Lukas Burget, Mirko Hannemann, Hynek Hermansky, Sally Isaacoff, Sanjeev Khudanpur, Haizhou Li, Chin-Hui Lee, Pavel Matejka, Jon Nedel, Ariya Rastrow, Marco Siniscalchi, Petr Schwarz, Puneet Sahani, Rong Tong, Chris White , Geoff Zweig

WHAZWRONG ?

Reveal weaknesses of

Large Vocabulary Continuous Speech Recognition

9:05-9:20

Introduction

Hynek Hermansky

9:20-9:50

Detection of Out-Of-Vocabulary Words and Recognizer Confidence Estimation Using Parallel Strongly and Weakly Constrained Classifiers

Lukas Burget, Petr Schwarz, Mirko Hannemann, Puneet Sahani

9:50-10:10

Estimation of Effectiveness of Features for Classification Based On Mutual Information Measure

Aryia Rostow and Sanjeev Khudanpur

10:10-10:20

Comparison of Strongly and Weakly Constrained Phoneme Posterior Streams Using Trained Artificial Neural Net

Pavel Matejka

10:20- 10:30

BREAK for Schmoozing

Everybody

10:30-11:00

Phoneme Recognizers and Transducers in Detection of Out-Of-Vocabulary Words and Recognizer Confidence Estimation

Sally Isaacoff, John Nedel, Chris White and Geoff Zweig

11:00-11:30

Detection of English Words in Mandarin Utterances and Improved Detection through Universal Phone Models

Rong Tong, Chin-Hui Lee, Haizhou Li, Marco Siniscalchi

11:30:11:35

Summary

Hynek Hermansky

11:35-12:00

General Discussion about How To Make Recognizers to Say “I Do Not Know”

Everybody

12:00

Catching a Plane Back To Switzerland

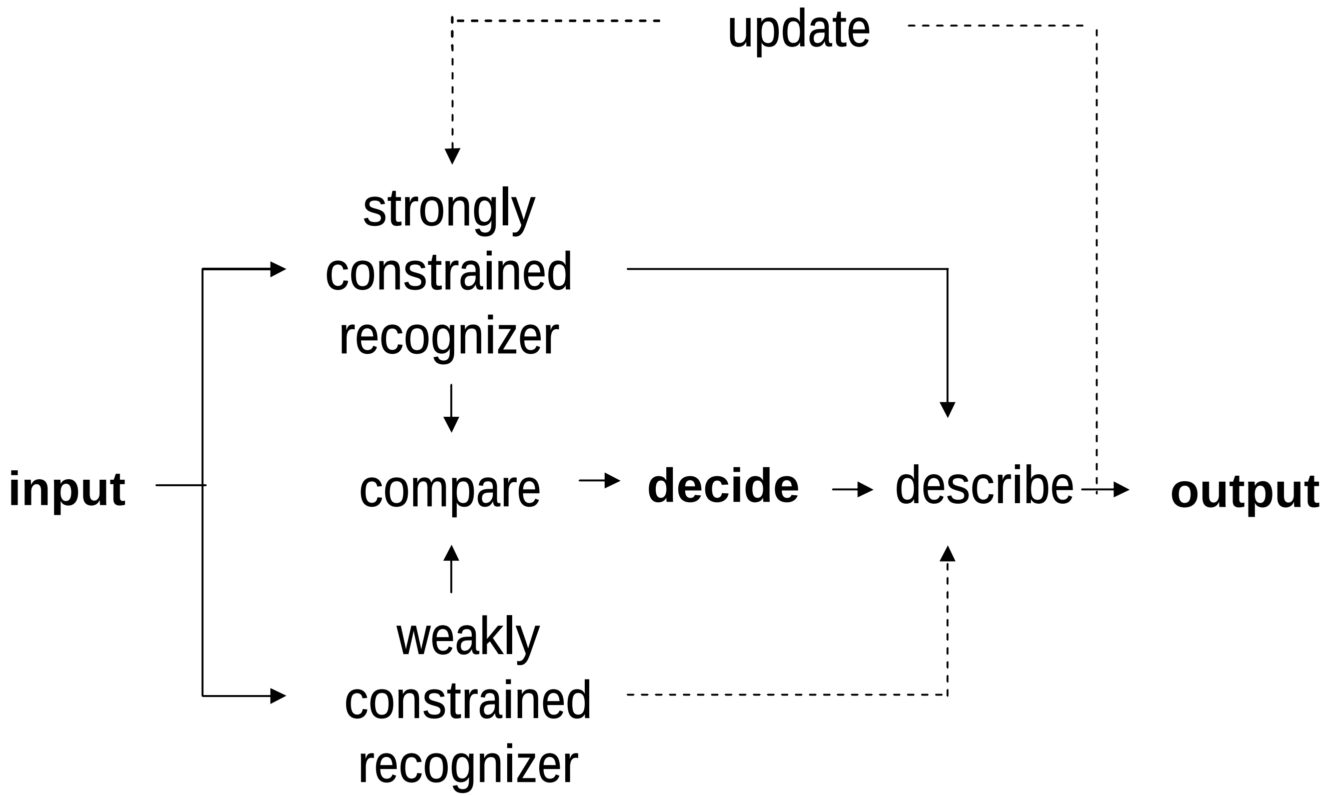
Hynek

Introduction

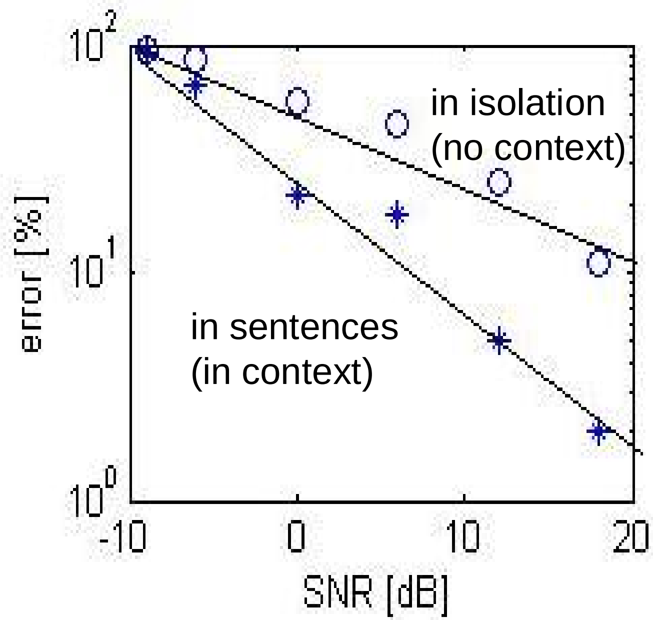
Hynek Hermansky

Dealing with unexpected inputs

- Out-of-vocabulary (OOV) words
  - unavoidable
    - proper names, out-of-language, invented, …
  - damaging
    - errors spread (each OOV causes in average 2 word errors)
  - rare
    - lower impact of the final WER **J**
  - **unexpected**
    - **therefore information-rich L**


update
strongly constrained recognizer
input
compare
decide
describe
output
weakly constrained recognizer

Word errors in human recognition of speech



$error_{context} = error^{k}_{no\ context}$

$\boldsymbol{error_{context} = error_{no\ context}\ error_{context\ channel}}$

errors multiply

– context (top-down) channel **is in parallel** with the acoustic (bottom-up) channel

