SRI's NIST 2005 Speaker Recognition Evaluation System

Sachin Kajarekar, Luciana Ferrer, Elizabeth Shriberg,
Kemal Sönmez, Andreas Stolcke, Anand Venkataraman

SRI International, Menlo Park, CA, USA

Acknowledgements

Collaborators: Harry Bratt (SRI), Yang Liu (ICSI)

ICSI systems: N. Mirghafori, B. Peskin, A. Hatch, S. Stafford

Work funded by KDD and NSF

Outline

- Overview of submissions
- Commonalities
  - Dataset
  - ASR
  - Combination
- Individual Systems
  - Acoustic systems
  - Stylistic systems
- System combination
- Overall analysis
- Summary and Conclusions

Overview of Submitted Systems

Individual Systems

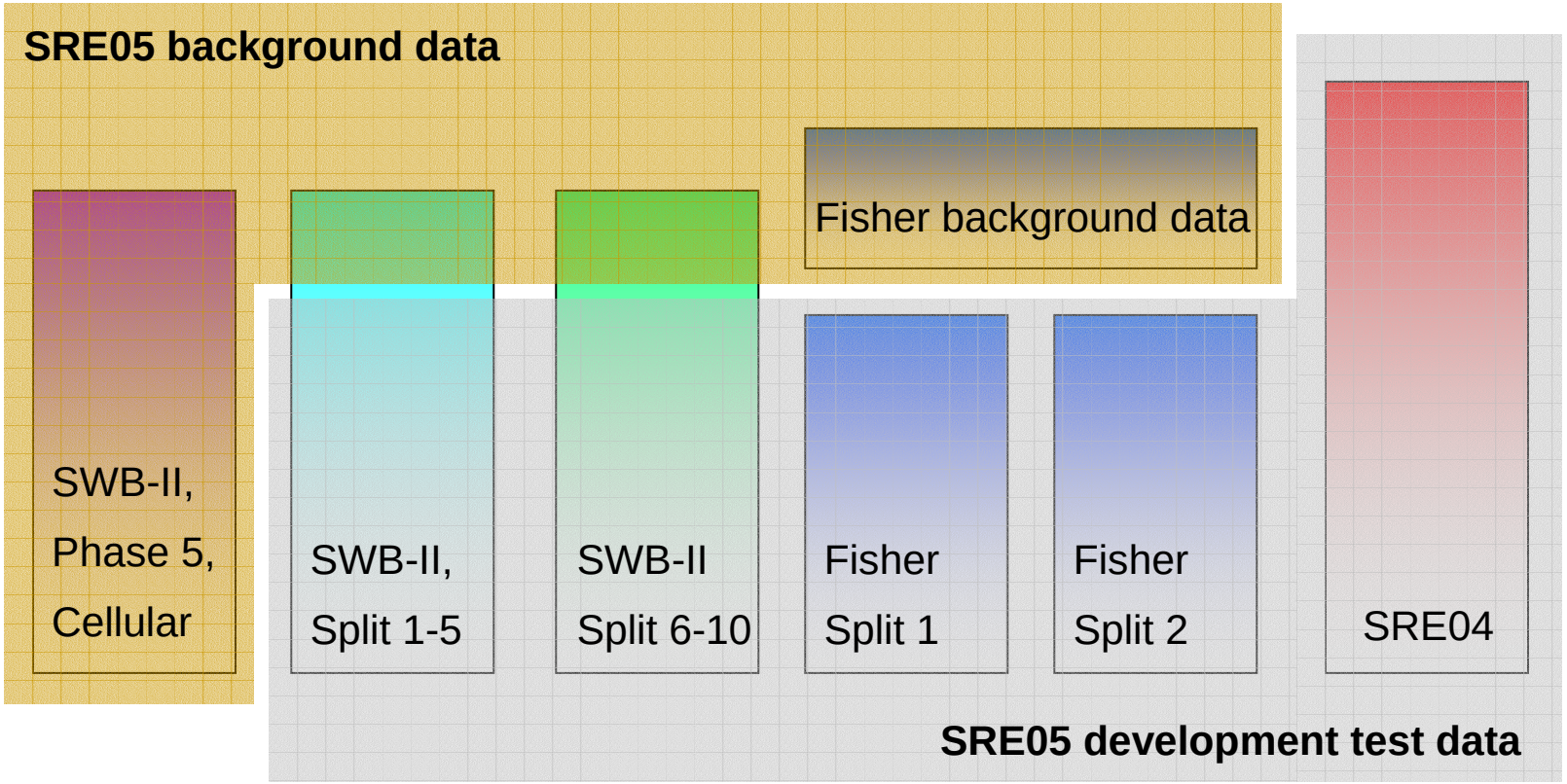
Type	Feature	Statistical Model	Trials Scored
Acoustic	MFCC	GMM	ALL
	MFCC	SVM	ALL
	MLLR Transform	SVM	ALL
Stylistic	State Duration	GMM	English-only
	Word Duration	GMM	English-only
	Word N-gram	SVM	English-only
	WNERFs + SNERF	SVM	English-only

Submission used all systems with different combiners

Submission	Systems	Combiner
SRI_1 (primary)	SRI (7)	Neural network + Class Dependent + SVM
SRI_2	SRI (7)	Neural network
SRI_3	SRI (7) + ICSI (3)	Neural network

All submissions include results for 1conv4w-1conv4w and 8conv4w-1conv4w

Development Datasets



- Part of SWB-II data was ignored because it had overlap with ASR training data
- TNORM for SRE05 was used from Fisher split 1
- Combiner for SRE05 was retrained on SRE04 (after development)

Automatic Speech Recognizer

- Ran a speech/non-speech classification
  - Used a two-state HMM
  - Output used directly in baseline MFCC-GMM and MFCC-SVM systems
- Obtained word-, phone-, state-level transcriptions from SRI's conversational telephone speech (CTS) recognition system
- For all development and evaluation data:
  - Multi-pass transcription with SRI CTS system
  - 3xRT on Intel 3.4 GHz Xeon hyper-threading processor
  - WER = 24.1% on RT-03 eval set (20.7% on Fisher subset)
- ASR system unchanged from SRE-04
  - No Fisher data used in training (unbiased output on Fisher dev data!)
  - Reprocessed SRE-03 and SWB2-cellular data with this system

Score Combination

- English and non-English trials were combined separately
  - SRE04 data was used for training the combiner and generating dev scores
  - Combiners were run on SRE05 data and the prediction was z-normed using dev scores
  - These scores were thresholded using min-DCF threshold for the dev scores for the same condition
  - Normalized and thresholded scores for English and non-English trials were concatenated to generate complete submission

For each individual subsystem (model), we will report results on:

- Common condition 1conv4w-1conv4w trials: **1convComC**
- Common condition 8conv4w-1conv4w trials: **8convComC**
- Neural Net combination with other systems

Effect of different combiners will be discussed at the end!

Outline

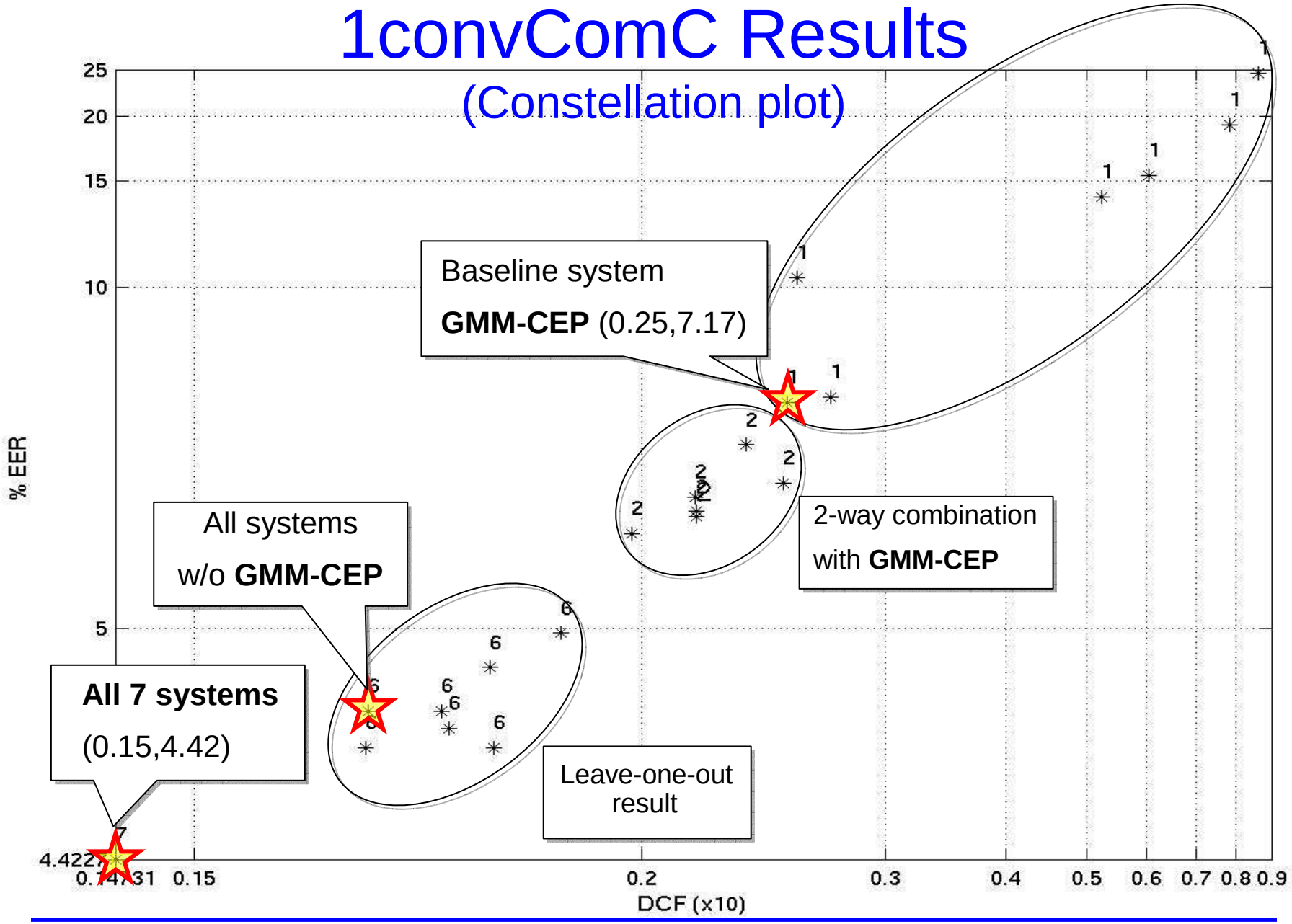
- Overview of submissions
- Commonalities
  - Dataset
  - ASR
  - Combination
- Individual Systems
  - Acoustic systems
  - Stylistic systems
- System combination
- Overall analysis
- Summary and Conclusions

MFCC-GMM (Baseline) System

- Features
  - 13 Mel frequency cepstral coefficients (C1-C13) after cepstral mean subtraction
  - Appended with delta, double-delta, and **triple-delta** coefficients
  - Feature normalization (Reynolds, 2003)
- Features modeled using conventional universal background model (UBM) and Gaussian mixture model (GMM) framework
- 2048-component gender and handset independent speaker independent (SI) model using gender and handset balanced data
  - NIST 1997 SRE data
  - Fisher-dev Bkg-set
  - SWB-II data from NIST 2003 SRE
- Used TNORM for score normalization

