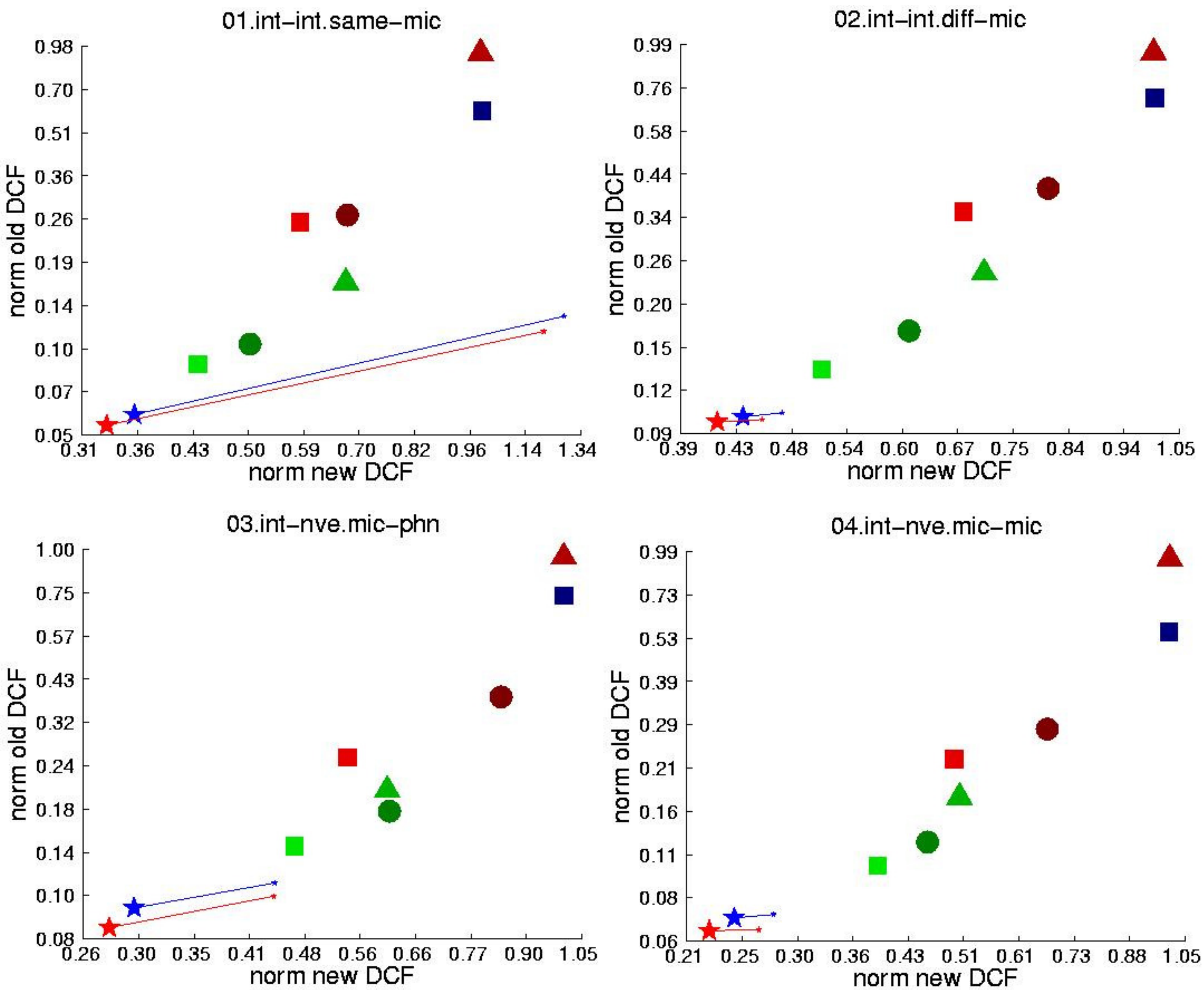
SRE Results and Analysis

Results for Condition 5

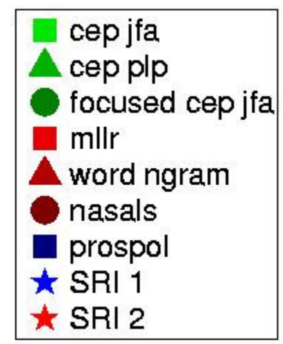


- Both combinations outperform individual systems by around 35%
- SRI_2 outperforms SRI_1 by around 5%