Miller 1962

-interpretation by Boothroyd and Nittrouer 1998

-credit to J. B. Allen

OOV detection in small (digit) task

- telephone quality continuous digits, one digit left out from the lexicon

speech

↓

neural net derived phoneme posteriors

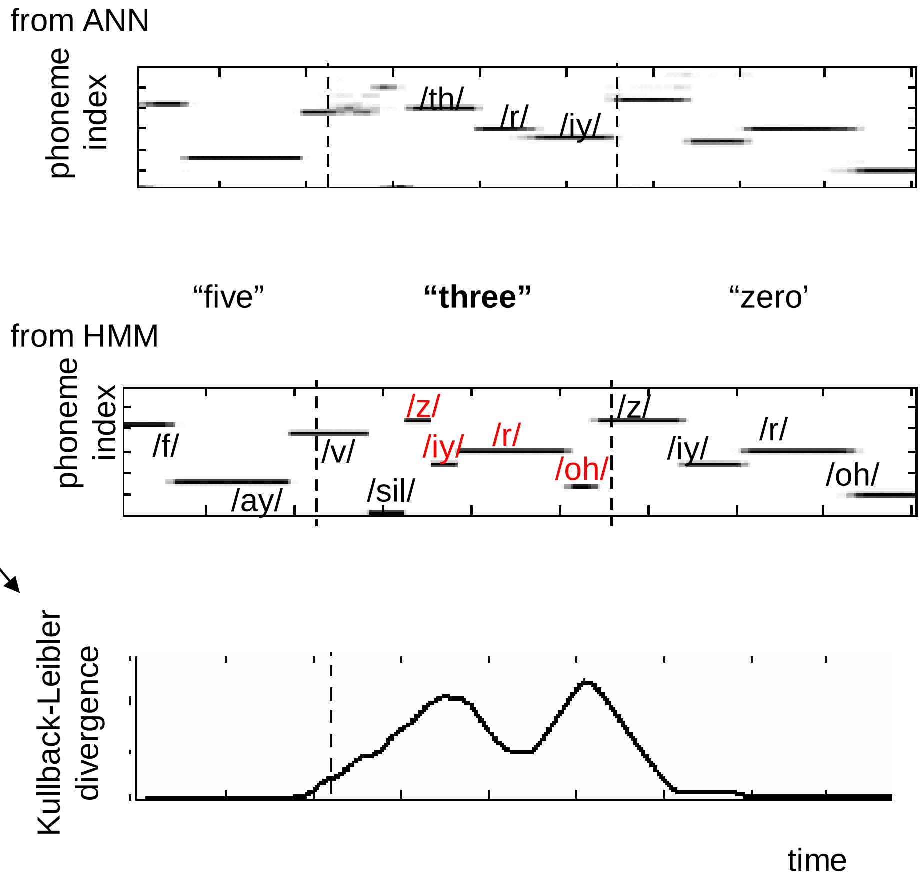
↓

global match on utterance level

↑

lexicon

compare posterior distributions



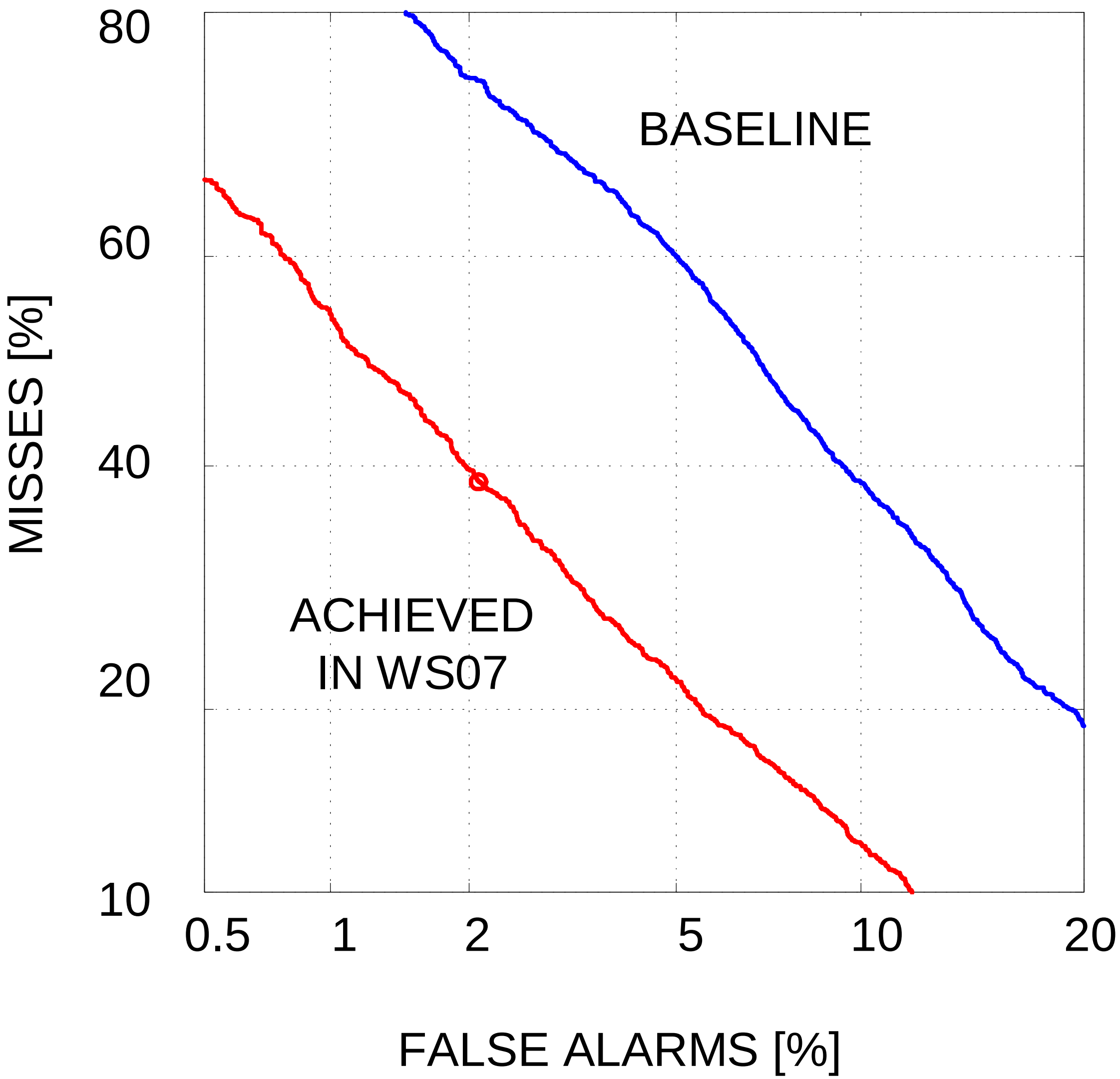
Ketabdar et al, Interspeech 2007

Decisions for WS07

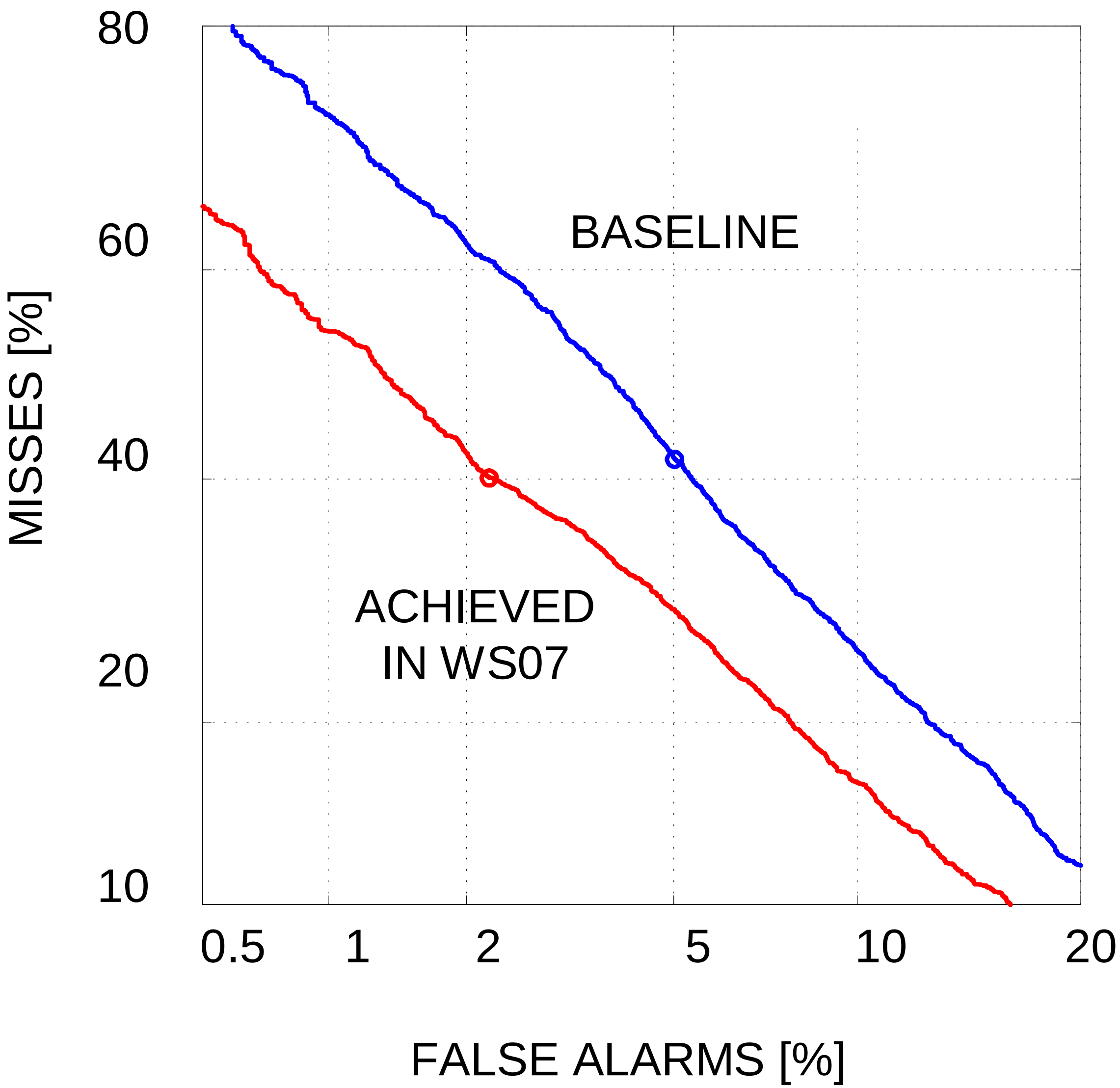
- 5 k word Wall Street Journal task, down-sampled to 4 kHz bandwidth
  - development and test sub-sets
    - no OOVs
    - 20% OOVs
- “strong” posteriors from the word lattice
- “weak” posteriors
  - directly from a neural network
  - from a phoneme recognizer
- both recognizers trained on Switchboard
  - can generalize to other tasks
  - some bias of the weak recognizer towards American English
- transducer-based approach
  - phoneme recognizer as “weak” recognizer
  - “strong” constraints from phoneme-based transducer
  - compare phoneme strings
- efforts towards Mandarin system
  - adaptation of Universal Phoneme Recognizer
  - training of “strong” recognizer on Mandarin
- towards 6 language Call home with transducer-based approach

And What Happened at WS07 ?

Detection of Out-Of-Vocabulary words (OOV)



Detection of Plain Errors



Main Achievements

- Demonstrated viability of out-of-vocabulary word (OOV) detection using parallel streams from strongly and weakly constrained recognizers
- Significantly improved the state of the art in confidence estimation in ASR of read speech
- Build phoneme-word transducers for 5 languages
  - 6 IPA lexica available
- Out-of-language (OOL) system built
- SMAP improvements over Universal Phoneme Recognizer for Mandarin
- Attribute-based phoneme recognizer applied to 4 languages
- New tool for estimating goodness of features for classification
- Built GMM and NN phone recognizers
- Built Mandarin LVCSR system