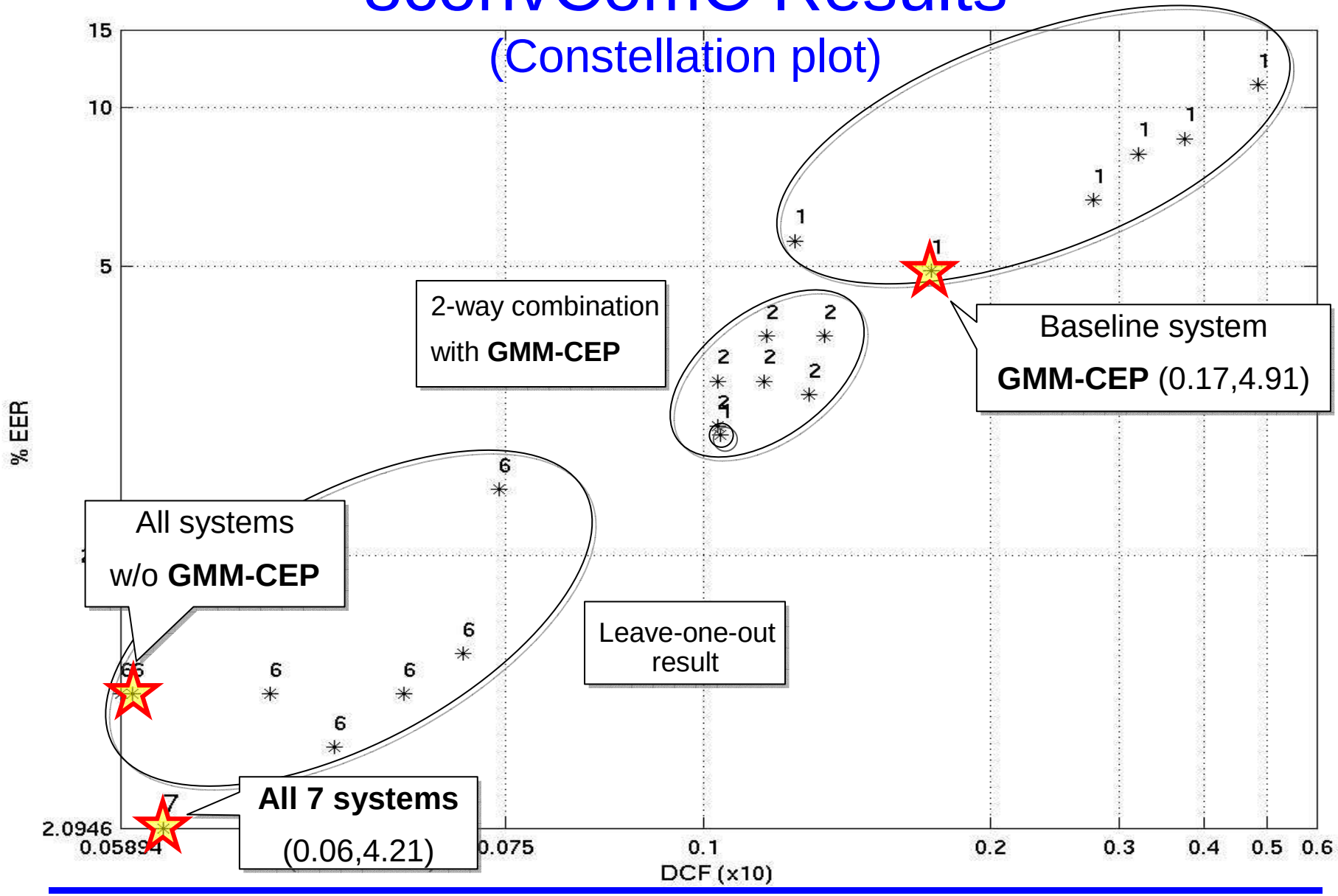
1convComC Results

(Constellation plot)



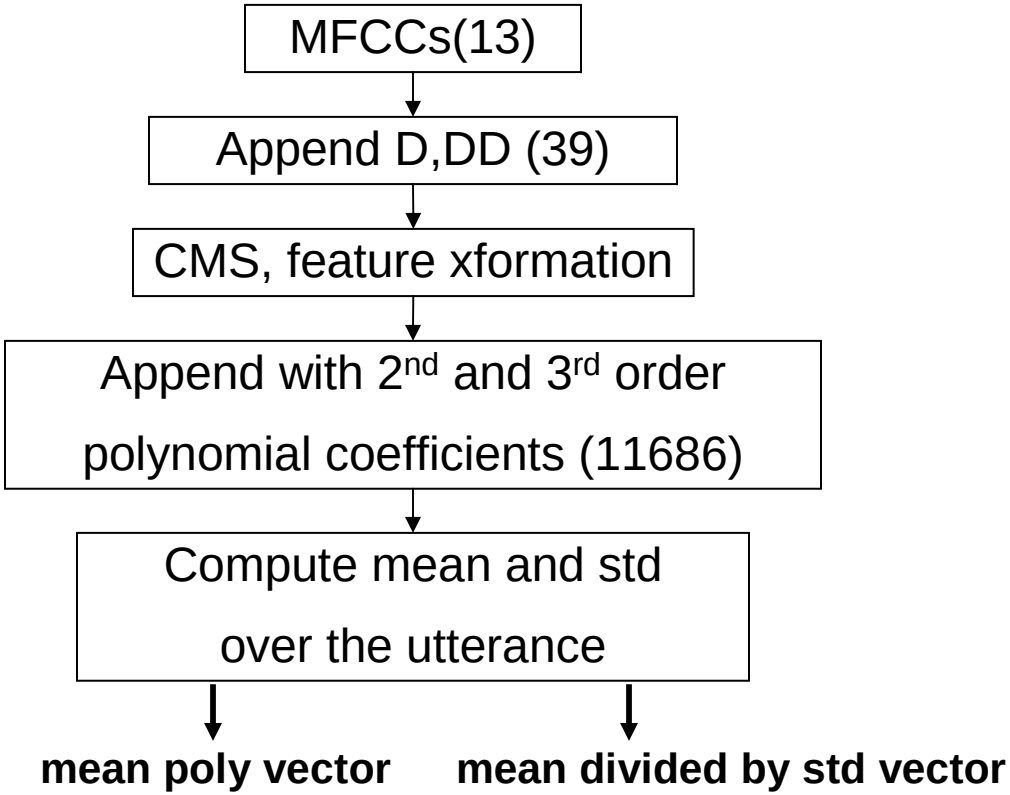
8convComC Results

(Constellation plot)

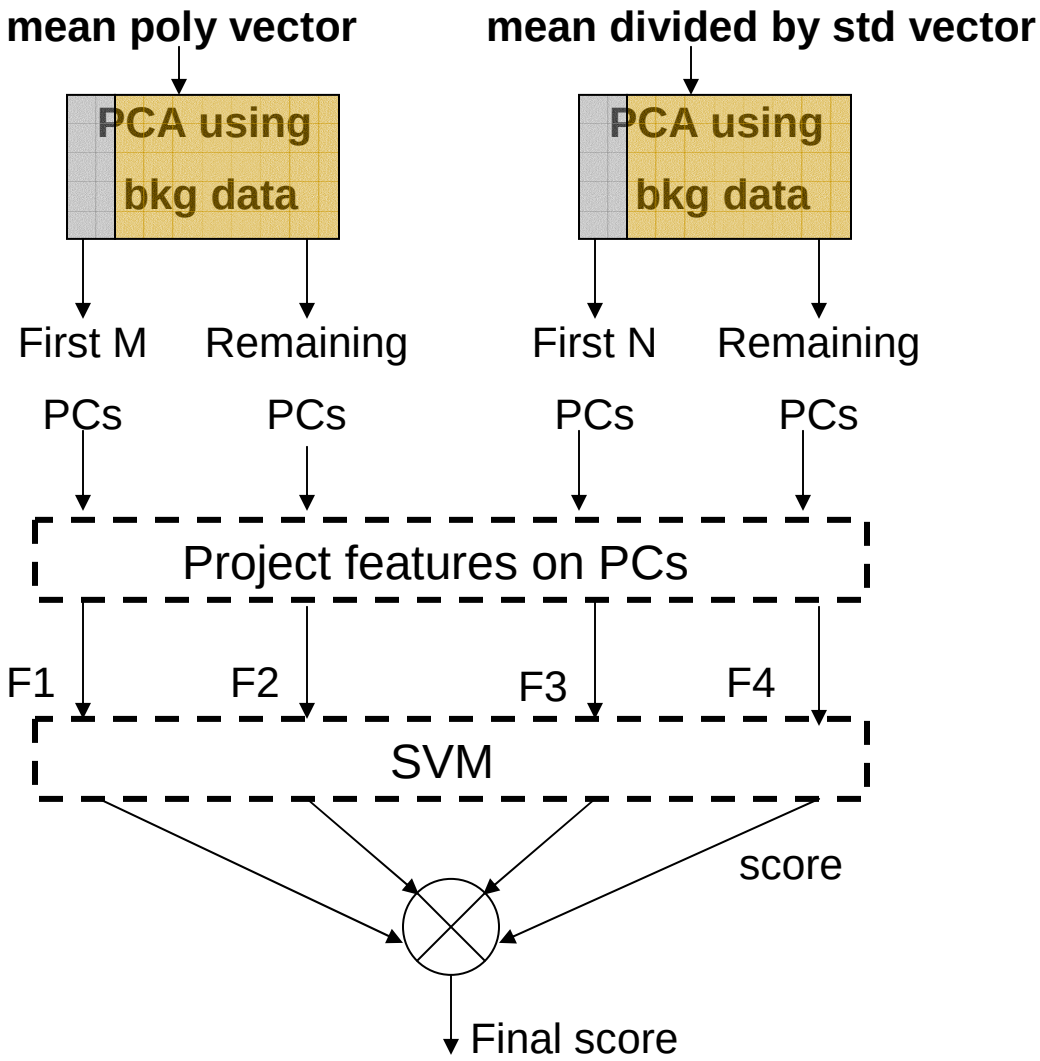


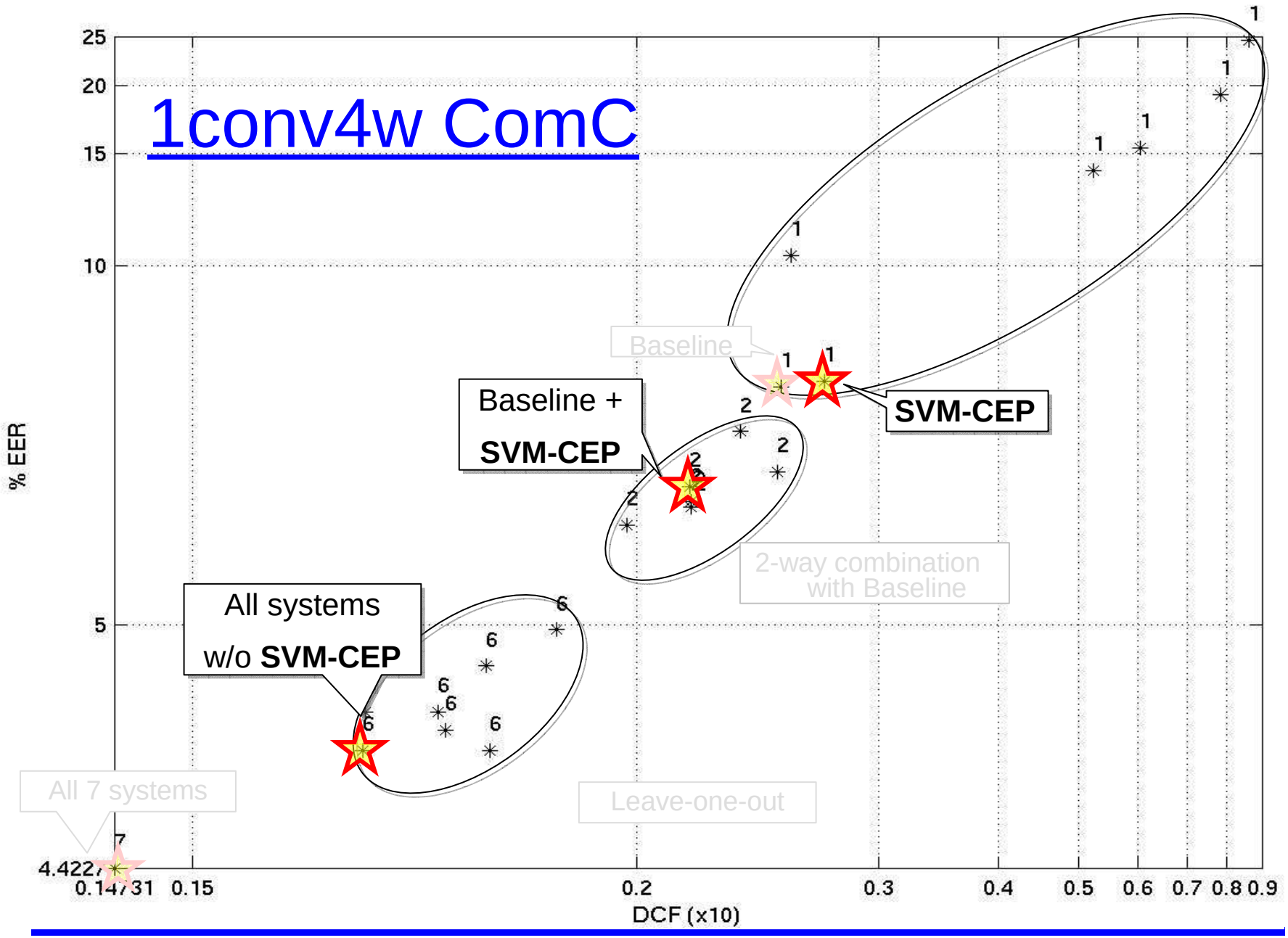
MFCC-SVM System

- The system is a equally-weighted combination of 4 systems
- Its all about features, SVM configuration does not change from system to system
  - Linear kernel, FR error weighted 500 times more costly than FA