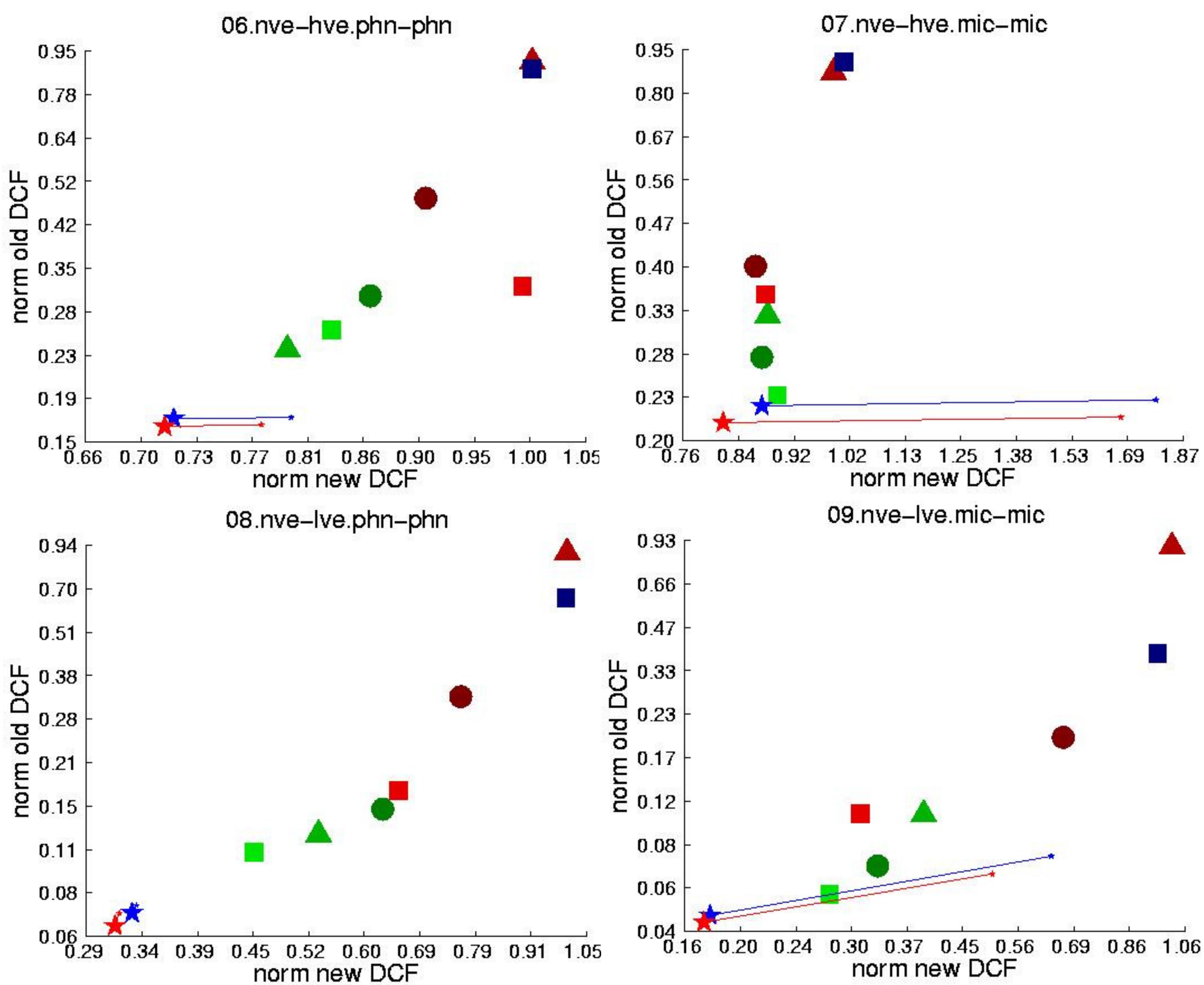
Results for Conditions 1-4



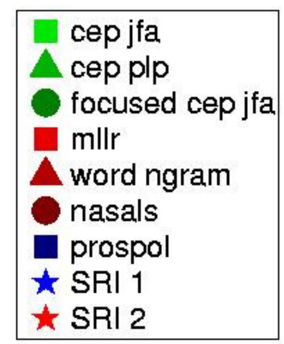
- Reasonable calibration for all conditions, except for 01
- This was expected, since we did not calibrate with same-mic data