Conclusions

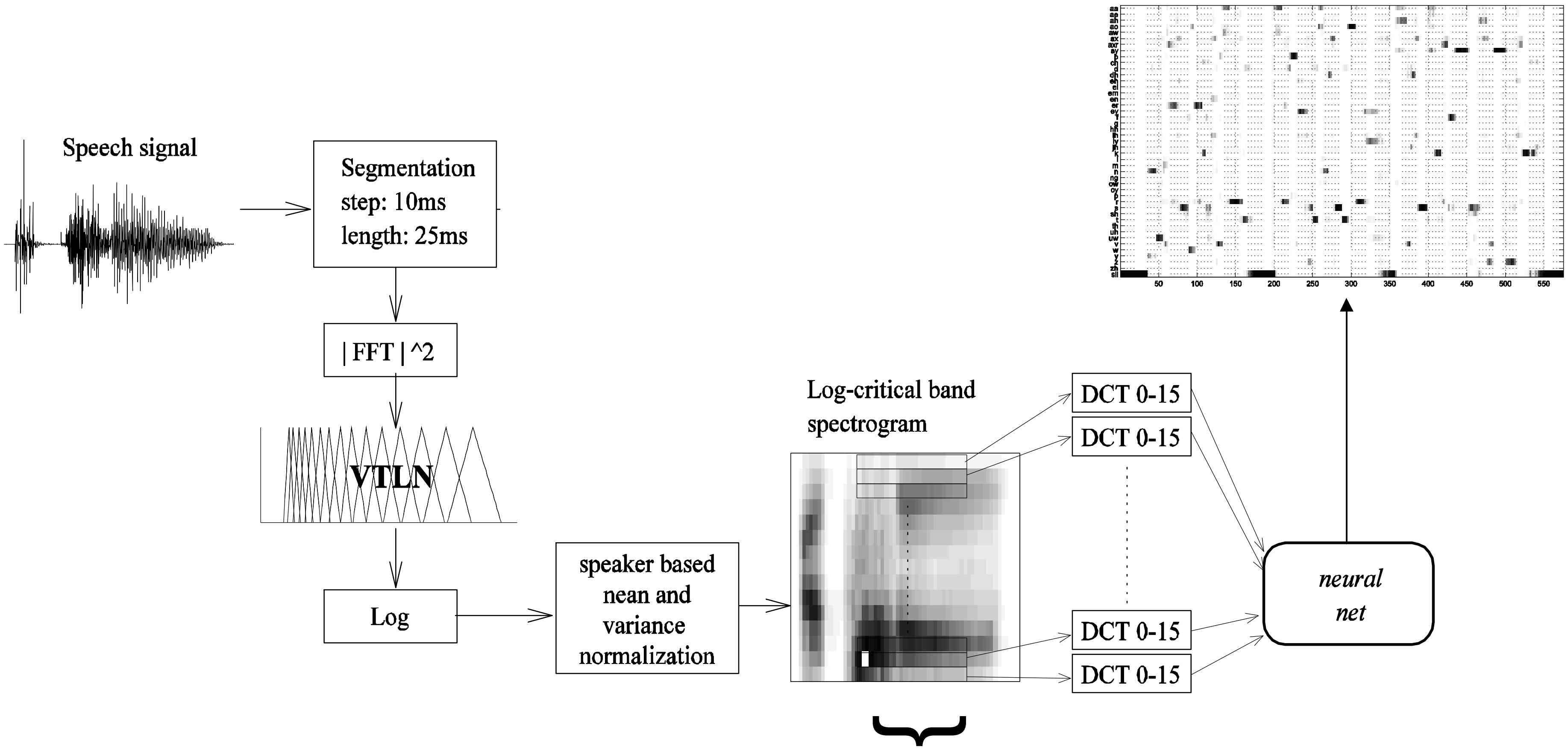
- Parallel strong and weak posterior streams are efficient for detection of OOVs
- OOVs and plain errors can be differentiated
- Trained statistical classifiers as a good alterative to divergence measures
- Mutual information is a viable measure of goodness of features for classification
- Improvements in phone accuracy important in many applications
- High error on spontaneous Mandarin makes OOL word detection hard
- SMAP works for adapting UPR models
- Attribute-based phoneme recognition promising as alternative to UPR

9:20-9:50

Detection of Out-Of-Vocabulary Words and Recognizer Confidence Estimation Using Parallel Strongly and Weakly Constrained Classifiers

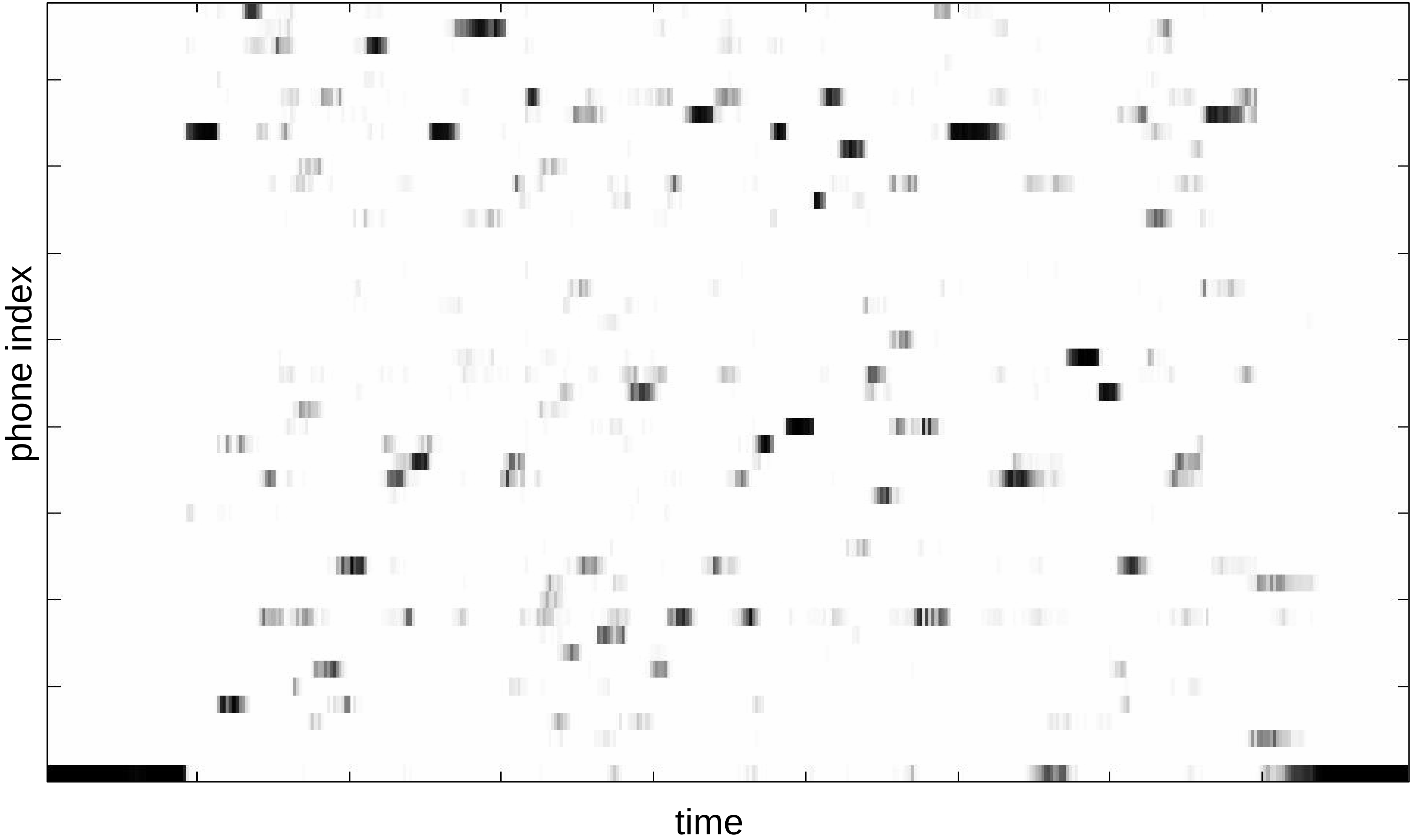
Lukas Burget, Petr Schwarz, Mirko Hannemann, Puneet Sahani