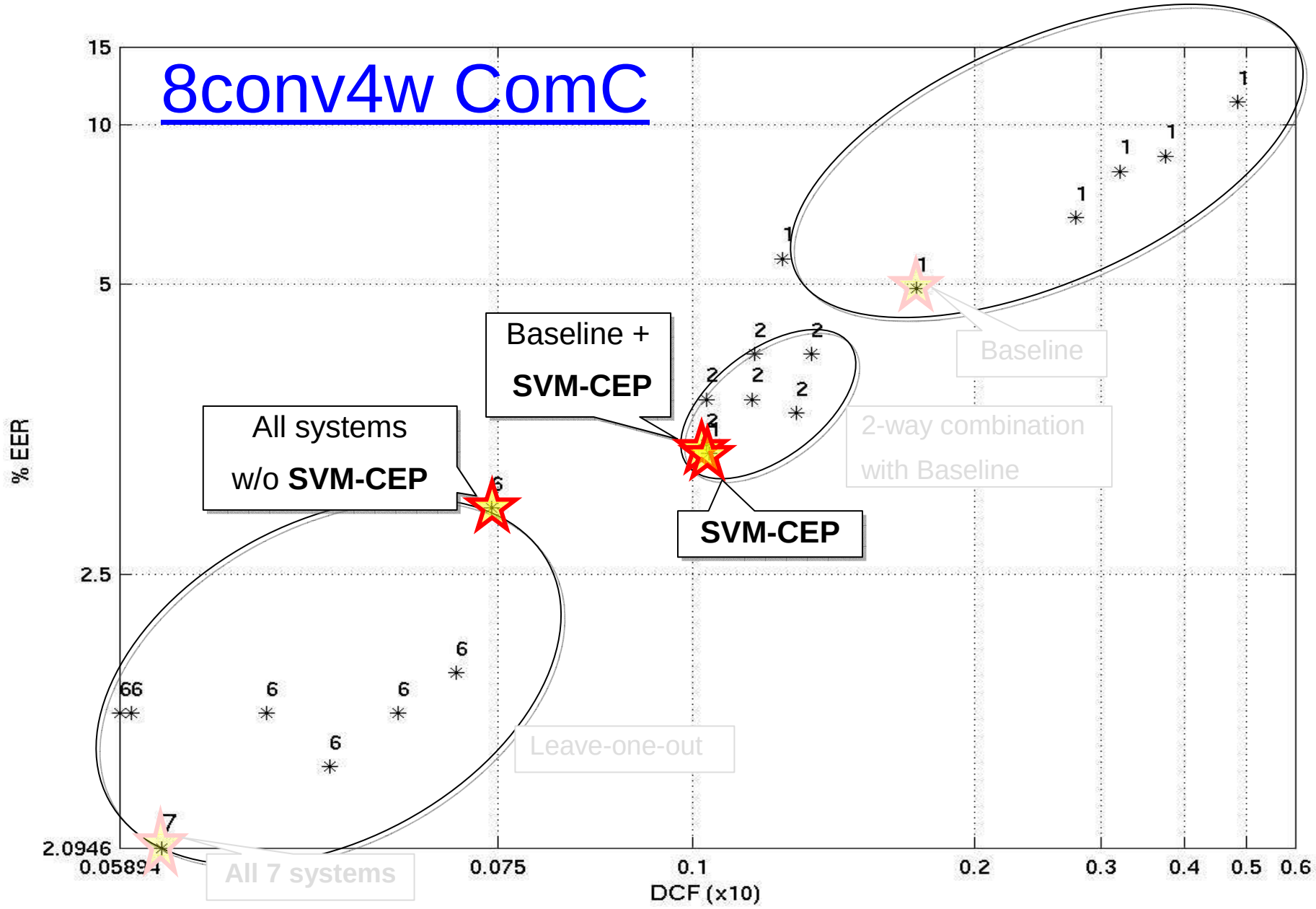


MFCC-SVM System




1conv4w ComC
% EER
DCF (x10)
25
20
15
10
5
0.15
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Baseline
Baseline +
SVM-CEP
SVM-CEP
2-way combination
with Baseline
All systems
w/o SVM-CEP
Leave-one-out
All 7 systems

8conv4w ComC



Rank Normalization for SVM Features

- SVMs are sensitive to relative scaling of feature dimensions
- Absent prior knowledge, ranges should be roughly equal on all dimensions
- Rank-norm: replace each sample by its rank in background speaker distribution
- Zero gets mapped to zero (assuming all values ≥ 0): Sparseness is preserved
- Maps reference distribution to a uniform distribution [0 … 1]
- Distance between two feature values = percentage of population that lies between them

Rank norm example:

Background data: 0 .34 .35 4.3 5.6 100

Data point: 7

Rank-normalized: 0 0.2 0.4 0.6 0.8 1.0

EERs on Fisher devtest			
Feature (old models)	No norm	Z-/Var-norm	Rank-norm
Word N-grams	16.7	20.1	13.4
SNERF-grams	18.1	14.6	14.0
MLLR coeffs	6.4	6.4	6.1

- Rank norm works well across a range of SVM systems
- Works best if background and test sets are well-matched

MLLR SVM System

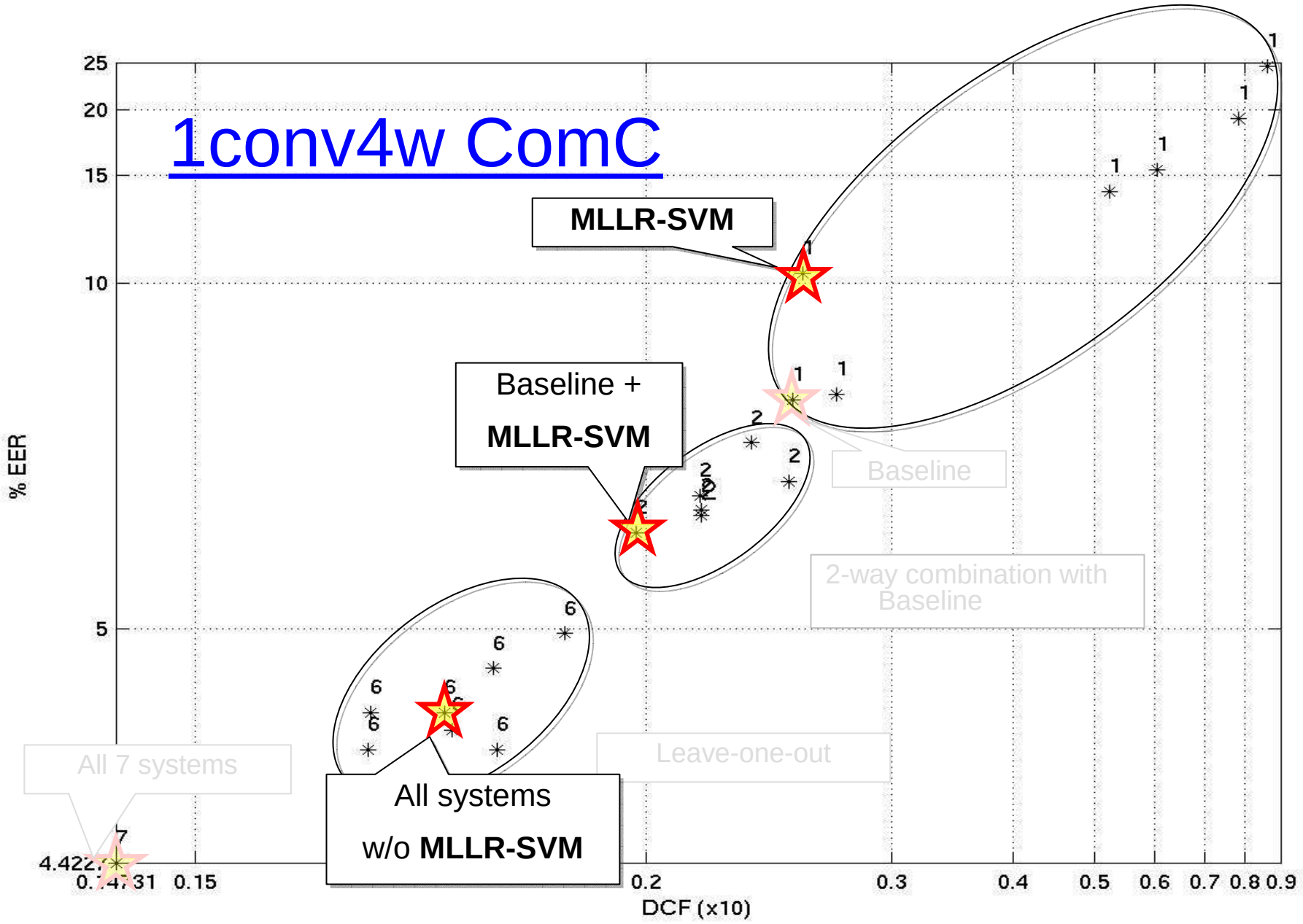
(submitted to Eurospeech '05)

- **Motivation:** normalize out text dependency in cepstral speaker modeling
- **Background:** Speaker adaptation in ASR system
  - Affine mapping of Gaussian means
  - Turns *speaker-independent* into *-dependent* models
  - Estimated with maximum likelihood linear regression
  - Uses phoneloop model or prior recognition output
- **Idea:** use MLLR coefficients as feature vectors and model with SVMs
- Side benefit: ASR front-end and feature transforms normalize out channel effects

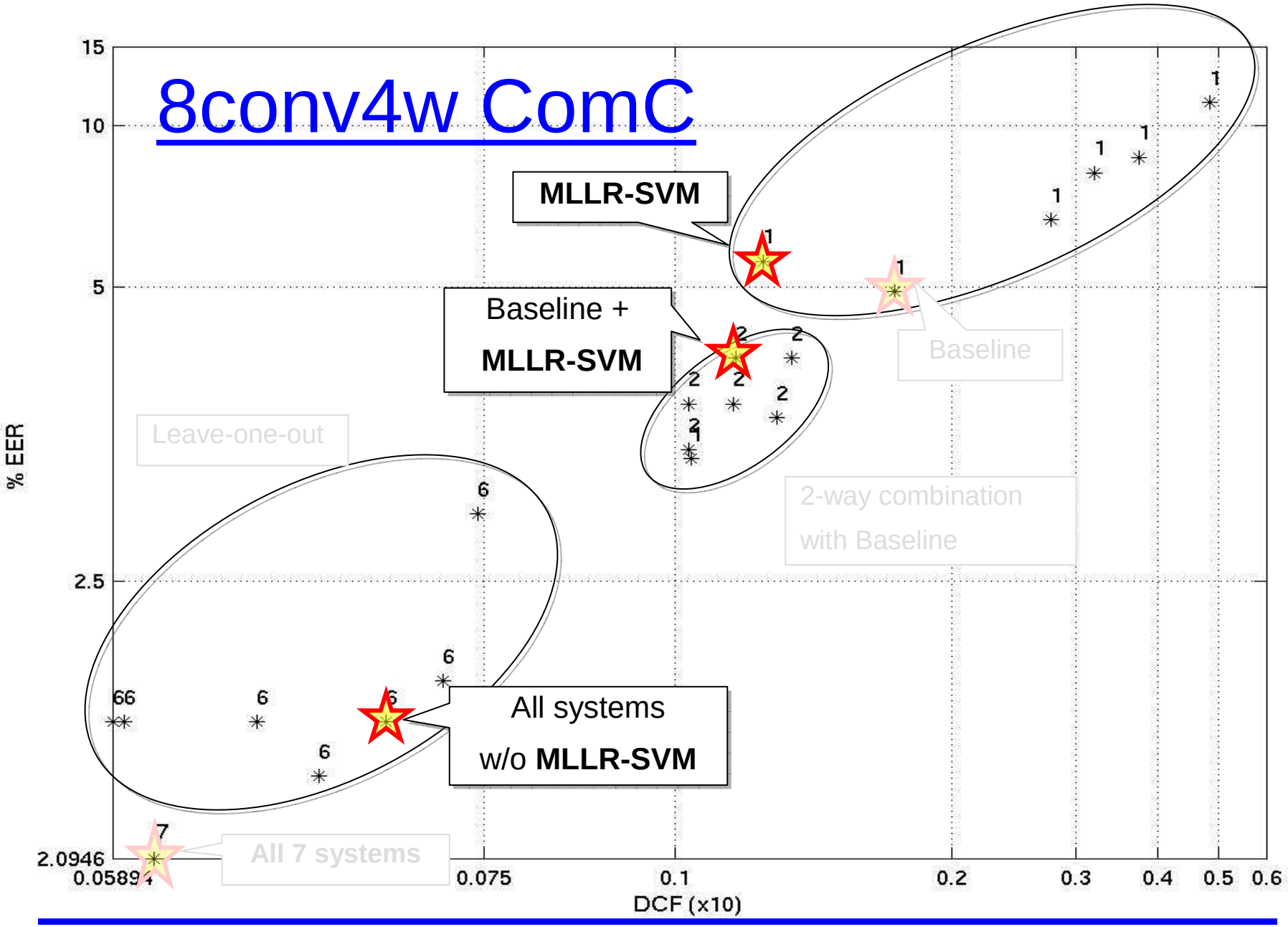
MLLR SVM Implementation & Results

- Combine MLLR transforms from two ASR stages:
  - 1st stage: MFCC, 2 phone classes, adapt to phoneloop model
  - 2nd stage: PLP, 8 phone classes, adapt to 1st recognition hyps
  - Discard nonspeech transforms
- 39 ASR features $\Rightarrow$ 15600 (10x39x40) MLLR features
- Fisher + SWB2 p2+3 background data, linear SVM kernel
- Non-English conversations use only phone-loop adaptation $\Rightarrow$ use 3120 feature components
- MLLR SVM system generally has lower DCF, higher EER than cepstral SVM and GMM systems
- Combines well with cepstral systems
  - 23% EER reduction over cepstral GMM on SRE04 and SRE05