Results for Conditions 6-9

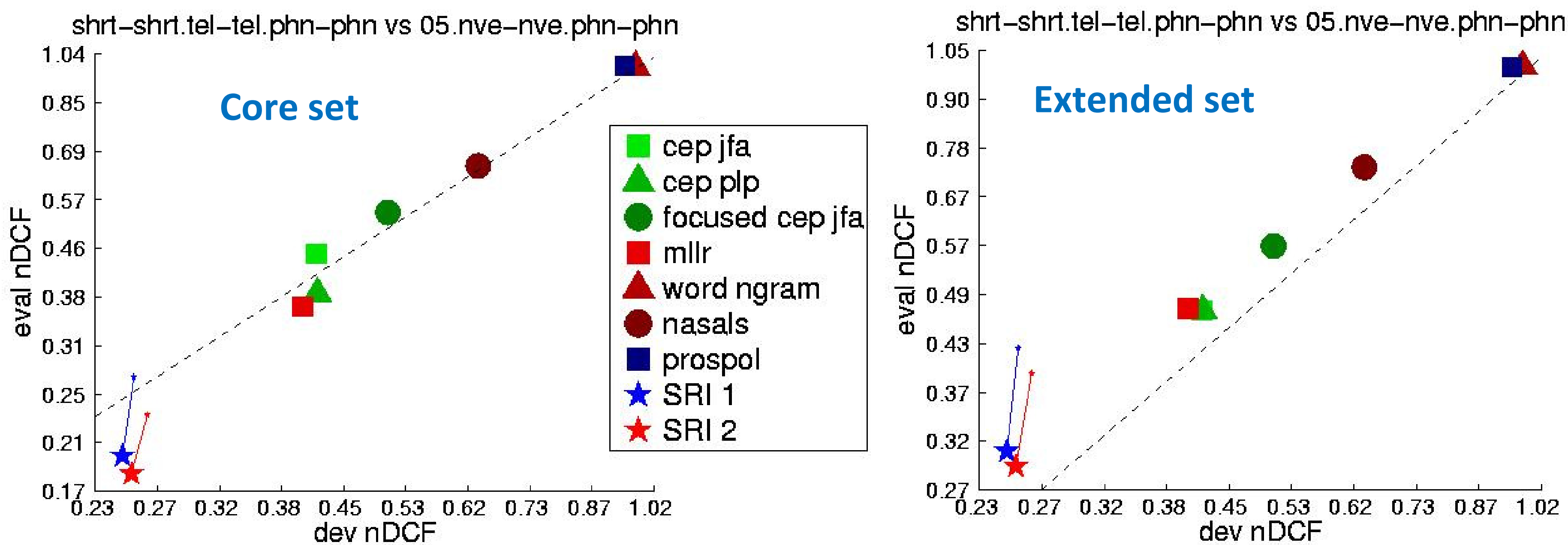


- Good calibration for phn-phn (surprising!)
- For mic-mic, we used mismatched style and matched channel
- Reversing this decision gives even worse calibration!