ANN Based Phone Posterior Estimator



300 ms

Example of Posteriogram

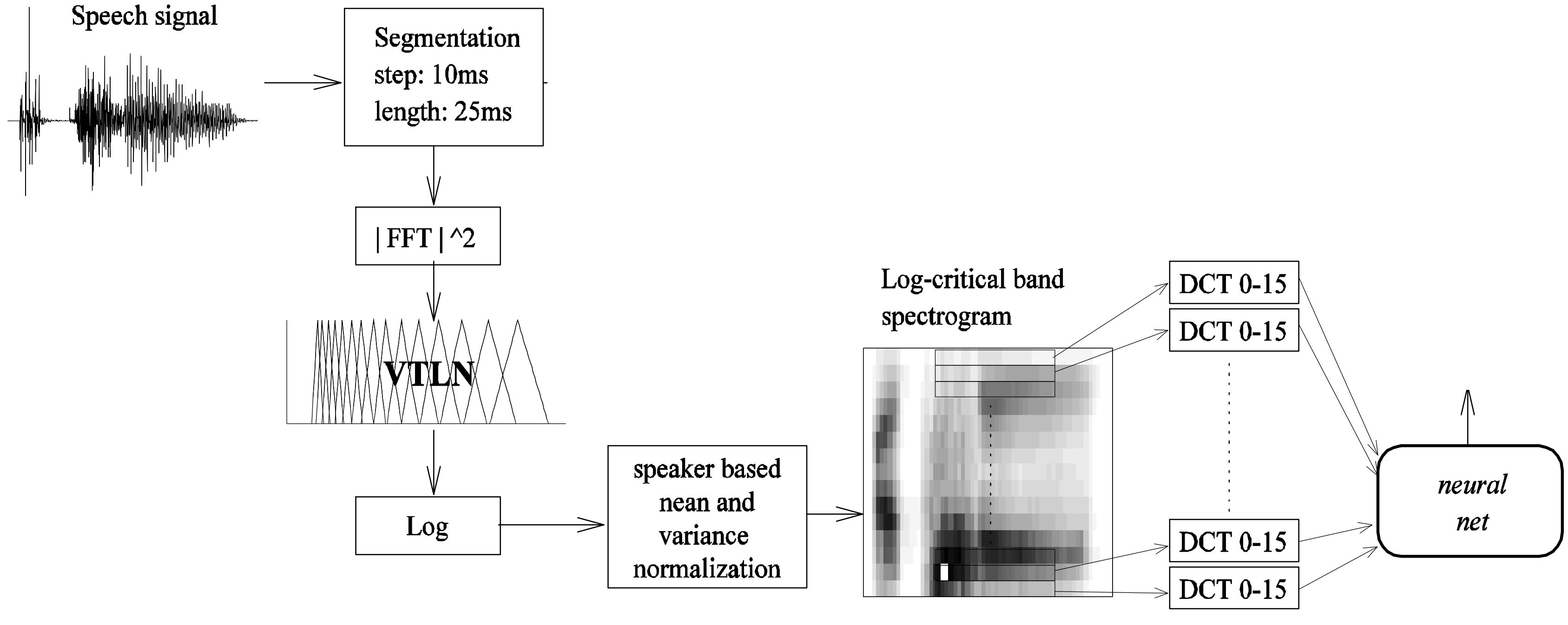


Hybrid ANN-HMM Phoneme Recognition

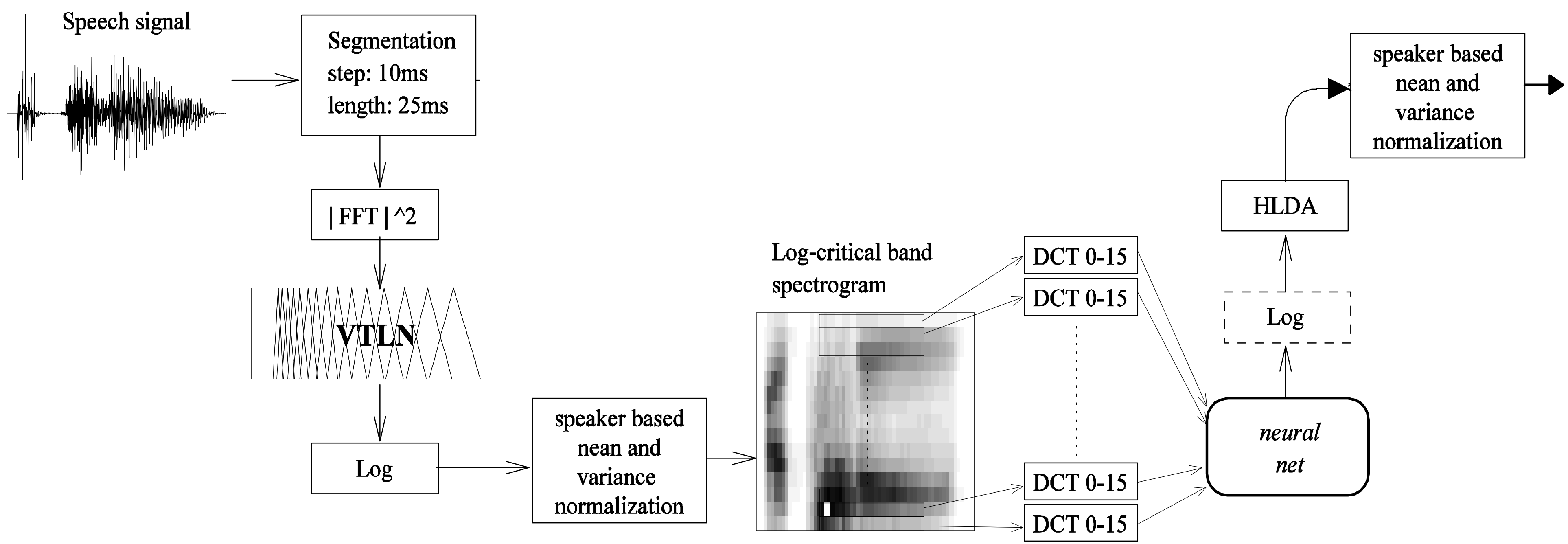
- **Effective for modeling speech sounds**
- Provides state-of-the-art phoneme recognition performance
- Performance of Brno phoneme recognizer on standard TIMIT database and comparison with other reported results for the same task:

Phoneme recognition system	Phone Error Rate
Schwarz: Hierarchical structure of neural nets	**21.8%**
Lamel: Triphone CD-HMM	27.1%
Ming: Bayesian Triphone HMM	27.1%
Deng: Hidden trajectory model	27.0%
Chang: Near--Miss modeling	25.5%
Robinson: Recurrent neural nets	25.0%
Halberstadt: Heterogenous Measurements	24.4%

Phoneme Posteriors as Features for LVCSR



Phoneme Posteriors as Features for LVCSR



LVCSR - System

- LVCSR is derived from the meeting data recognition system developed for European project AMI (University of Sheffield, Brno University of Technology, IDIAP, University of Edinburgh)
- System is trained on 250 hours of CTS data (SWB)
- Three pass decoding:

Pass1:
PLP+Δ+ΔΔ +ΔΔΔ
HLDA, MPE
3-gram decoding

→

Pass2:
VTLN, PLP+Δ+ΔΔ +ΔΔΔ
HLDA, MPE, CMLLR-SAT
2-gram decoding
3 or 4-gram lattice rescoring

→

Pass3:
VTLN, NN posteriors+Δ
HLDA, MPE, CMLLR-SAT
2-gram decoding
3 or 4-gram lattice rescoring

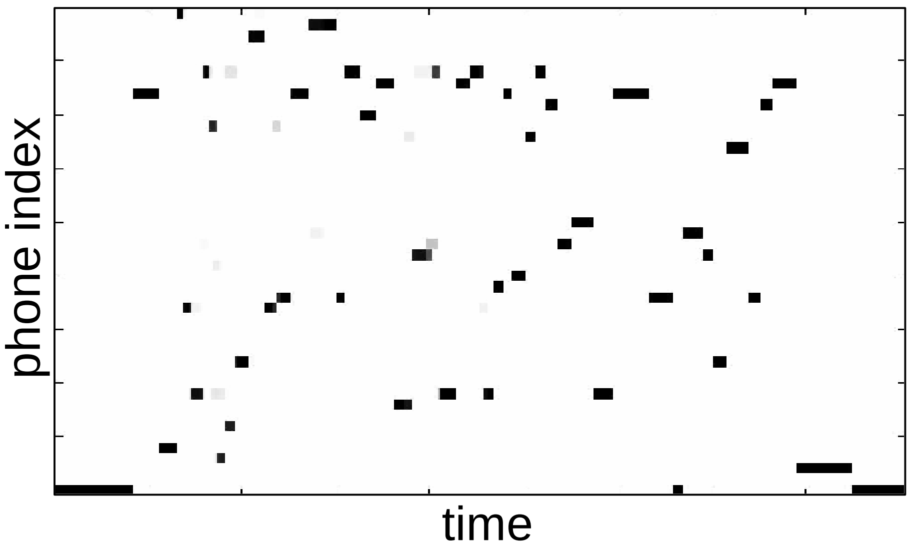
LVCSR system performance

- On WSJ0, November 92, Hub2 test set
  (330 utterances down-sampled to 8kHz)
  – closed set 5k wrord 3-gram LM: 2.9% WER
  – open set 20k words 4-gram CTS LM 12% WER
- On NIST RT01 (eval01) 50k word 4-gram LM
  – Pass2 (PLP features): 22.6% WER
  – Pass3 (NN features): 22.1% WER
- Possible improvements:
  – More elaborate NN based features mentioned above (- 2% WER)
  – Training on more date (fisher data **à** -2% WER)
  – Confusion network decoding and system combination (- 2% WER)
  – Corresponds to state-of-the-art performance on this task

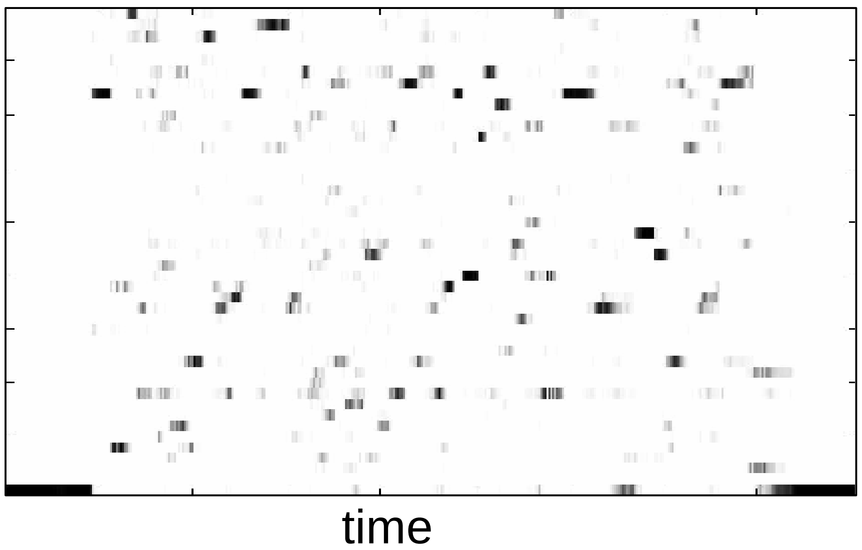
HMM Based Phone Recognizer

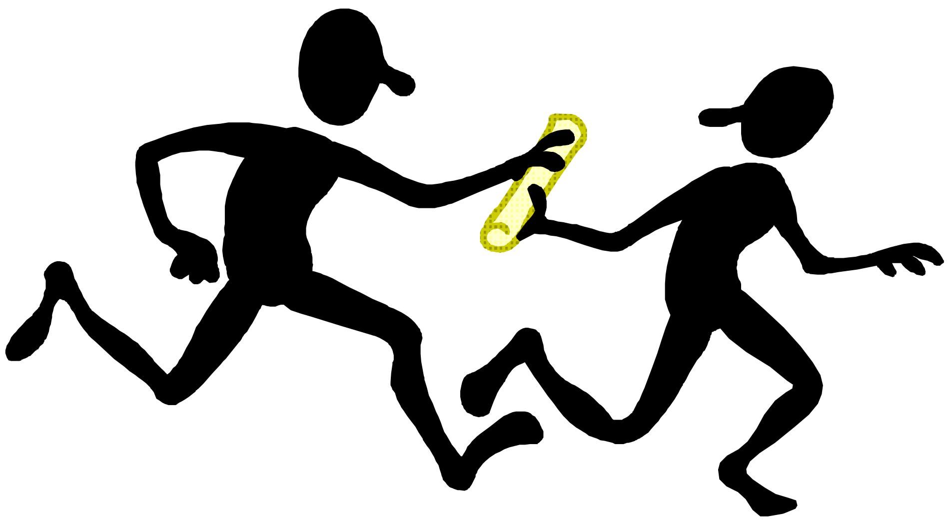
- Weakly constrained recognizer
- Acoustic models taken from LVCSR system
- Bigram phonotactic model
- 15.4% PER on WSJ0, November 92, Hub2 test set