1conv4w ComC



8conv4w ComC



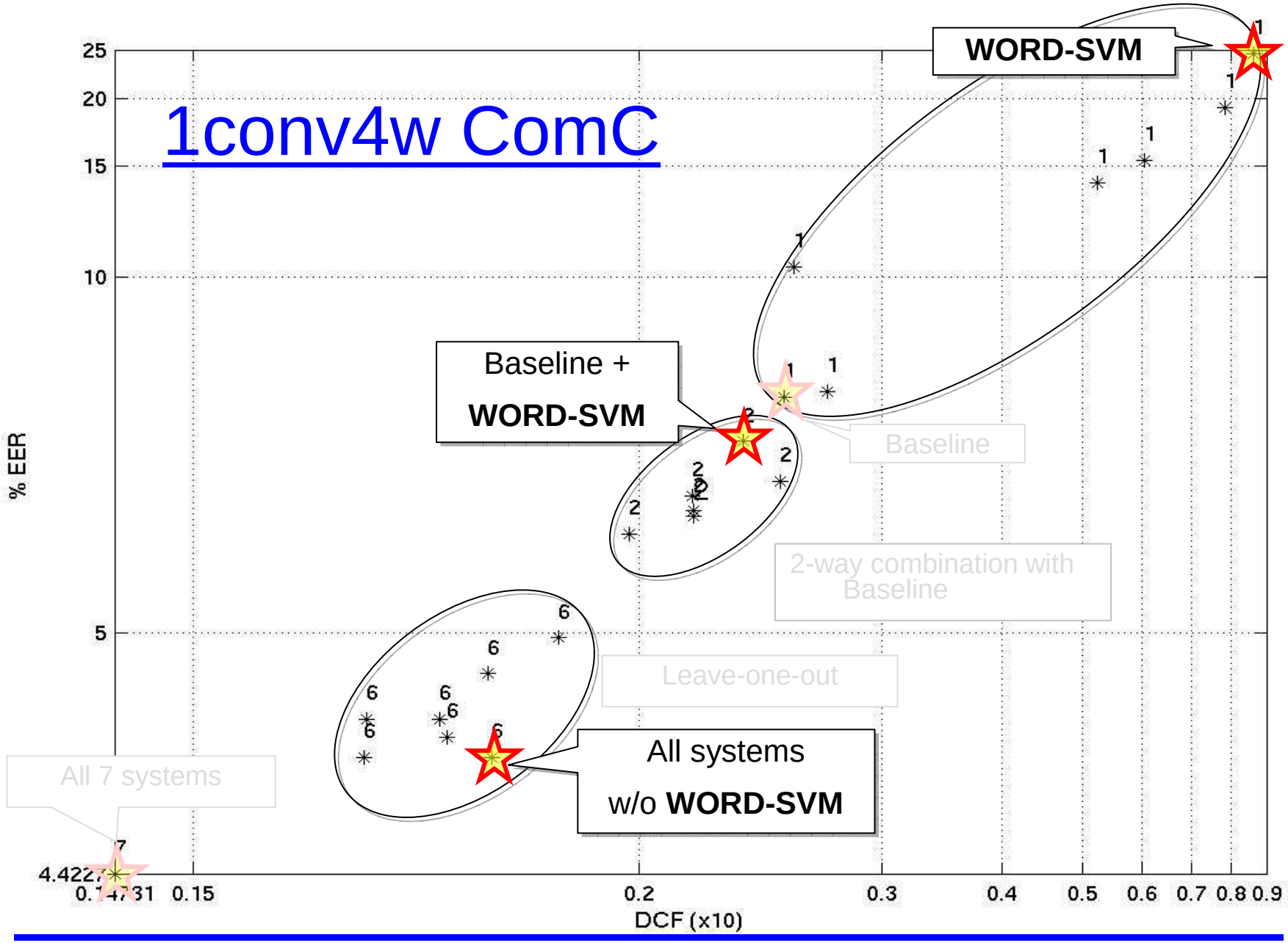
Outline

- Overview of submissions
- Commonalities
  - Dataset
  - ASR
  - Combination
- Individual Systems
  - Acoustic systems
  - Stylistic systems
- System combination
- Overall analysis
- Summary and Conclusions

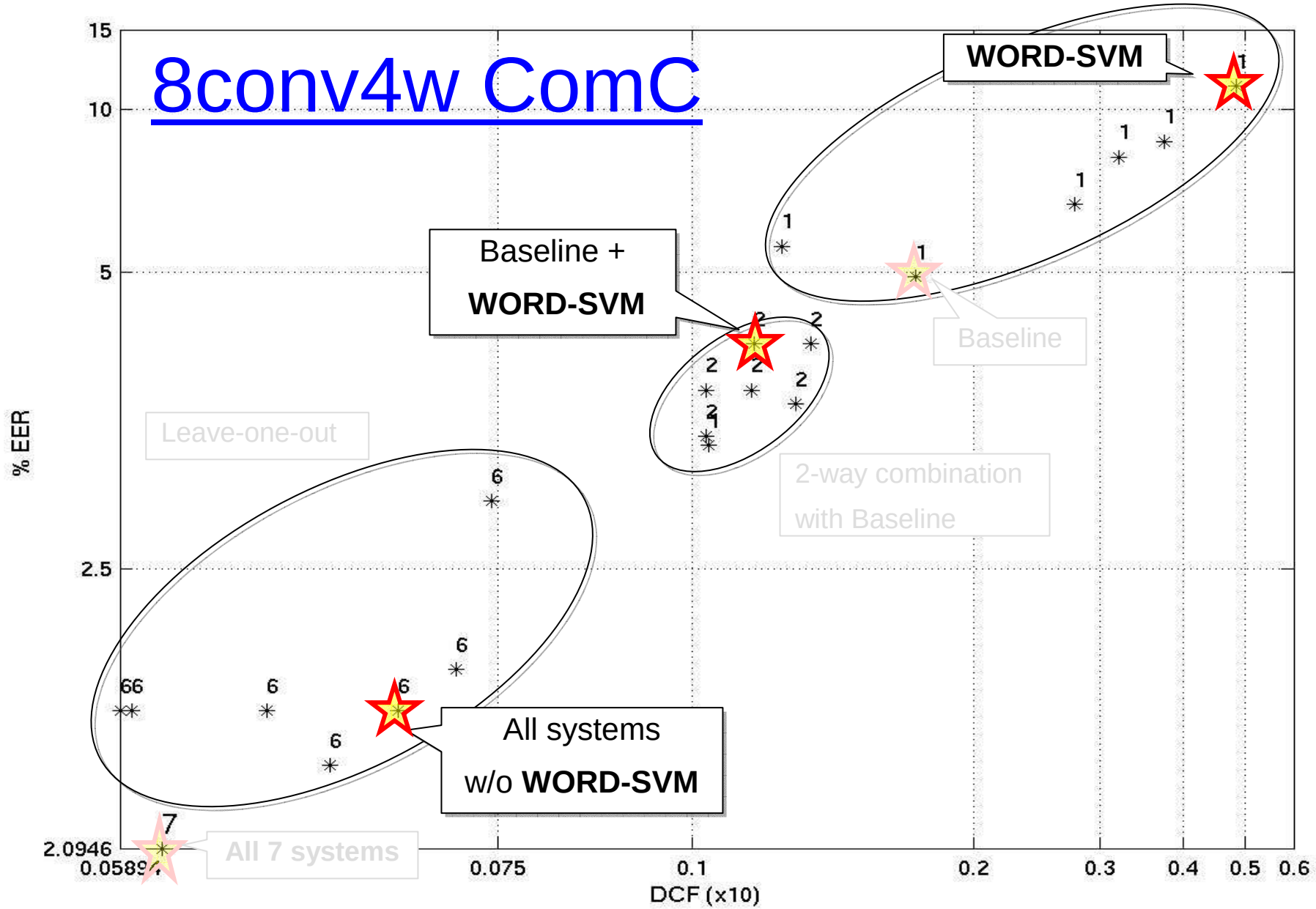
Word N-gram SVM

- Model idiosynchratic word usage patterns
  - Similar to Doddington model, but using SVMs instead of likelihood ratios
  - Based on unigram, bigram, and trigram conversation-level relative frequencies from final ASR output
- Background data: Fisher + SWB2 p2+3+5
  - Removed sides with duplicate speakers to reduce data
  - Unlike for other systems, SWB2 phase5 data helped
- Used all N-grams occurring ≥ 3 times (= 125,579)
- Feature vector = rank-normalized frequencies
- Linear kernel SVMs
- Word N-gram system has poor performance by itself, but combines well with acoustic models

1conv4w ComC



8conv4w ComC

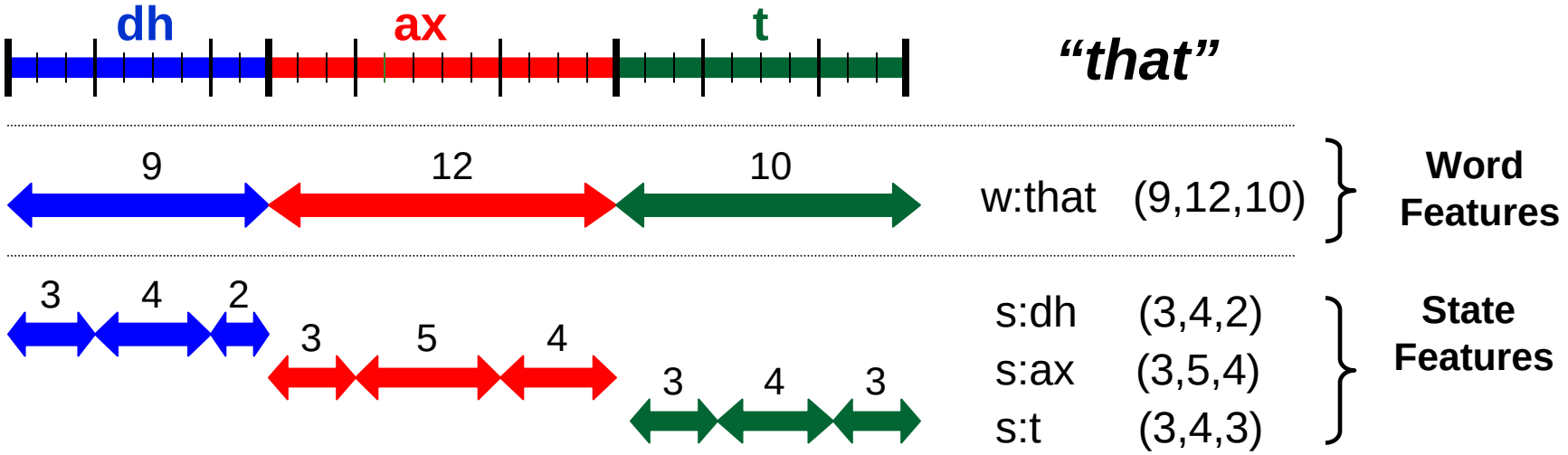


Word N-gram Enhancement Attempts

- Tried to improve word N-gram SVM system in several ways, but no wins so far!
- N-best and lattice-based modeling
  - Inspired by success of lattice-based phone N-grams
  - Compute expected N-gram counts from N-best or lattice output
  - Small gains in matched condition (Fisher for background & test)
  - Degradation in unmatched condition (Fisher $\rightarrow$ SRE04)
- LSA-type dimensionality reduction
  - SVD of background speaker/N-gram frequency matrix
  - Project sparse N-gram freqs to most important eigen-dimensions
  - Similar to Khan & Bayya (ICSLP'04), but using SVM models
  - No gains over baseline: as number of used dimensions increases, performance approaches original system