Development versus SRE Results

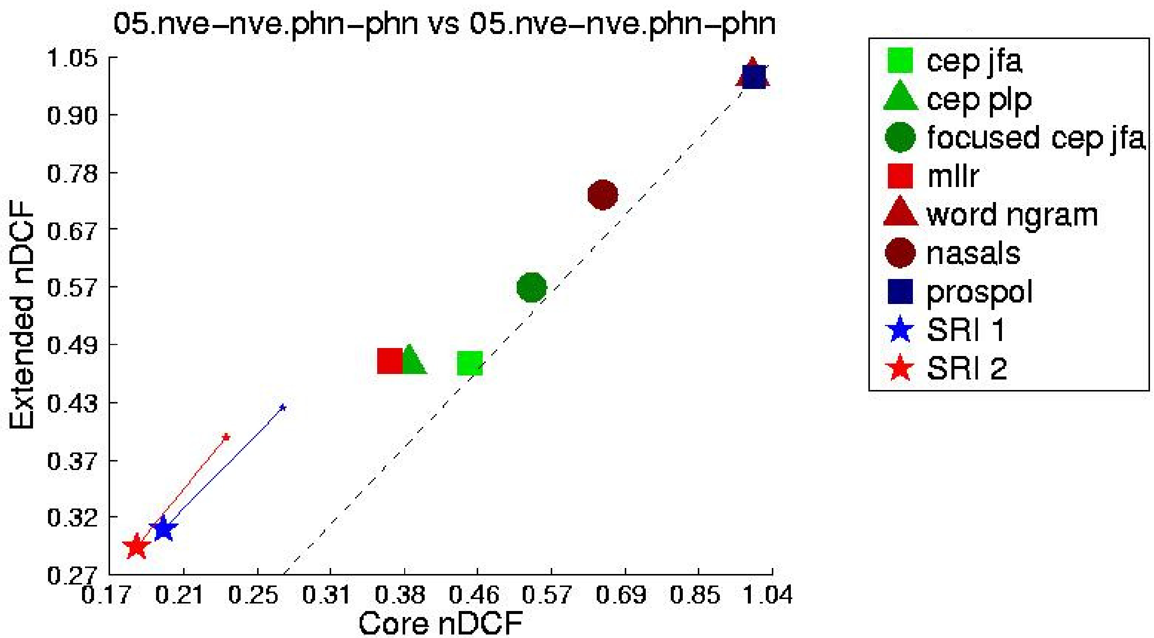
- How did individual subsystems and their combination generalize?
- Condition 5 has perfectly matched development set



- Reasonably good generalization of performance
- Core set easier than dev set for cep_plp and mllr systems
- Extended set harder than dev for all systems

Extended versus Core Results

- ❑ Our extended results on most conditions are worse than the core results (especially on conditions 5, 6, 7 and 8)



- ❑ Showing results on condition 5
- ❑ Figures for other conditions available in additional slides

- ❑ The two best systems degrade around 25% from core to extended set
- ❑ This results in a degradation of the combination performance
- ❑ From the better systems these are the two that rely on PLP and ASR.
  - Does the extended set contain more noisy sessions? More investigation needed …

N-Best Systems by Condition (New DCF)

All 7 systems

01.int-int.same-mic

.329	cep	mllr	nasal	foc
.432	X			
.309	X	X		
.284	X	X	X	
.279	X	X	X	X

02.int-int.diff-mic

.421	cep	mllr	nasal	foc
.514	X			
.404	X	X		
.395	X	X	X	
.389	X	X	X	X

03.int-nve.mic-phn

.298	cep	mllr	plp	foc
.468	X			
.333	X	X		
.308	X	X	X	
.298	X	X	X	X

04.int-nve.mic-mic

.237	cep	mllr	pros	plp
.388	X			
.273	X	X		
.256	X		X	X
.240	X	X	X	X

N-Best Systems by Condition (New DCF)