Posteriogram derived from phoneme recognizer



Posteriogram derived using NN phone posterior estimator





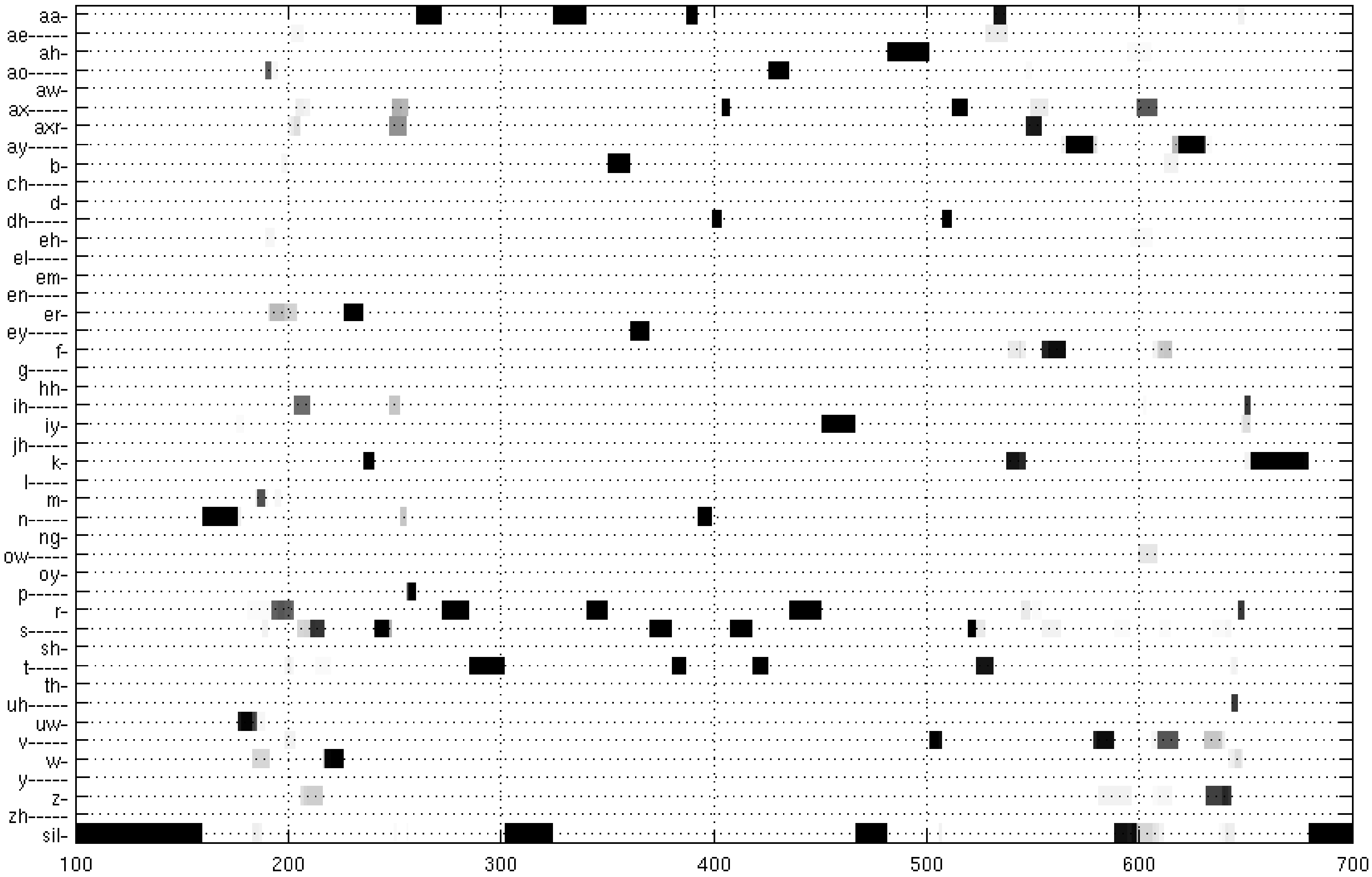
Visualizing the Data

to study Inconsistency Phenomena

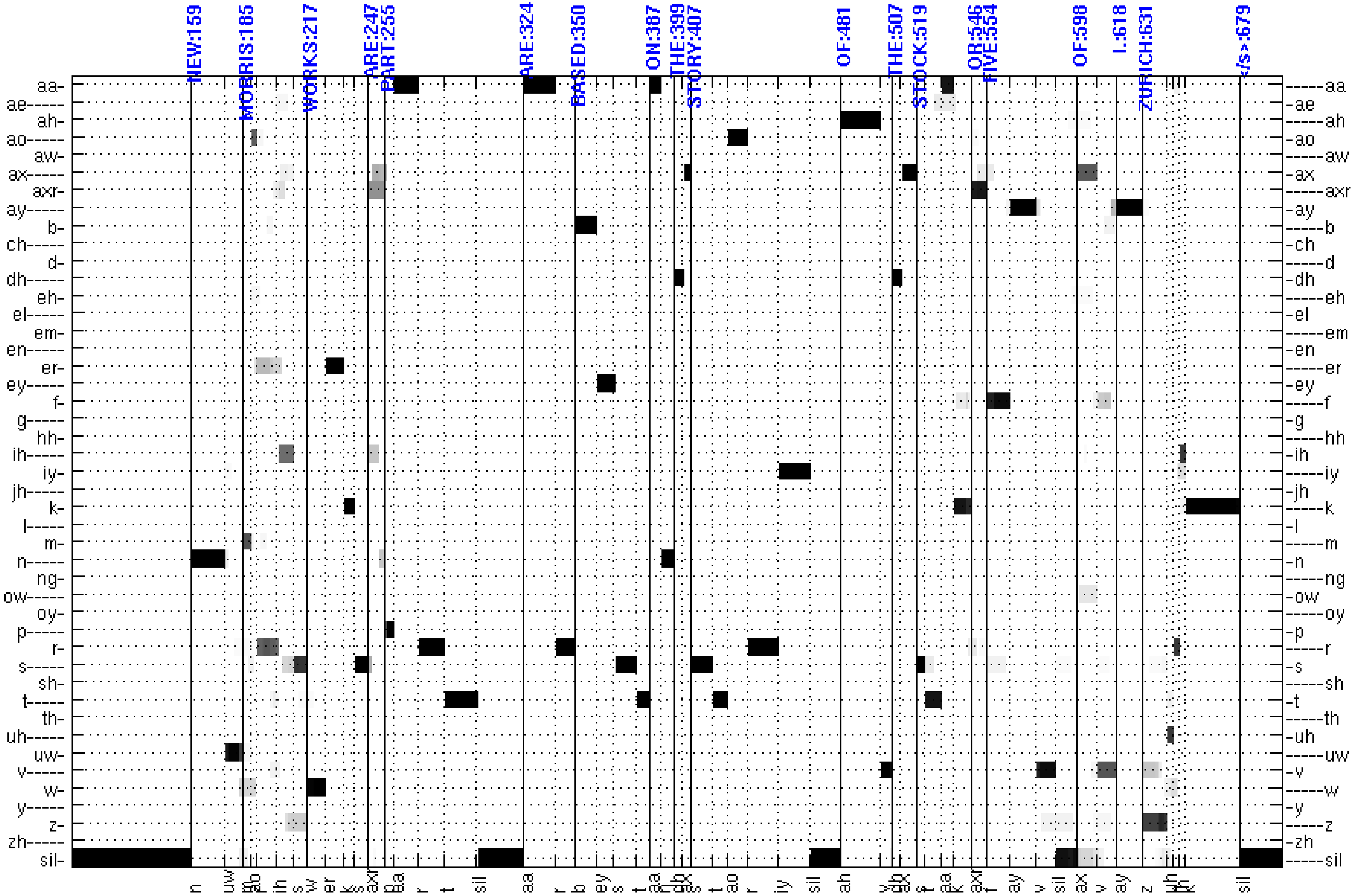
Puneet Sahani

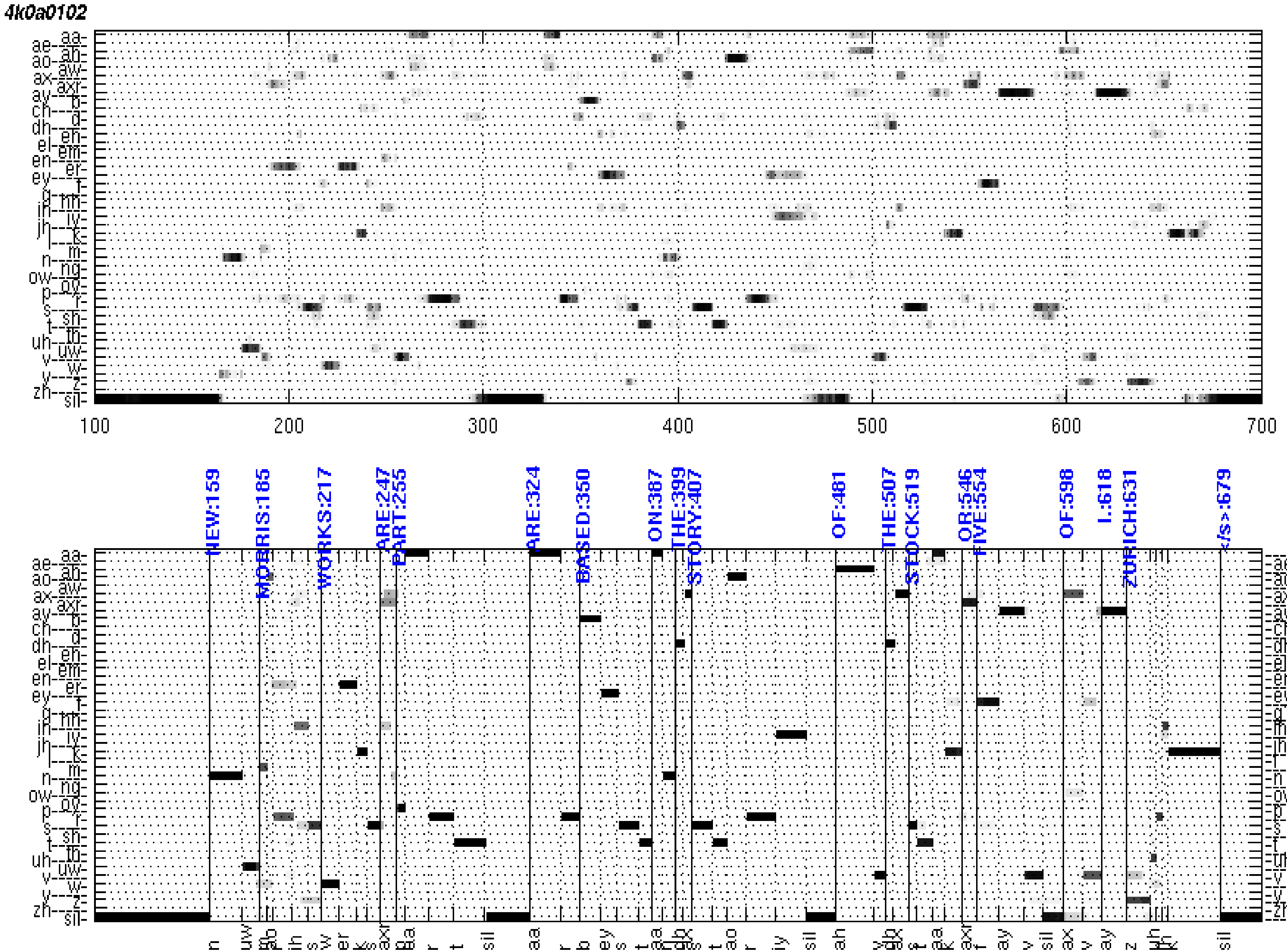
Mirko Hannemann

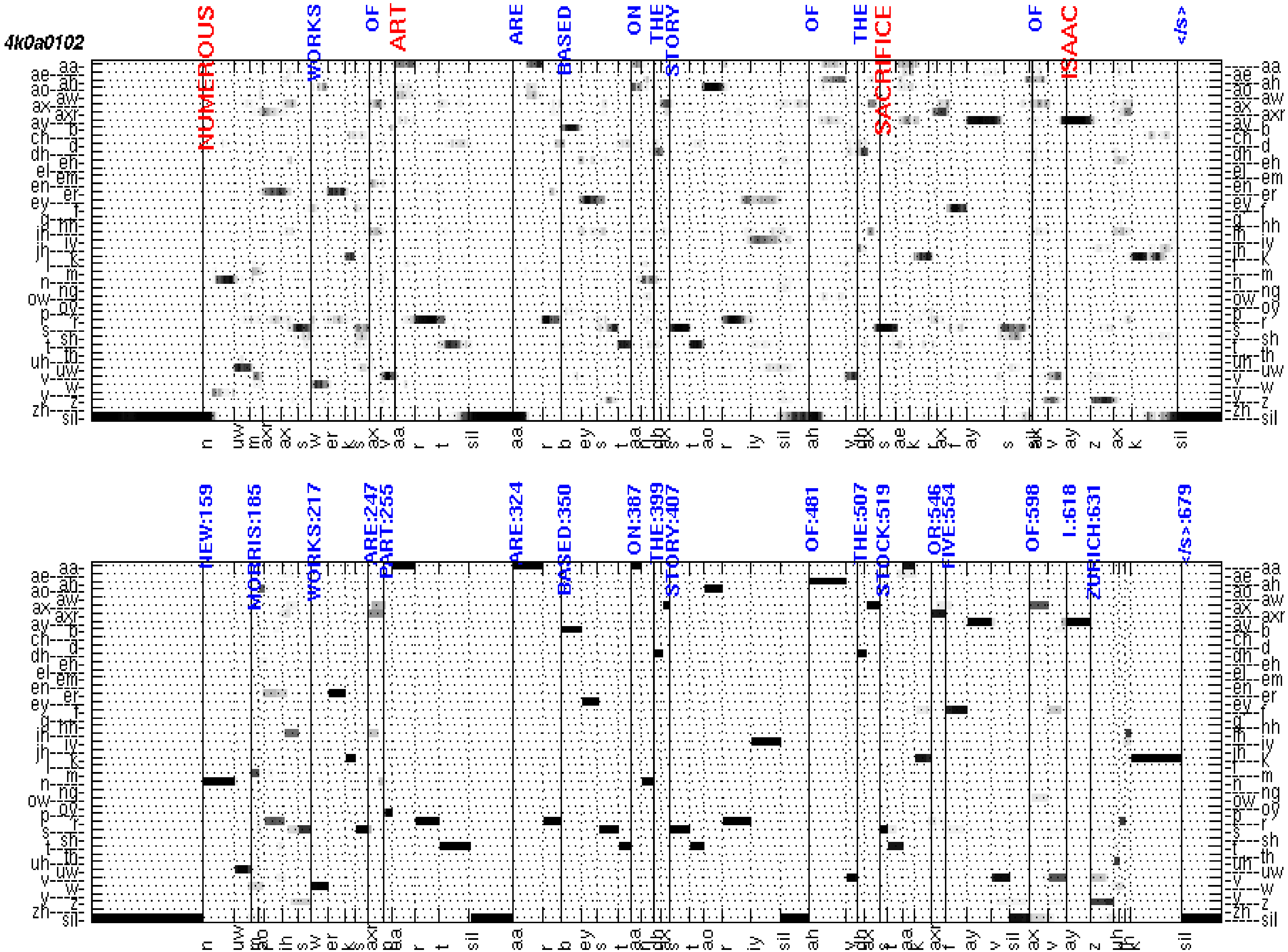
Strongly Constrained Recognizer

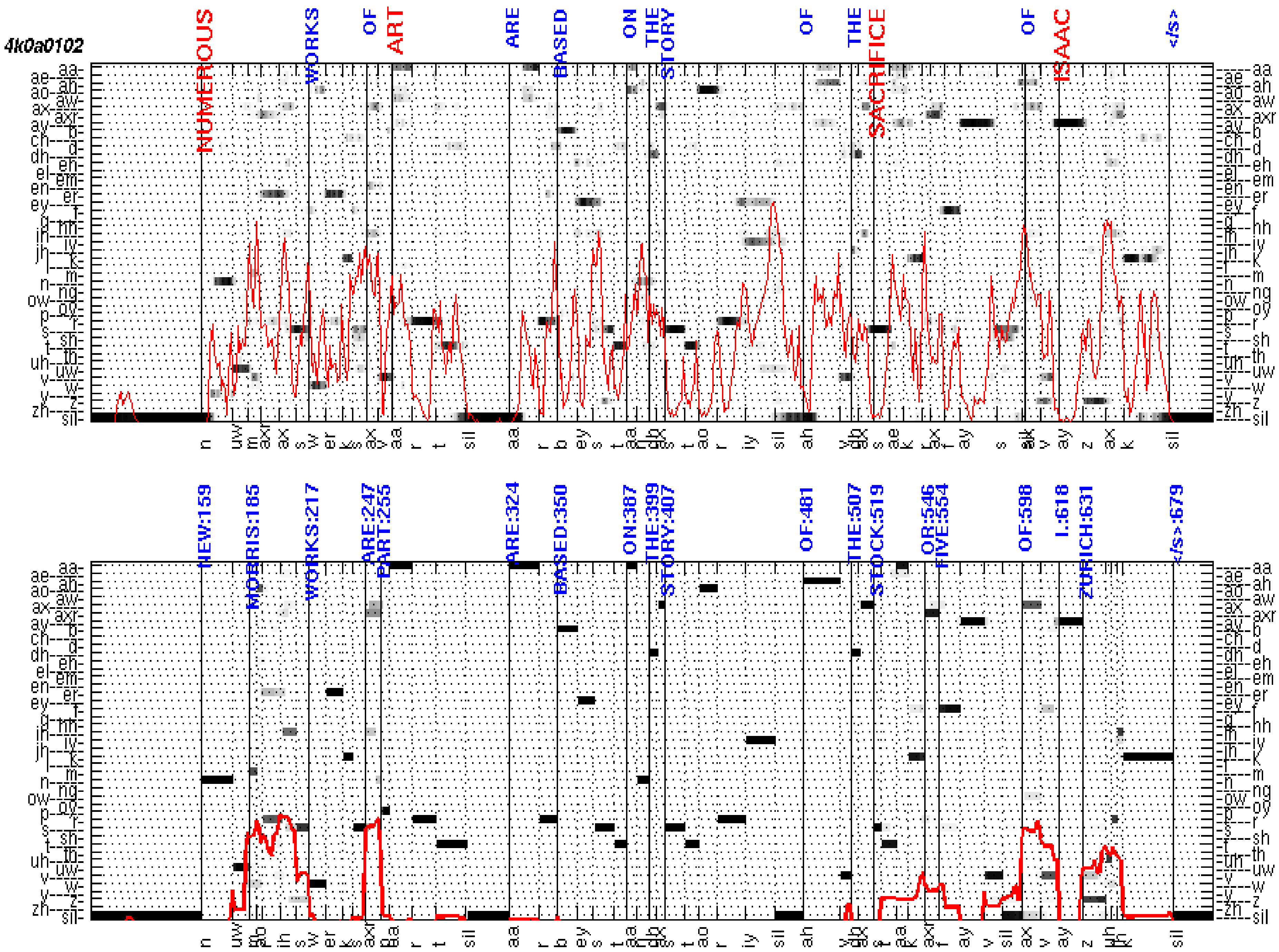


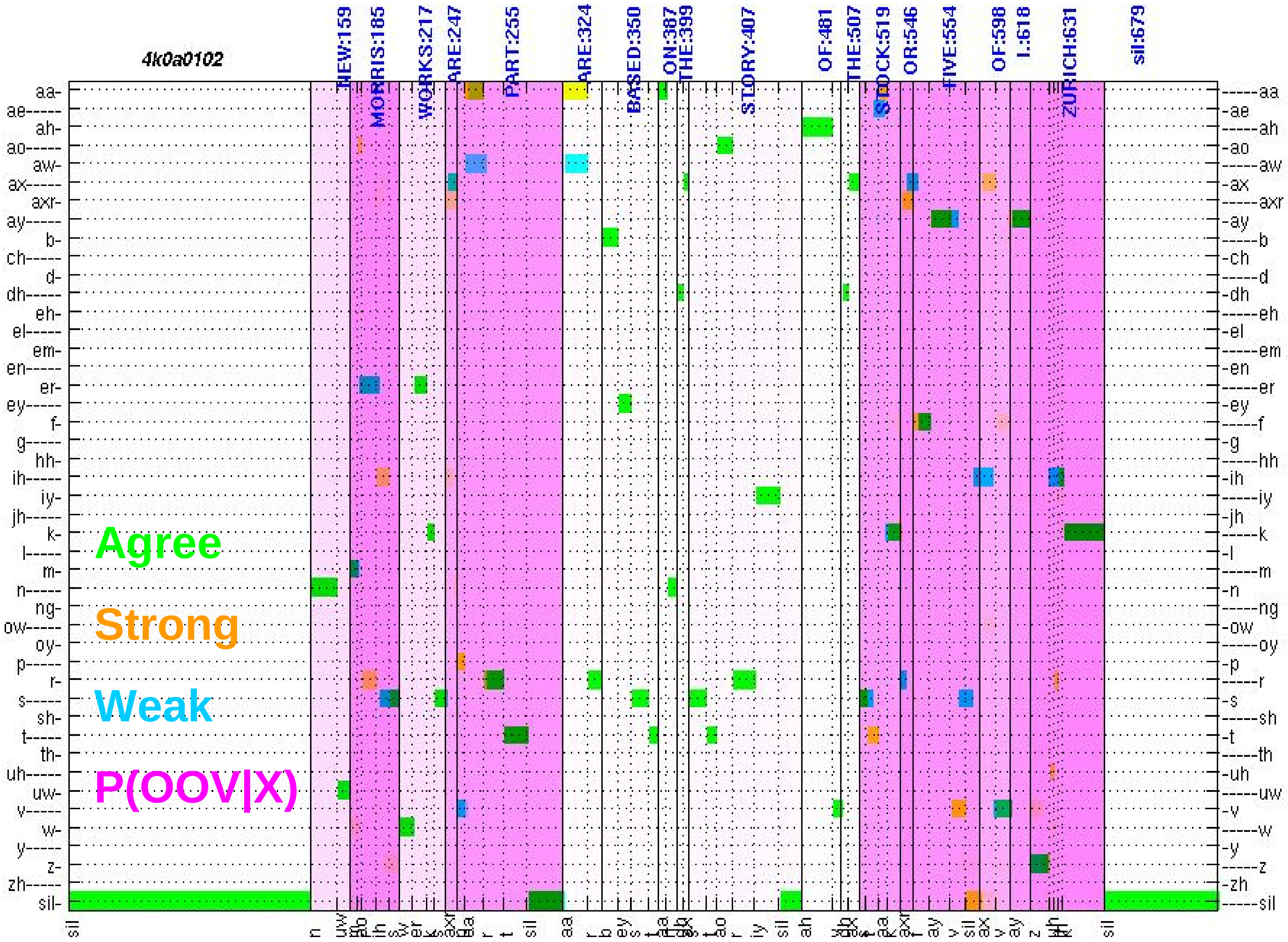
Strongly Constrained Recognizer




4k0a0102
NEW:159
MORRIS:185
WORKS:217
ARE:247
PART:255
ARE:324
BASED:350
ON:387
THE:399
STORY:407
OF:481
THE:507
STOCK:519
OR:546
FIVE:554
OF:598
I:618
ZURICH:631
</s>:679


4k0a0102
NUMEROUS WORKS OF ART ARE BASED ON THE STORY OF THE SACRIFICE OF ISAAC </s>