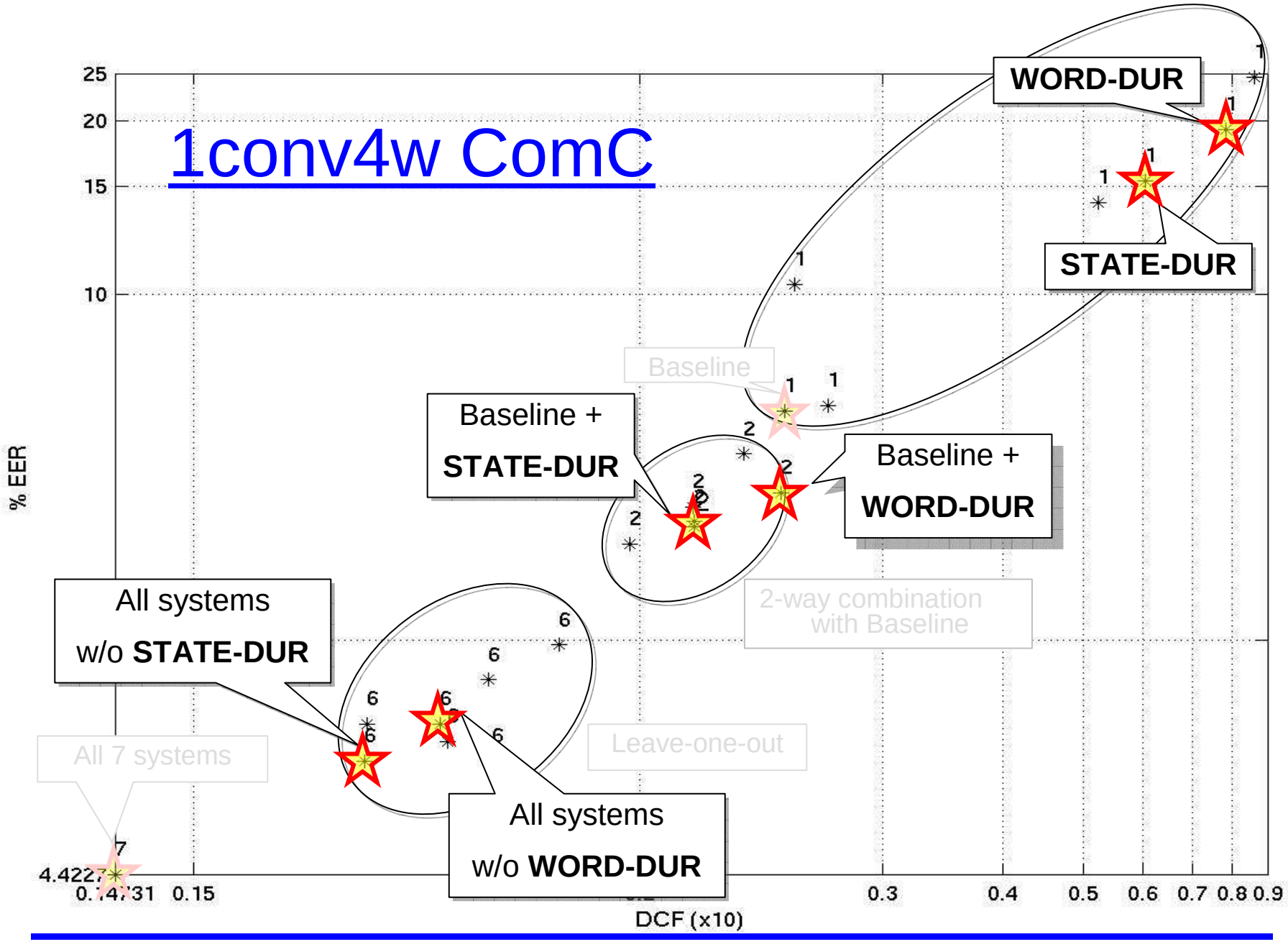
Duration Features
(Eurospeech 2003)

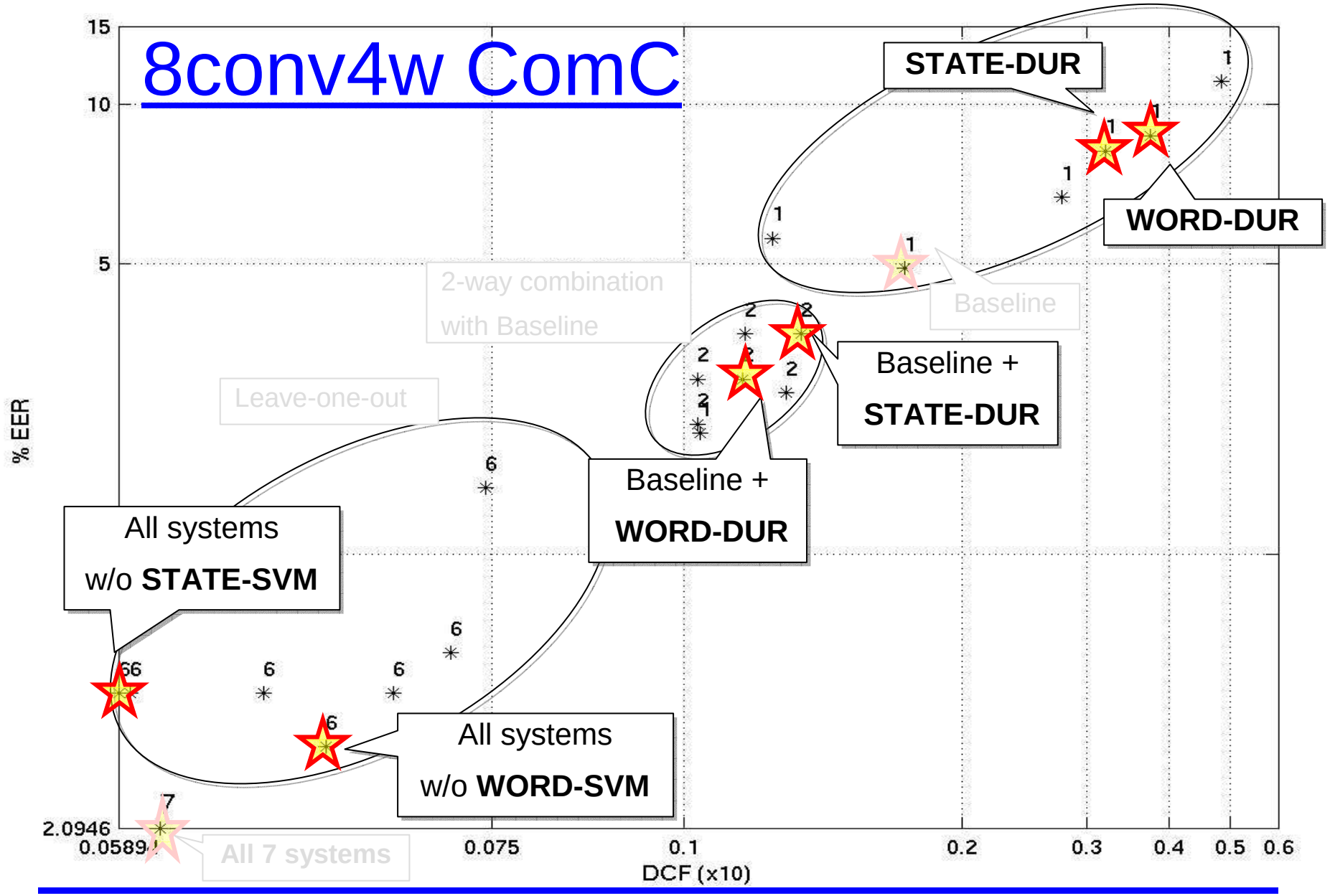
- Each word or phone is represented by a feature vector comprised of the durations of the individual phones or states inside it.



- Durations are obtained from alignments of recognized words
  - first pass for state features
  - final pass for word features
- UBM-GMM paradigm is used to model each word or phone.

1conv4w ComC

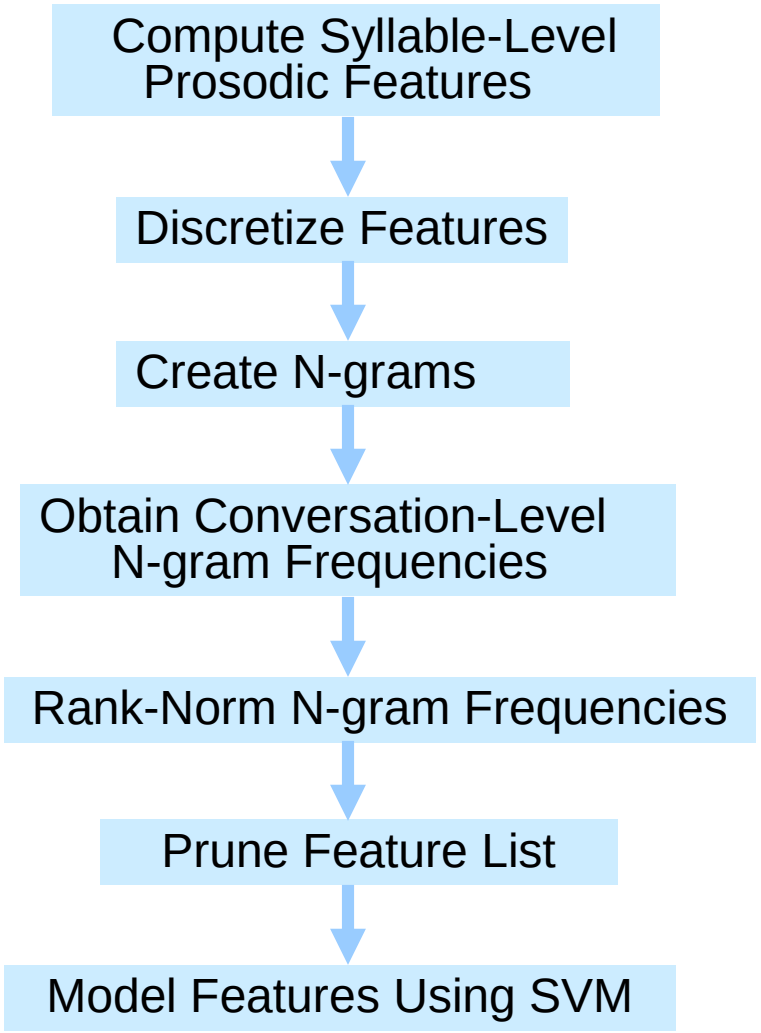



8conv4w ComC
STATE-DUR
WORD-DUR
Baseline
2-way combination with Baseline
Baseline + STATE-DUR
Leave-one-out
Baseline + WORD-DUR
All systems w/o STATE-SVM
All systems w/o WORD-SVM
All 7 systems
% EER
DCF (x10)
15
10
5
2.0946
0.0589
0.075
0.1
0.2
0.3
0.4
0.5
0.6

Nonuniform Extraction Region Features

- SVM feature-level combination of 2 sets of prosodic features.
  - SNERFs (syllable-based nonuniform extraction region features, extracted for all syllables)
  - WNERFs (similar, but extracted only for certain “word lists”)
- SNERF system used last year. Updated version for this year, see ICSLP 2004, and *Speech Communication* 2005 to appear
- No write-up yet on WNERFs

SNERFs



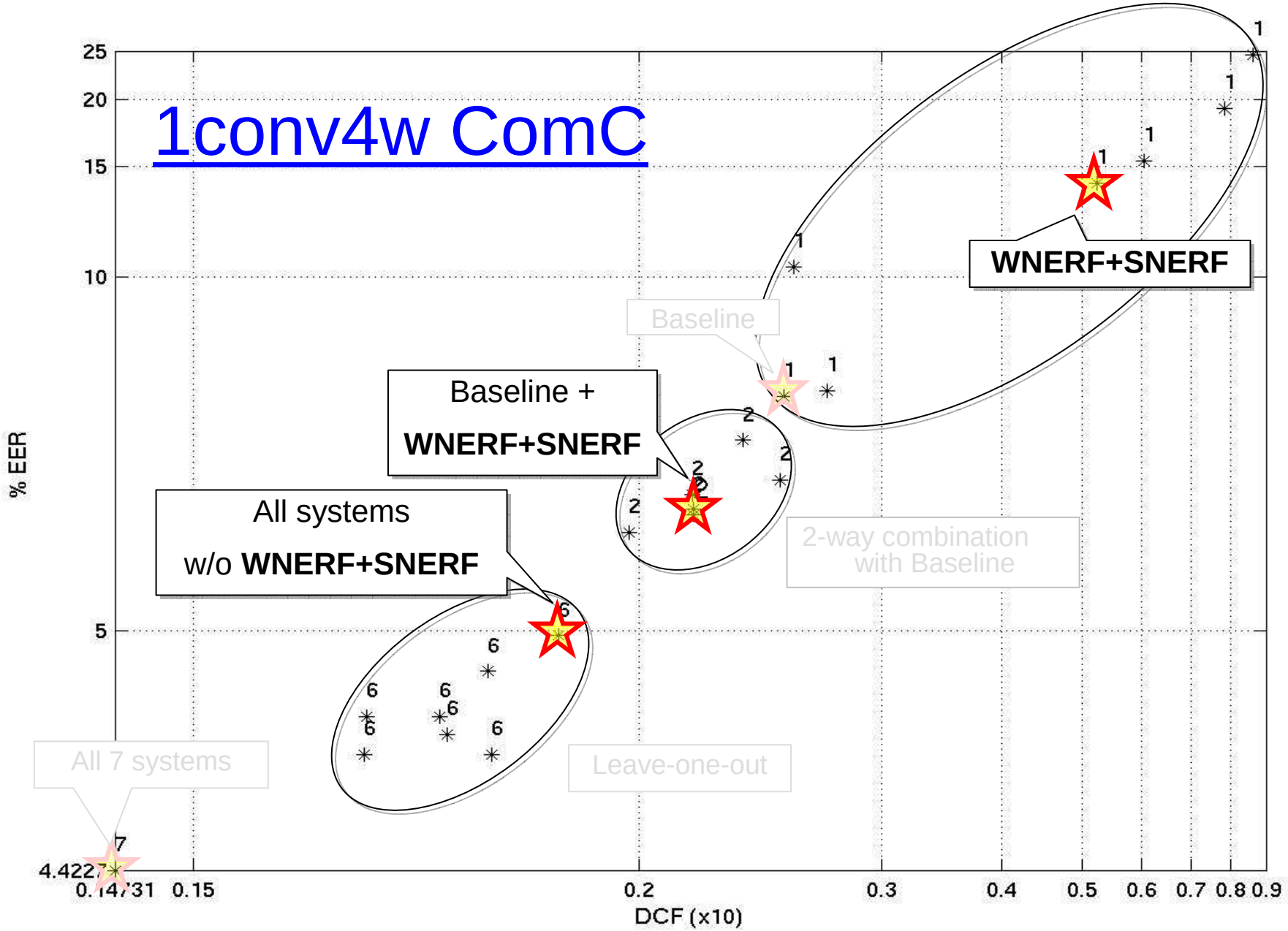
- Features: pitch, energy, duration (pitch most useful)
- Discretization: bin distribution to equally-fill bins; good number of bins across features: 5-10
- N-grams: helpful up to trigrams
- Keep only N-most-frequently occurring N-grams, where N-grams are sequences of binned feature values