05.nve-nve.phn-phn

.305	plp	mllr	foc	ngrm
.471	X			
.345	X	X		
.310	X	X	X	
.298	X	X	X	X

06.nve-hve.phn-phn

.713	plp	nasal	foc	ngrm
.798	X			
.710	X	X		
.658	X		X	X
.645	X	X	X	X

07.nve-hve.mic-mic

.858	nasal	plp	mllr
.862	X		
.777	X	X	
.768	X	X	X

08.nve-lve.phn-phn

.329	cep	plp	mllr	ngrm
.450	X			
.372	X	X		
.346	X	X	X	
.332	X	X	X	X

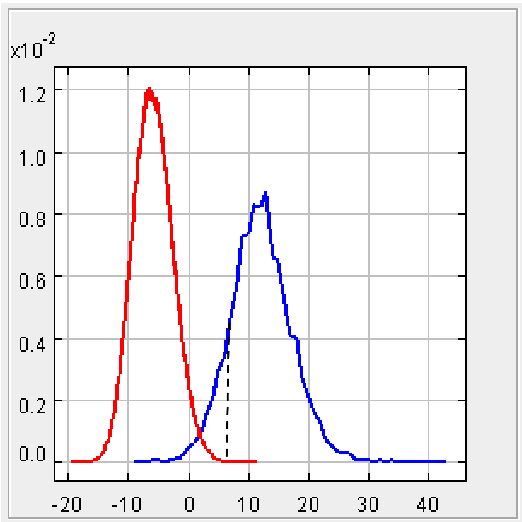
09.nve-lve.mic-mic

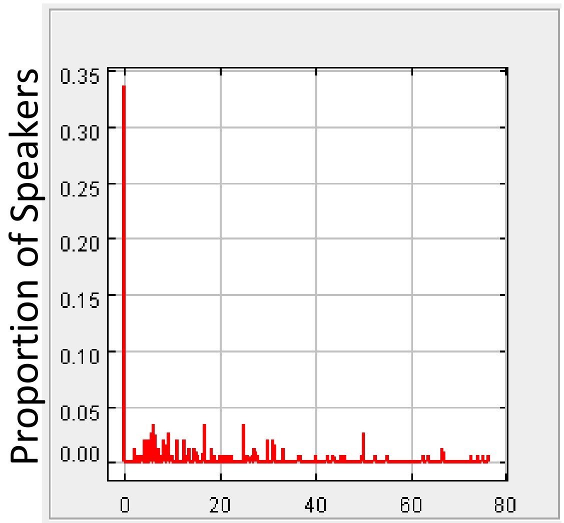
.166	cep	mllr	pros
.274	X		
.187	X	X	
.145	X	X	X

N-Best Analyses: Summary

- For many conditions, < 7 systems better than all 7 systems (best usually about 4 systems)
- But, different systems good at different conds.
- System ordering usually cumulative
- CEP_JFA or CEP_PLP usually the best single system – except for cond. 7
- CEP_PLP superior on telephone data (PLP frontend was optimized for telephone ASR)
- Focused cepstral system can help when only one other cepstral system present
- MLLR best 2nd or 3rd system, except cond 6
- Prosody, Nasals, Word N-gram complement Cepstral and MLLR systems
- Nasals seem to help high vocal effort → try other constraints, vocal effort as side info

Analysis of Errors on Condition 5





- Histogram of %misses per speaker (at the new DCF threshold)
  - Only showing speakers that have at least 10 target trials
- Around 34% of speakers have 0% misses
- For other speakers, up to 75% of the target trials are missed

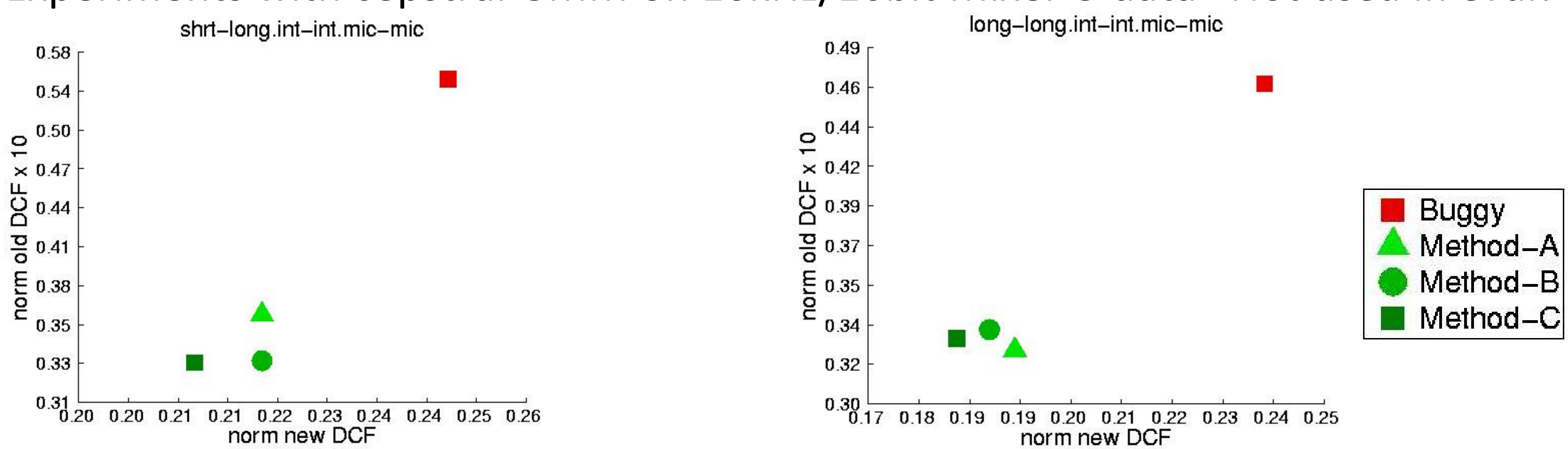
- Hence: misses produced by systems are highly correlated
  - Significance measures that assume independence are too optimistic
- Nevertheless, false alarms do seem to be pretty independent of speaker and session
  - From the 78 false alarms, 62 come from different speaker pairs (even though, on average, there are 4 trials per speaker pair)
  - Worth creating the extended set, which mainly generates additional impostor samples