NEW:159 MOBRIS:185 WORKS:217 ARE:247 PART:255 ARE:324 BASED:350 ON:387 THE:399 STORY:407 OF:481 THE:507 STOCK:519 OR:546 FIVE:554 OF:598 I:618 ZURICH:631 </s>:679


4k0a0102
NUMEROUS WORKS OF ART ARE BASED ON THE STORY OF THE SACRIFICE OF ISAAC </s>
NEW:159 MORRIS:185 WORKS:217 ARE:247 PART:255 ARE:324 BASED:350 ON:387 THE:399 STORY:407 OF:481 THE:507 STOCK:519 OR:546 FIVE:554 OF:598 I:618 ZURICH:631 </s>:679



4k0a0102
Agree
Strong
Weak
P(OOV|X)

Detection of OOV words

Lukáš Burget

Mirko Hannemann

Data Sets for Test and Development

- Wall Street Journal Corpus
- “to make a healthy animal sick”
- Introduce OOV by limiting dictionary
  to most frequent 5000 words
- Decode WSJ files with a state-of-the-art LVCSR system, using this 5K dictionary
- Word Error Detection task:
  detect words in ASR output which are wrong

Selecting Sets for Different Tasks

- Evaluate Word Error and OOV detection performance
- Provide recognition examples as a development set to train classifiers on Word Error and OOV detection

- Provide different partitions of test sets with varying OOV token rates:
- Whole test set: ~5% OOV token
- Subsets with 20% OOV token

OOV Detection

- Focusing on:
  - Detecting misrecognized words

ASR out	Television	shows	such	as	let's	hope	or	E.	O.	</s>
Reference	Television	films	such	as	Little	Gloria				sil

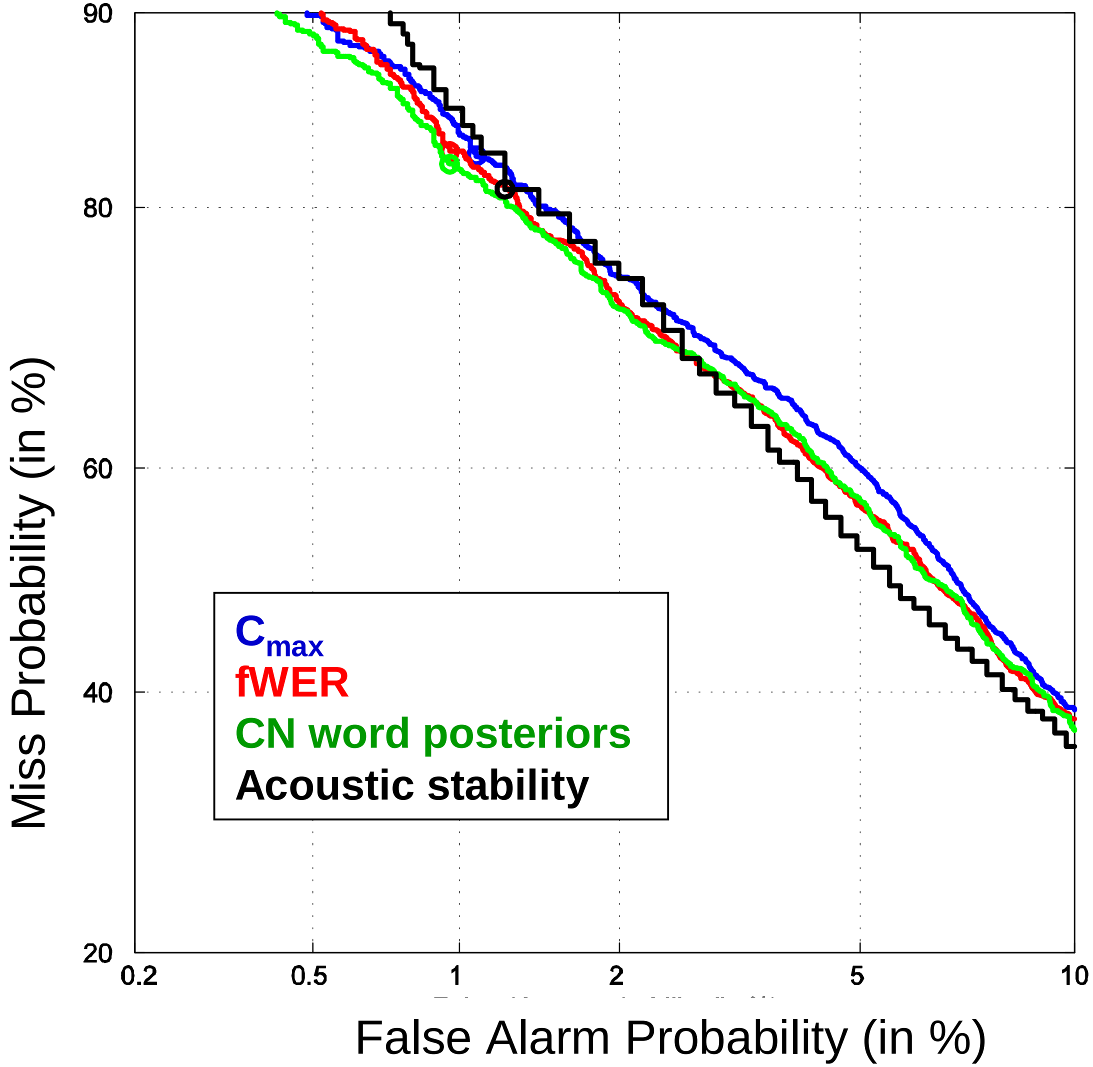
OOV Detection

- Focusing on:
  - Detecting misrecognized words
  - Detecting misrecognized words overlapping with OOV words

ASR out	Television	shows	such	as	let's	hope	or	E.	O.	</s>

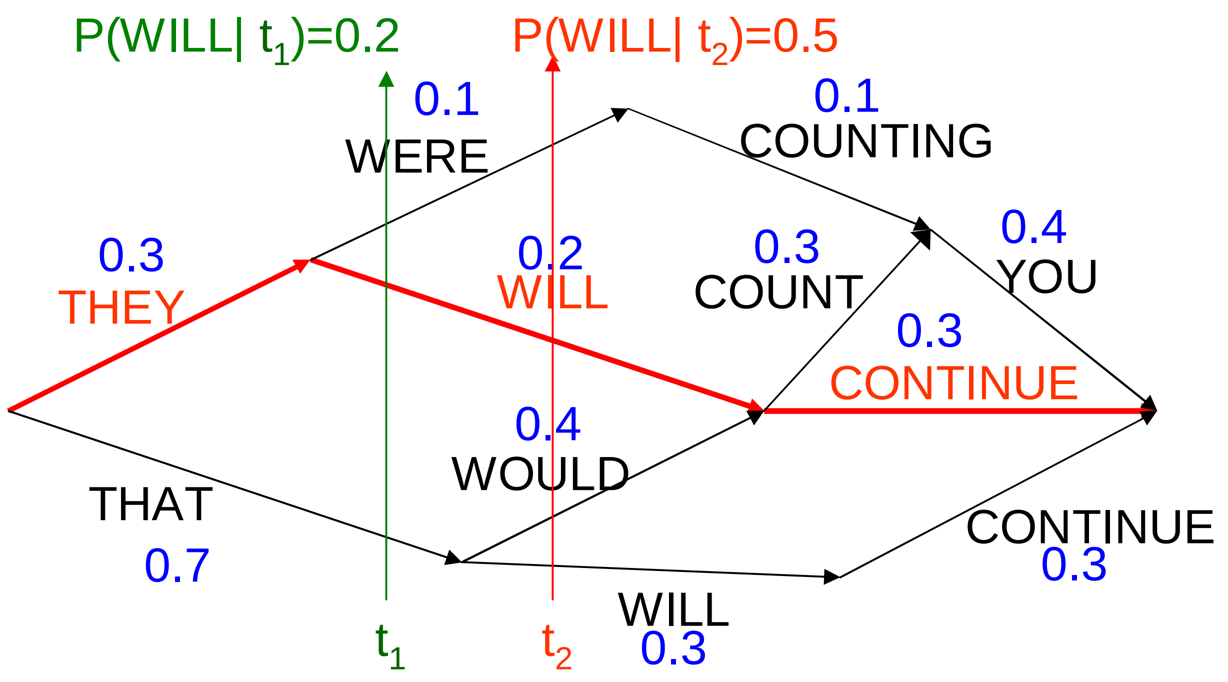
Reference	Television	films	such	as	Little	Gloria	sil