WNERFs

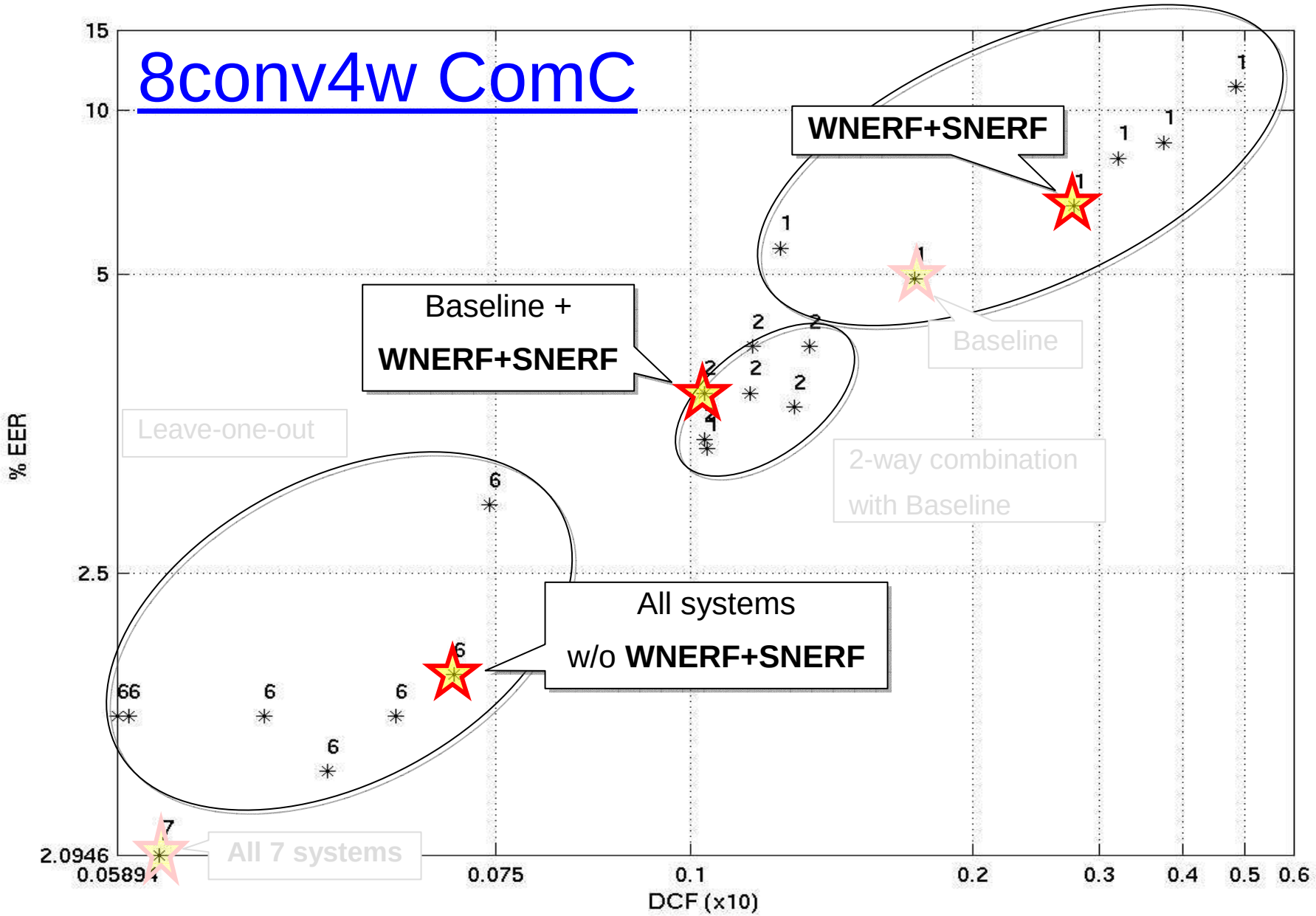
- Same features as SNERFs, but only use data from small set of words
- Models *how the words are said* (given that they are said)
- Created 13 “wordlists” containing total of 24 unique words (words reused), based on linguistic knowledge, “plowing” & “praying”. E.g.:
  - uh um and
  - [pause>100msec] [uh um and but so well]
  - uhhuh sure right okay huh cool good no wow yeah . . .
- Best to group words similar in discourse function or phrase location
- Despite tiny word set & no new features, WNERFs help SNERFs:

DCF, 1 side training	Fisher	SRE04	SRE05
SNERFS only	0.342	0.669	0.595
SNERFS+WNERFS	0.290	0.569	0.522

1conv4w ComC



8conv4w ComC



"Algemy" System Using Metadata

Thanks to Harry Bratt (Algemy) and Yang Liu (metadata)

- Also started work on features in regions longer than 1 syllable, using new software called "Algemy". Similar idea to NERFS, but here, we:
  - Used not only pauses but also estimated sentence boundaries (EARS MDE system)
  - Explored range of region lengths
  - Used N-grams (of values in consecutive regions)
  - Modeling using an SVM (previous NERFs used GMM)
- Algemy system performed in similar range as word duration system, and better than word N-grams, on dev data.
- However: did not generalize that well to SRE04, and we did not have ideal method for combining with other SVM features
- Nevertheless, learned that:
  1. Automatic sentence boundaries are better than pauses for region cuts
  2. Shorter regions are better than longer ones
  3. To find shorter regions for metadata, lower threshold probability of event
  4. N-grams help considerably

Outline

- Overview of submissions
- Commonalities
  - Dataset
  - ASR
  - Combination
- Individual Systems
  - Acoustic systems
  - Stylistic systems
- System combination
- Overall analysis
- Summary and Conclusions

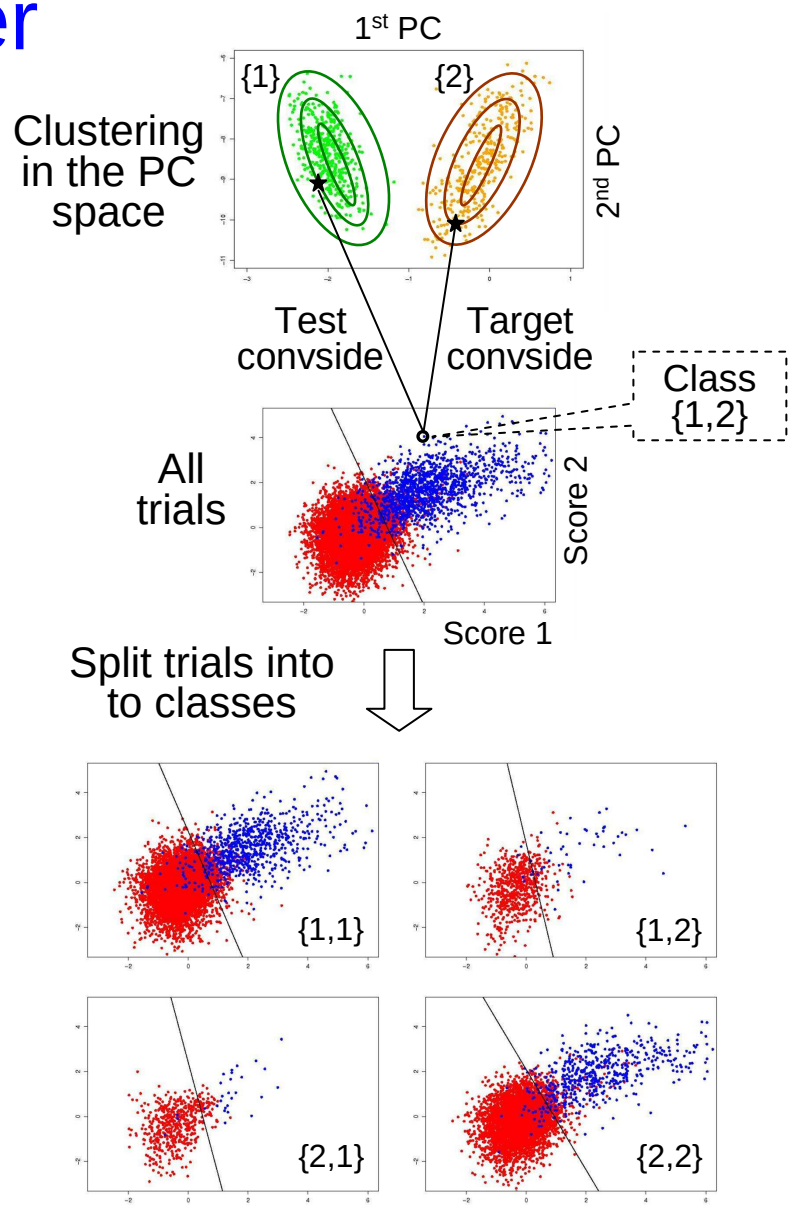
Class-dependent Combiner

(submitted to Eurospeech 2005, inspired by Solewicz et al.)

- Motivation:
  - Combiner parameters may depend on SNR, channel type, speaker characteristics, etc.
  - Auxiliary features can be used to determine these classes.
  - A different combiner can be trained for each class.
- Currently using:
  - MLLR features (2 transforms from 1st ASR stage) as auxiliary features
  - Two clusters
  - Weighted least squares regression to train each class-dependent combiner
- Other auxiliary features were tried (no significant improvements over NN)

Class-dependent Combiner

- Using the first PCs of the MLLR features for the background data, obtain a GMM (gaussian = cluster)
- Each trial belongs to a class given by: {test cluster, target cluster}
- **Training**: Fit a classifier for each class
- **Testing**: Use corresponding classifier to obtain score
- **Enhancement**: Use probabilities given by GMM to make soft decisions instead of hard ones.
  - Use all samples to train all classifiers with weights given by probs
  - Use all classifiers for testing, averaging them by the corresponding probs



Other Combination Models We Tried

For each system *i*
  Let τ(*i*) = Threshold(*i*)
  Z-norm score file (μ=0, σ=1)
  For each trial *t*
    Ask "Is score(*t*,i) < τ(*i*)?"
        "Is |score(*t*,*i*)| > 0.5?"
        "Is |score(*t*,*i*)| > 0.75?"
        etc.
  end
end

Maximum entropy (Doesn't generalize)

The combined score for each trial is
$S(t,i) = \Sigma\, w(i)\, \text{score}(t,i) + c$
$w(i)$ is the weight of system i
$s(t,i)$ is the score of trial t using system i

Weights are initially set to 1 - EER(*i*)
Trained by GD using DCF to reduce effect of errors due to false positives.

Weighted Cost Sum (Good, but not enough)

Each trial is a feature vector
Elements = individual system scores

Do boosting of hundreds of small trees to predict class (0 or 1).

Current performance close to NN performance.

Decision Tree (Need further research)

Each trial is a point in hyperspace
Coords = individual system scores

Train SVM to separate –ves and +ves

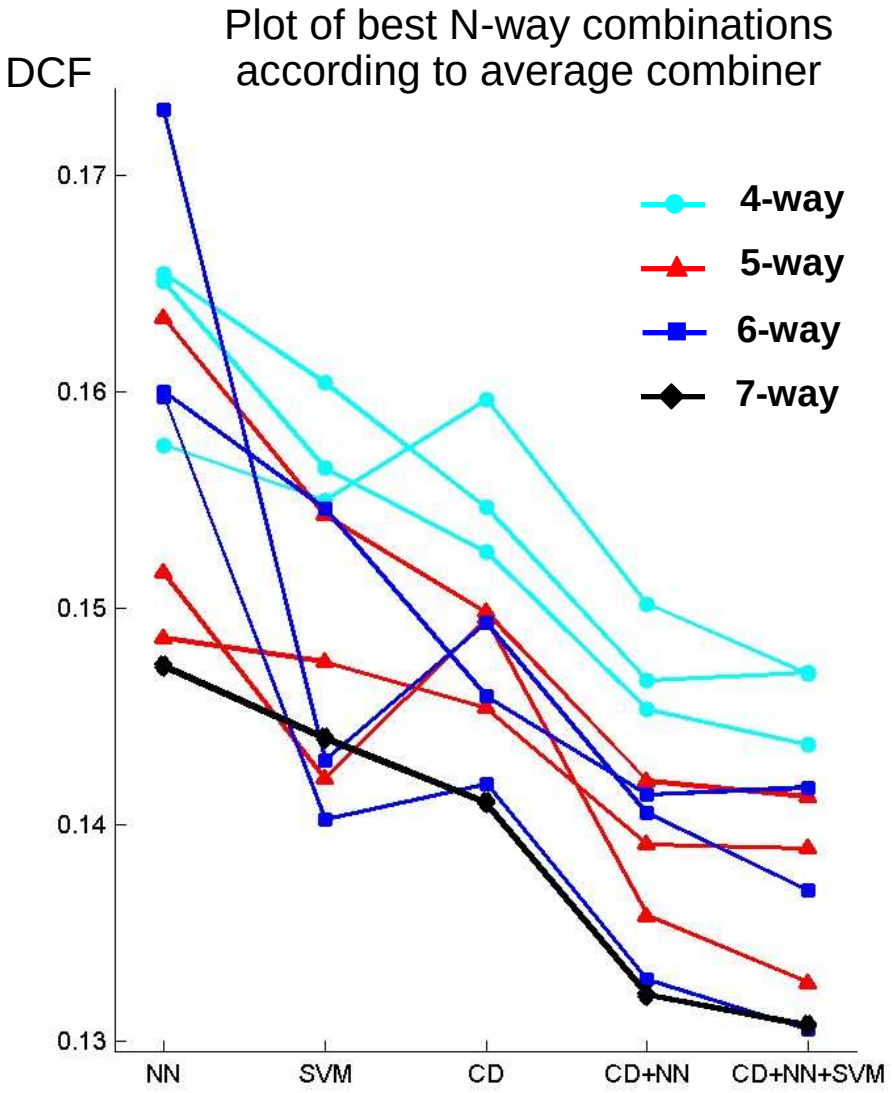
During test, interpret signed distance from hyperplane as the combined score.