Bandwidth/Coding of Interview Data

- 4 days before submission, found a bug in our microphone data processing:
  16kHz/16bit $\Rightarrow$ 8kHz/16bit $\Rightarrow$ 8kHz/8bit-μlaw $\Rightarrow$ 8kHz/16bit $\Rightarrow$ Wienerfilter $\Rightarrow$ 8kHz/**8bit-μlaw**
  Low amplitude signals are coded using only 1-2 bits, leading to bad distortion **(Buggy)**

@NIST

- Correct processing: **(Method A)**
  16kHz/16bit $\Rightarrow$ 8kHz/16bit $\Rightarrow$ 8kHz/**8bit-μlaw** $\Rightarrow$ 8kHz/16bit $\Rightarrow$ Wienerfilter $\Rightarrow$ 8kHz/16bit
  But: coding of low amplitudes still potentially problematic!
- Better yet (proposed for future SREs):
  16kHz/16bit $\Rightarrow$ **8kHz**/16bit $\Rightarrow$ Wienerfilter $\Rightarrow$ 8kHz/16bit **(Method B)**
  16kHz/16bit $\Rightarrow$ Wienerfilter $\Rightarrow$ 16kHz/16bit $\Rightarrow$ 8kHz/16bit **(Method C)**
- Experiments with cepstral GMM on 16kHz/16bit Mixer-5 data *Not used in eval!*

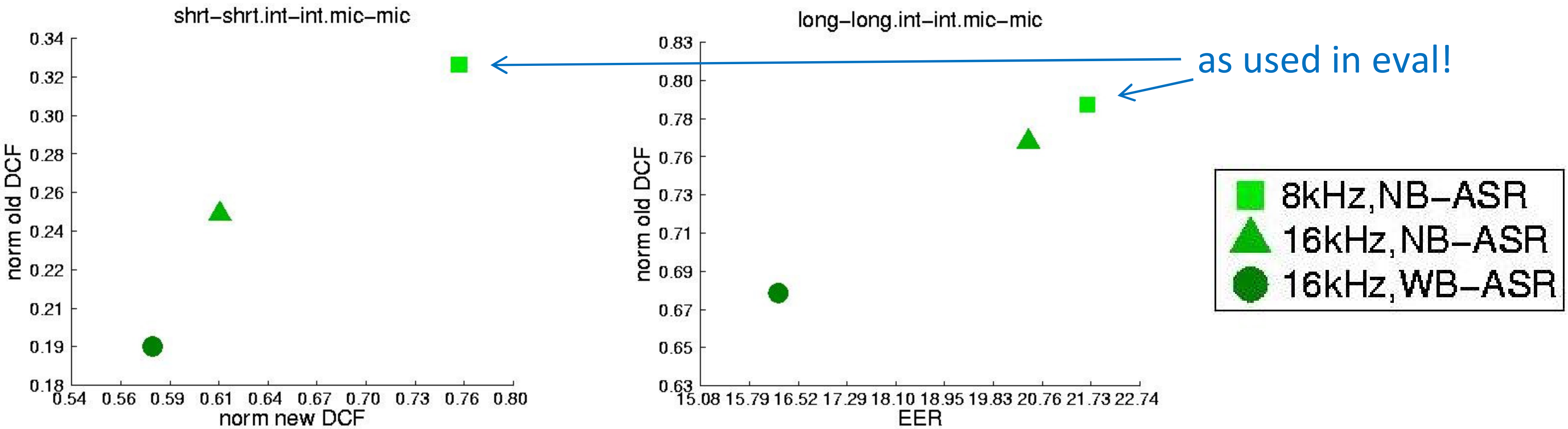


SRI International

BW/Coding & ASR-based Systems

- ASR-based systems can benefit twofold from wideband data:
  - Less lossy coding (no μlaw coding, better noise filtering at 16kHz)
  - Better ASR using wideband (WB) recognition models
  - Even though cepstral speaker models need to be narrowband (NB) for compatibility with telephone data in training
- Experiments using WB ASR system trained on meeting data
  - Showing one representative condition each for 2 ASR-dependent systems

MLLR SVM **Word N-gram SVM**

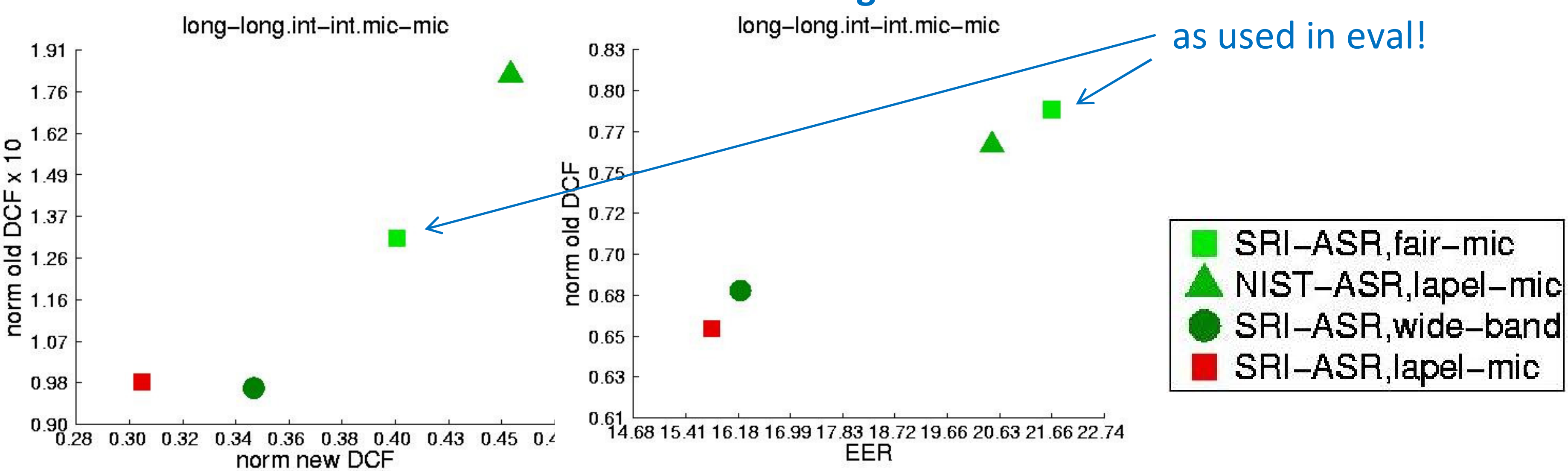


SRI International

Effect of ASR Quality

- What is effect of ASR quality on high-level speaker models?
  - How much “cheating” is it to use lapel microphone for ASR?
  - But segmentation is held constant, so we’re underestimating the effect
- Result: using lapel mic (CH02) for ASR leads to dramatic improvements, similar to using wideband ASR on true mic
- Using NIST ASR gives poor results by comparison (not sure why)

MLLR SVM **Word N-gram SVM**



Summary

- Created dev set (shared with sre10 Google group)
  - Tried to match eval conditions
  - Generated additional trials for evaluating new DCF more reliably
  - Fixed labeling errors (as confirmed by LDC)
- System description
  - Improved interview VAD by utilizing both distant-mic speech models and NIST ASR
  - Subsystems: 3 cepstral, mllr, prosody, nasals, word n-gram
  - PLP system excels on telephone data
  - Prosody modeling improvements (Ferrer et al. ICASSP paper)
  - System combination with side information (sample duration, channel/genre, no. words, SNR)
- Results by condition good to excellent
  - Poor calibration for same-mic interview and LVE/HVE mic-mic conditions (due to lack of matched training data)
  - SRE_2 system validates benefit of constrained (nasals) GMM & combiner side info
  - Extended set harder than core in most conditions (still trying to figure out why)