SVM (Robust, performs well)

Combiners We Used

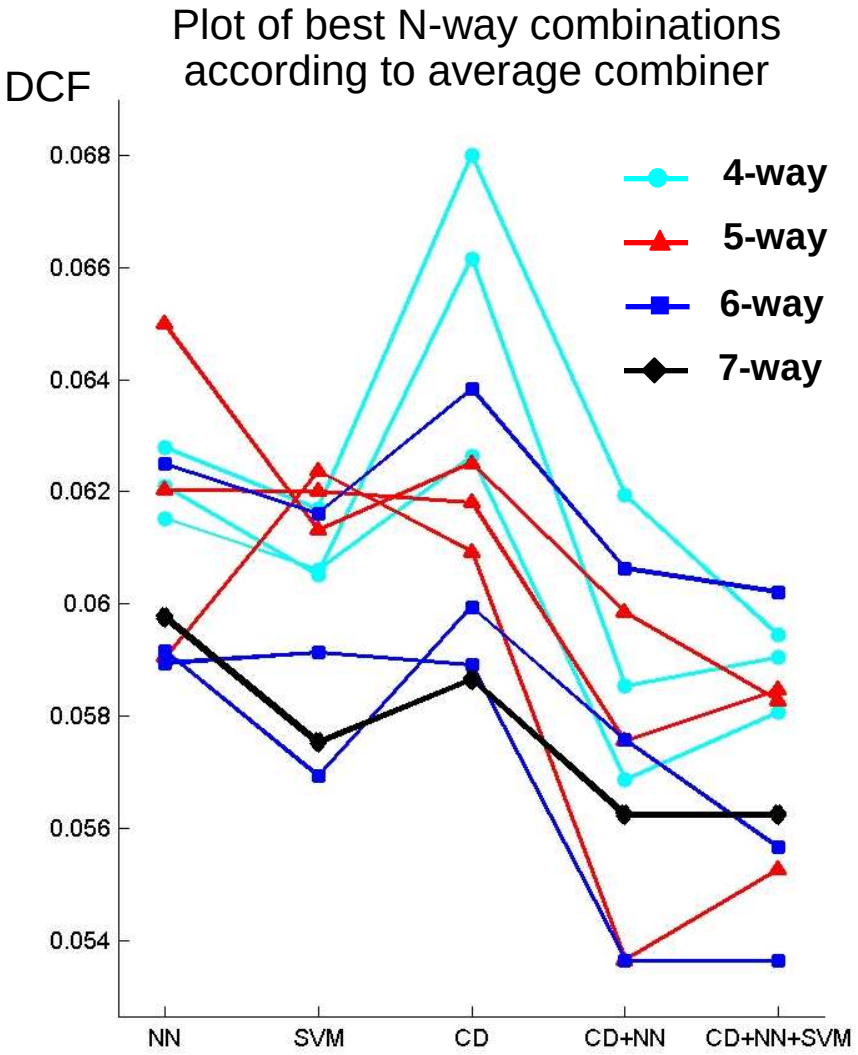
- **Neural network combiner**
  - LNKnet software (MIT-LL)
  - Neural network classifier without hidden layer
  - Classifier trained to optimize decision cost function (DCF) value
  - Target and impostor priors adjusted to 0.09 and 0.91
  - Output node was modified to generate output on the linear scale
- SVM combiner (Garcia Romero et al.)
  - Polynomial kernel, with equal penalty for false acceptances and false rejection
  - Trained the three combiners with orders 1,2, and 3
  - Averaged the outputs from these three combiners
- Class-dependent combiner
- **Combination of combiners**
  - Average of 3 combiners above
  - First normalize scores to zero mean, unit variance (using SRE04 data)
  - Other weights were tried but equal weight proved to be a good choice.

Results for 1conv-side Condition



- Class-dependent combiner is in many cases the best of the individual combiners (this was not true when training and testing on SRE04 !)
- Average of CD+NN gets further improvement over each of them
- Adding SVM tends to help in most cases

Results for 8conv-side condition



- Class-dep combiner is not as good as for 1conv cond. It was not designed for this task. Need more research.
- But in combination with NN is better than any individual combiner.
- DCF of the 7-way combination is worse than best 6- and 5-way L

Outline

- Overview of submissions
- Commonalities
  - Dataset
  - ASR
  - Combination
- Individual Systems
  - Acoustic systems
  - Stylistic systems
- System combination
- Overall analysis
- Summary and Conclusions

Overall Analysis

- Questions:
  - Which *systems* are most important?
  - Does system importance depend on *training data size*?
  - Does system importance depend on *combiner approach*?
  - Did we *improve from last year*?
  - Can we improve results further by adding *new systems*?
- Abbreviations:

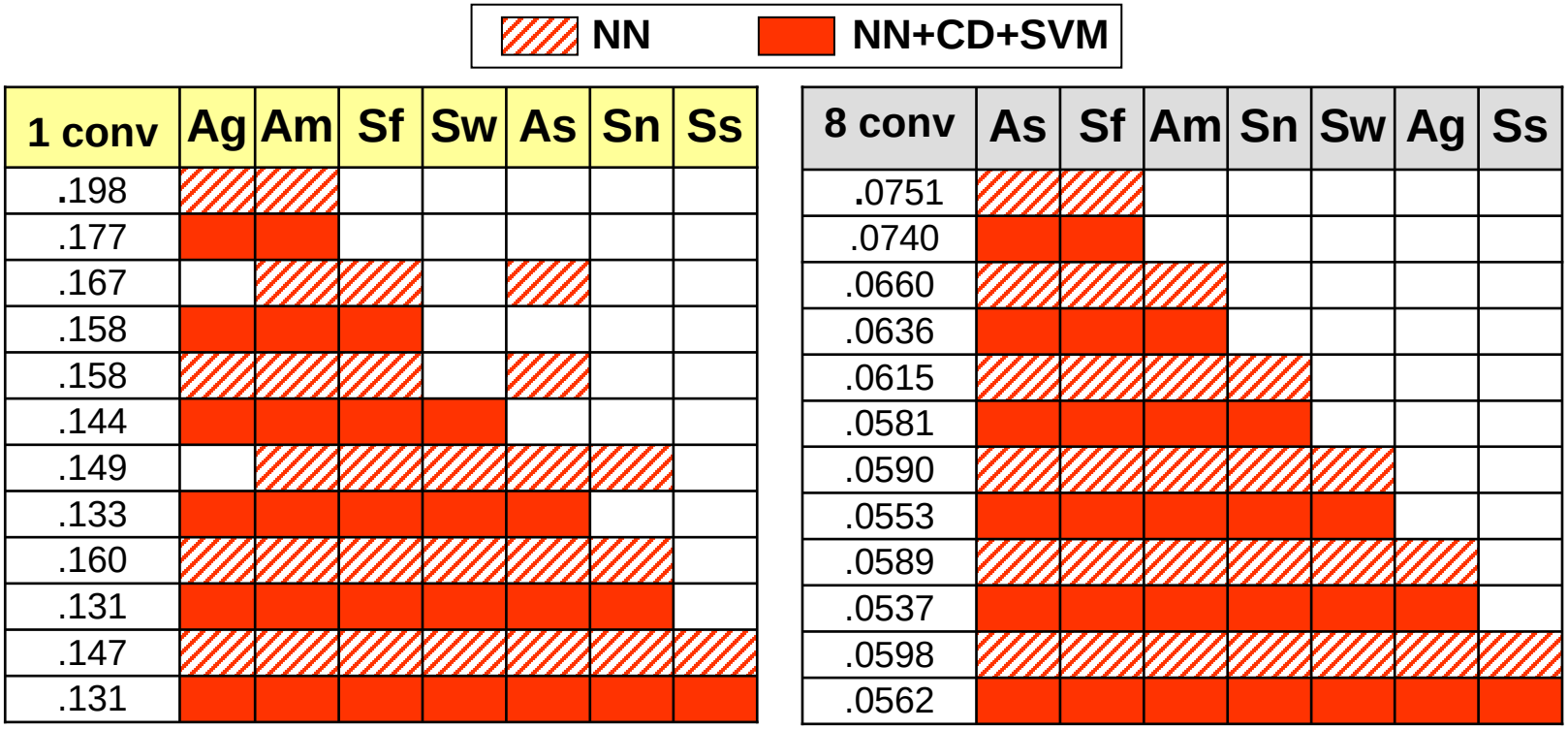
Acoustic Systems (A)	Stylistic Systems (S)
Ag cepstral GMM	**Sn** word N-grams
As cepstral SVM	**Sw** word duration
Am MLLR SVM	**Ss** state duration
	Sf SNERFs+WNERFs

For Reference Only: System Characteristics

	System	Unit	Features	Text Dependent (on Unit)	All vs. Select Regions	Models Bag vs. Sequence	Model
"acoustic"	**Ag**: cepstral gmm	frame	cepstrum	none	all	bag	GMM
	As: cepstral svm	frame	cepstrum	none	all	bag	SVM
	Am: MLLR	frame	cep-transforms	triphone	all	bag	SVM
"stylistic"	**Sn**: word ngrams	word	rel. frequencies	word	select	sequence	SVM
	Sw: word duration	phone	duration	word	backoff	sequence (within word)	GMM
	Ss: state duration	state	duration	phone	all	sequence (within word)	GMM
	Sf: SNERFs	syllable	rel. frequencies prosody values	syllable, phone	all	sequence	SVM
	WNERFs	syllable	rel. frequencies prosody values	word, syllable, phone	select	depends on wordlist	SVM

Analysis: Systems x Combiners x TrainSize

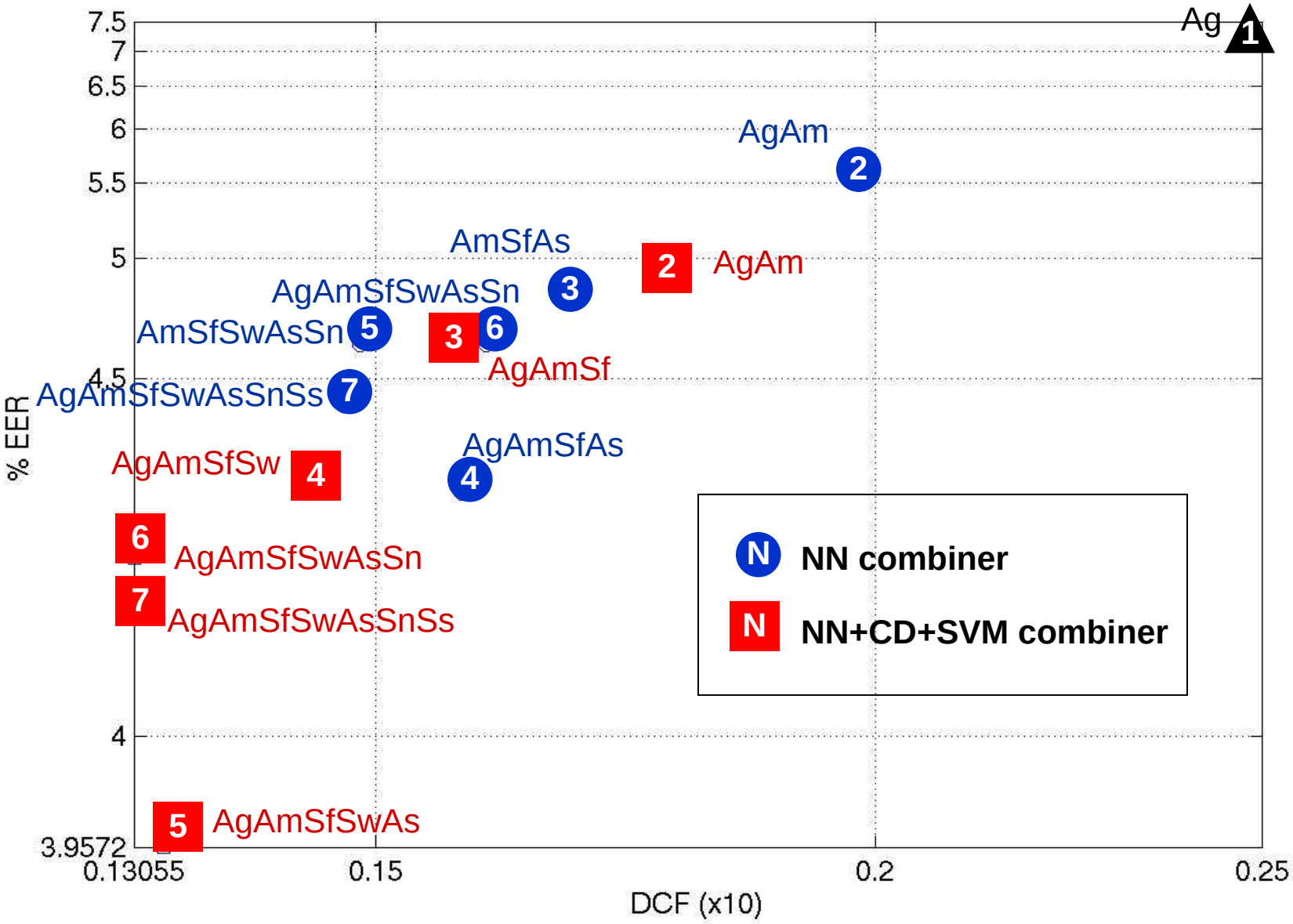
- 1best DCF results for *N* systems, for neural net vs. NN+CD+SVM combiner