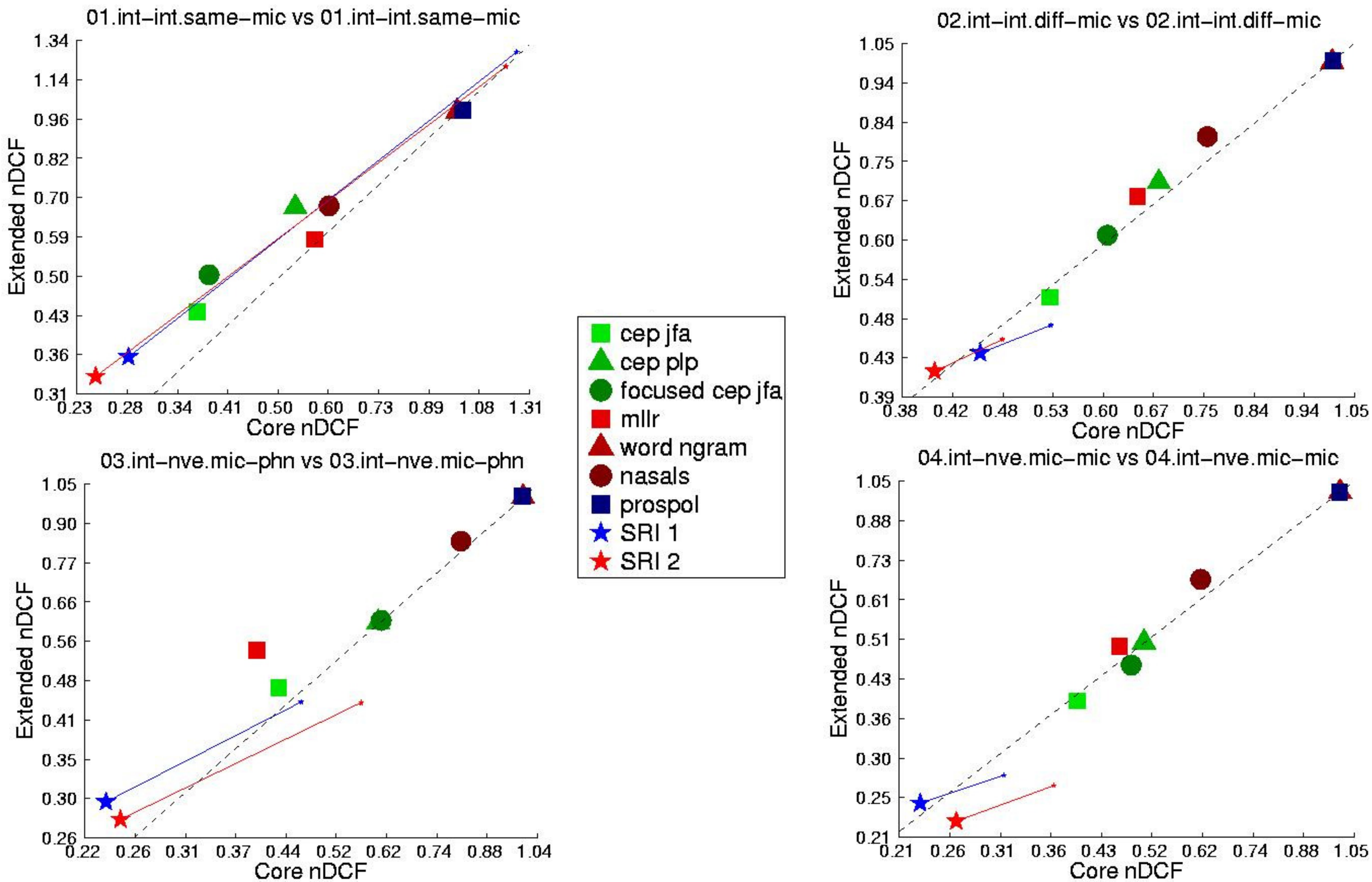
Post-Eval Analyses

- N-best system combinations:
  - Different subsystems are good for different conditions
  - Typical pattern: 1 cepstral plus MLLR, followed by other systems
  - Using all systems everywhere hurt us, but different subsets by condition was considered too complicated
  - Interesting correlations between subsystems and conditions worthy of more study
- Miss errors highly correlated as a function of speaker
  - But false alarms fairly independent of speaker and session
- Bandwidth and μlaw coding hurts performance on interviews significantly
  - We advocate NIST distribute full-band data in the future
- Using only close-talking mics for ASR is overly optimistic
  - ASR-based models perform much better than in realistic conditions

Thank You

http://www.speech.sri.com/projects/verification/SRI-SRE10-presentation.pdf

Extended versus Core Results



Extended versus Core Results