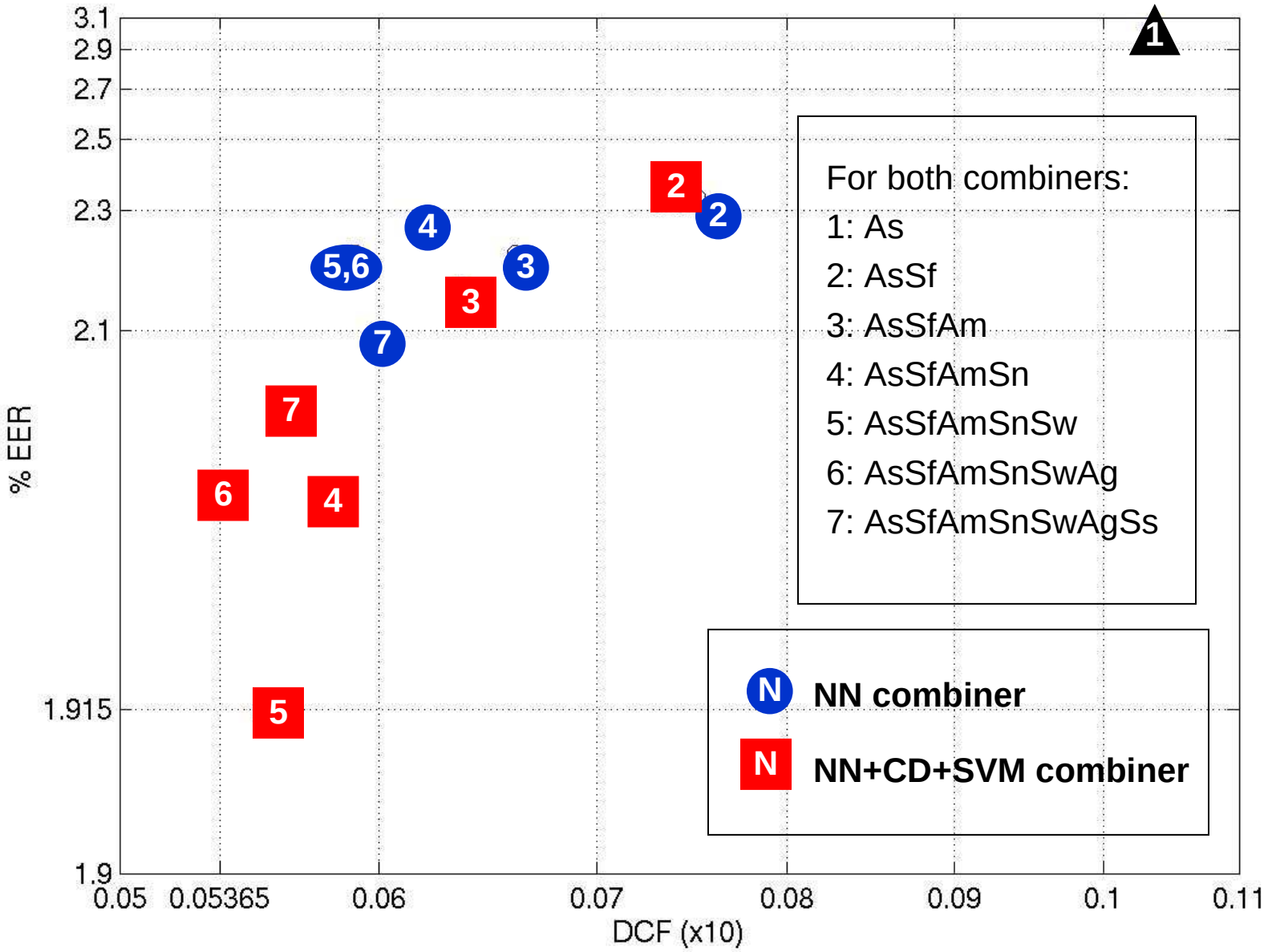
- Fancy combiner always beats NN; also always shows cumulative pattern
- 8 conv training more stable: both combiners have same pattern of systems
- System order changes from 1 to 8 (unlikely to be noise given cumulative pattern): **Ag in 1 > As in 8**, **more stylistic in 8**; hurt by Ss

For Reference: 1-Best, *N* Systems, 1 side

System listing order based on cumulative results for 3-way combiner



For Reference: 1-Best, *N* Systems, 8 sides



Contribution from Acoustic vs. Stylistic Systems

- Compared adding new acoustic vs. stylistic systems to a baseline GMM. (Omitted Ss from 8s since known to hurt.)
- Used 3way combiner (general results similar for NN)

Systems Included	1 side DCF	1 side EER	8 sides DCF	8 sides EER
Baseline Only (cepstral GMM)	.24781	7.1695	.16886	4.9072
Baseline + newAcoustic	.16612	4.6089	.08043	2.4536
Baseline + Stylistic	.17664	4.8883	.07721	2.4536
Baseline + newAcoustic + Stylistic	.13073	4.0968	.05365	1.9749

- Interestingly, adding acoustic gives about same performance as adding stylistic (acoustic a bit better for 1s, stylistic for 8s)
- Yet combining them gives a large win, in both conditions

Adding ICSI Systems

- Used 3-way combiner to add ICSI systems to 6 SRI systems (no state dur)

Systems Included	1 side DCF	1 side EER	8 sides DCF	8 sides EER
SRI systems only	.13055	4.1434	.05365	1.9749
+ wordHMM + phoneNgram + SNP	.13228	4.1899	.05665	2.1871
+ wordHMM + phoneNgram	.13070	4.1434	.05340	2.0347
+ wordHMM + SNP	.12949	4.2831	.05657	2.1264
+ wordHMM	.12841	4.3296	.05349	2.0347

- Hard to improve over $N \approx 6$ systems (also for SRI-only systems), especially for 8 sides, so promising that adding selected ICSI systems can give gains:
  - Adding just the wordHMM helped DCF for 1 side
  - Adding wordHMM+phone Ngrams shows slight gain for 8 sides
  - Such differences in results by training size not inconsistent with differences seen for SRI
- Also, some subselection analyses show some ICSI systems chosen before some SRI systems. E.g., 8 sides, N=6, wordHMM replaces cepstral GMM.

Overall Improvement Since Last Year:
Results on SRE04 Data

- Due to amount of work and limited resources, could not run mothball system
- However, we can run this year's system on SRE04 data
  - Not strictly comparable, since this year had a better-matched tuning set than we had in last year's eval.
  - Nevertheless, not "cheating" since we didn't use SRE04 for background data or for TNORM.
  - Used English-only trials
  - Combiner trained on Fisher (for this experiment)

	1 side DCF	1 side EER	8 sides DCF	8 sides EER
Last year NN	0.32604	7.57	0.16001	3.50
This year NN	0.22016	5.27	0.09109	2.91
This year NN+CD+SVM	0.21805	4.84	0.09115	2.63
Total Relative Improvement	**33.1%**	36.1%	**43.0%**	24.8%

Analysis: Take-Home Messages

- The two cepstral systems are (not unexpectedly) largely redundant. The SVM is superior to the GMM cepstral system, but only if there is enough training data (e.g. 8 sides). For only 1 side, GMM is better.
- Performance of a system alone does not predict (often inversely related to) importance in a larger combination.
- Both acoustic and stylistic features are important
- System importance does not seem to depend *inherently* on combiner
- System importance *does* depend on amount of training data
- More training data means:
  - Greater use of SVM cepstral system
  - Greater use of stylistic features, esp. WNERFs+SNERFs and word N-grams
- Systems useful in previous evaluations may not always be useful in current ones (can even hurt) L . Example: state duration.
  Need to reevaluate combined system with each system change.

Outline

- Overview of submissions
- Commonalities
  - Dataset
  - ASR
  - Combination
- Individual Systems
  - Acoustic systems
  - Stylistic systems
- System combination
- Overall analysis
- Summary and Conclusions

Summary and Conclusions

- Submitted results for 1 & 8 side conditions, using 7 systems, **3 of them novel**
  - Acoustic: cepstral GMM, **cepstral SVM, MLLR SVM**
  - Stylistic: word N-gram, word dur, state dur, **WNERFs+SNERFs**
- All but one system (state duration) helped us.
- Use of rank normalization in various SVMs also helped.
- Both acoustic and stylistic features important
- Non-English trials used 3 acoustic systems, merged with full system for Eng.
- Class-dependent combiner, averaged with NN and SVM gives consistent gain.
- Found 1-best N-system results cumulative for both 1 & 8s using new combiner.
- Relative system importance changes from 1s to 8s:
  - Cepstral features: move from GMM to SVM
  - Increased usage of stylistic systems
- Looked at combination of SRI and ICSI systems.
- Year's progress (DCF reduction on SRE04): **33.1% (1side), 43.0% (8 sides)**

References

- W. M. Campbell, “Generalized linear discriminant sequence kernels for speaker recognition,”, *Proc. IEEE ICASSP*, Orlando, FL, vol. 1, pp. 161-164, 2002.
- L.Ferrer, H. Bratt, V. R. R. Gadde, S. Kajarekar, E. Shriberg, K. Sonmez, A. Stolcke, and A. Venkataraman, “Modeling duration patterns for speaker recognition,” *Proc. Eurospeech*, Geneva, pp.2017-2020, 2003.
- L. Ferrer, K. Sönmez, and S. Kajarekar, “Class-dependent Score Combination for Speaker Recognition”, submitted to *Eurospeech*, 2005.
- T. Joachims, “Text categorization with support vector machines: Learning with many relevant features”, *Proc. European Conference on Machine Learning*, 1998.
- S. Kajarekar, L. Ferrer, K. Sonmez, J. Zheng, E. Shriberg, and A. Stolcke, “Modeling NERFs for Speaker Recognition”, Proc. Odyssey 2004: The Speaker and Language Recognition Workshop, pp. 51-56, Toledo, Spain, 2004.
- C. Leggetter and P. Woodland, “Maximum likelihood linear regression for speaker adaptation of HMMs,” *Computer Speech and Language*, vol. 9, pp. 171-186, 1995.
- D. A. Reynolds, "Channel Robust Speaker Verification via Channel Mapping", *Proc. IEEE ICASSP*, vol. 2, pp. 53-56, Hong Kong, 2003.

References

- E. Shriberg, L. Ferrer, A. Venkataraman, and S. Kajarekar, “SVM Modeling of SNERF-Grams for Speaker Recognition”, Proc. ICSLP, Jeju, Korea, 2004.
- E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, Modeling Prosodic Feature Sequences for Speaker Recognition. *Speech Communication*, 2005 (in press).
- A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman, “MLLR Transforms as Features in Speaker Recognition”, submitted to *Eurospeech*, 2005.
- A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauche, C. Richey, E. Shriberg, K. Sonmez, F. Weng, and J. Zheng, The SRI arch 2000 Hub-5 Conversational Speech Transcription System,” *Proc. NIST Speech Transcription Workshop*, College Park, MD, 2000.
- V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995.
- Y. Solewicz, M. Koppel, "Enhanced Fusion Methods for Speaker Verification", SPECOM, Saint-Petersburg, September, 2004.
- D. Garcia-Romero, J. Fierrez-Aguilar, J. Ortega-Garcia and J. Gonzalez-Rodriguez, “Support Vector Machine fusion of idiolectal and acoustic speaker information in Spanish conversational speech”, *ICASSP, Hong Kong, April 2003.*