DET Curves for Standard Confidence Measures (OOV Detection)



Computing the Word Posterior Distribution

- For given frame of a hypothesized word, sum posterior probabilities of that word a lattice
- e.q. standard confidence measures
  - Cmax – maximum of the word posteriors over the hypothesized word
  - fWER – sum the word posteriors over the hypothesized word



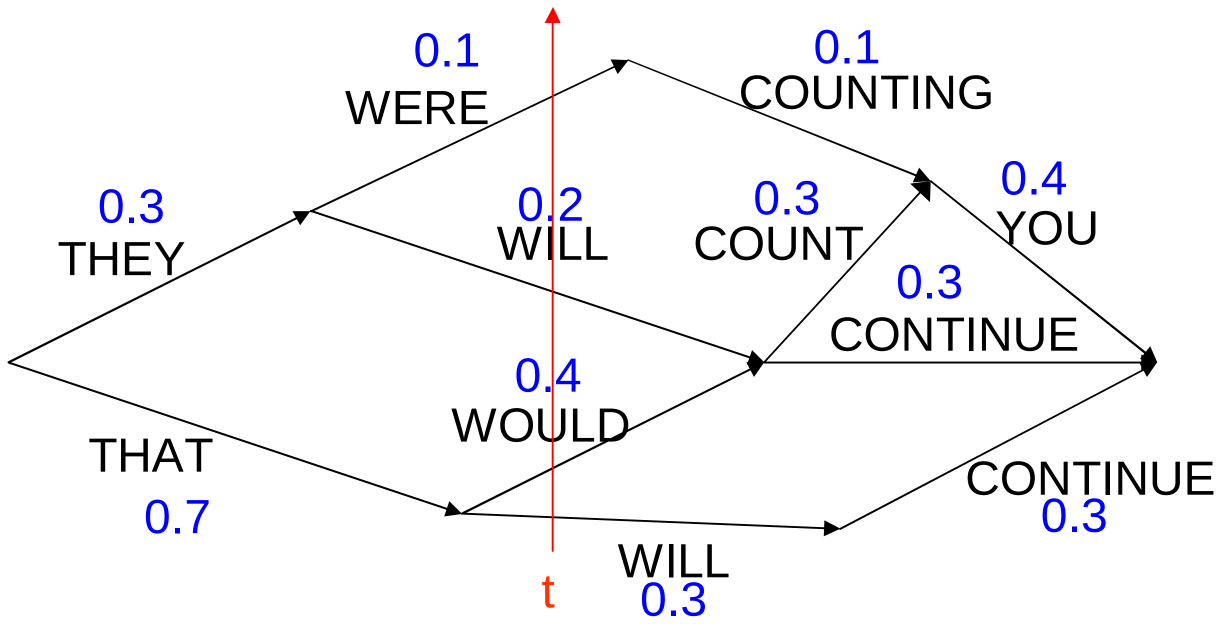
Computing Word Entropy

- For a given frame, sum the posterior probabilities of each word in a lattice and estimate entropy across words

P(WERE | t) =0.1

P(WILL | t) =0.5 è H(t) = 0.65

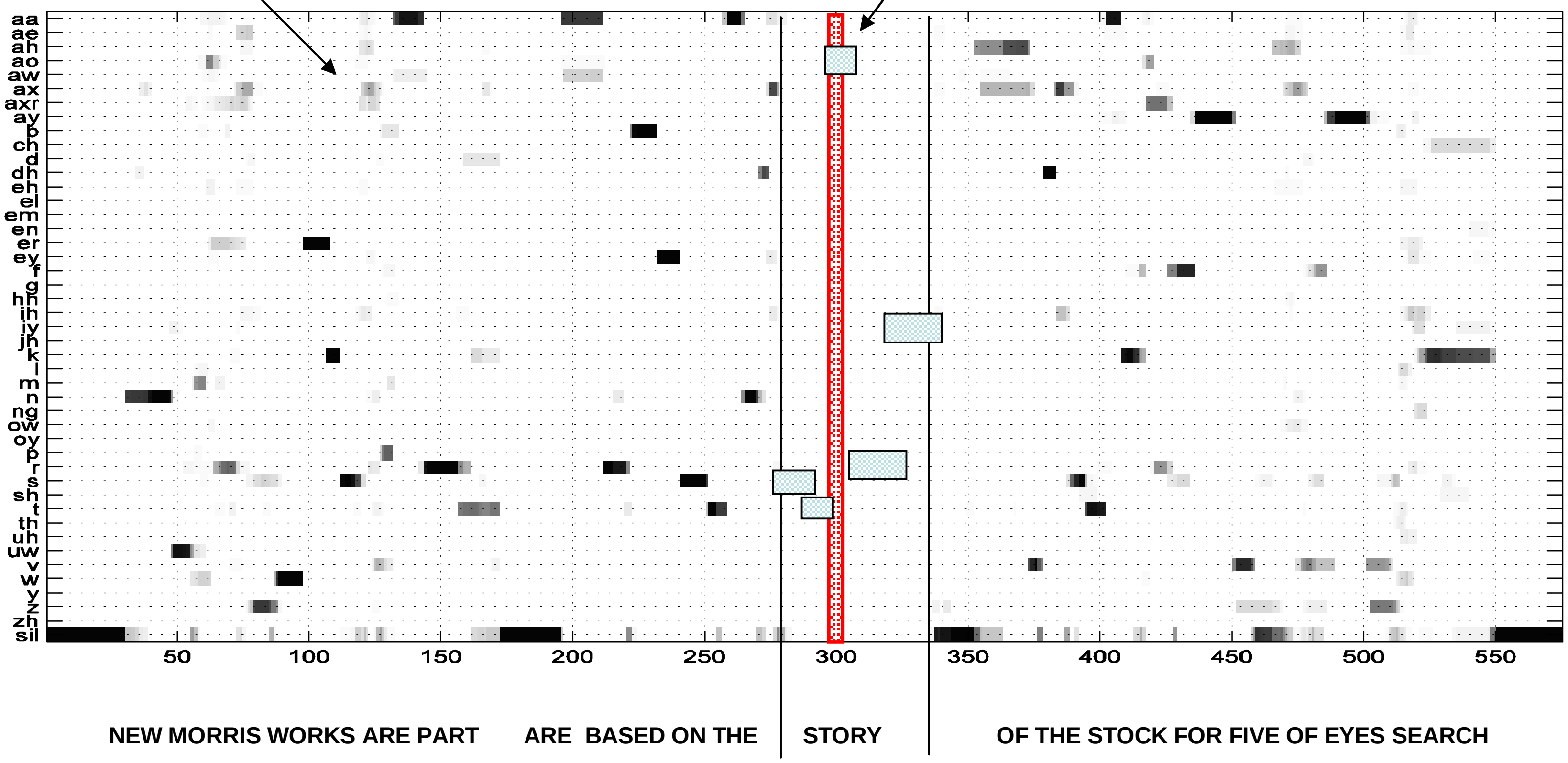
P(WOULD | t) =0.4



Posterior of a Hypothesized Phone

Posteriogram from strongly or weakly constrained recognizer

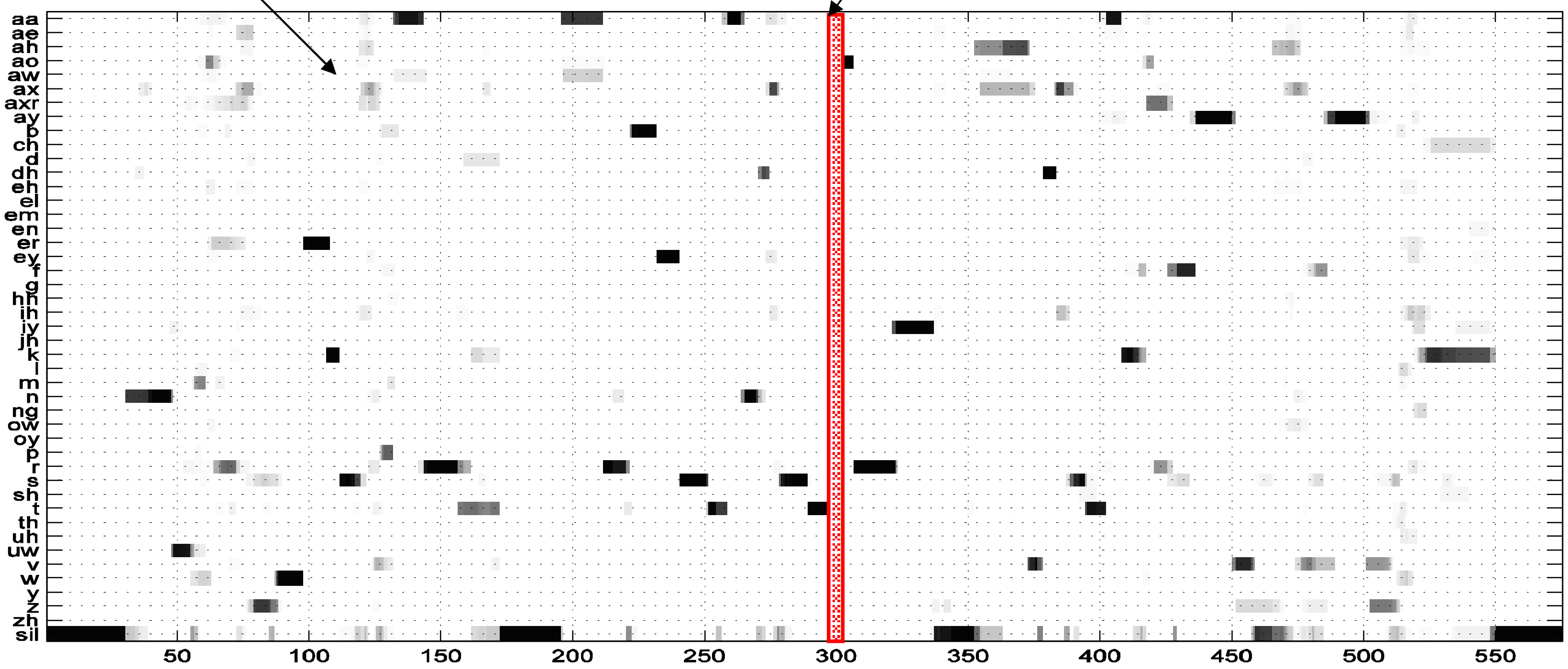
Hypothesized phone frame posterior probability



Computing Phone Entropy

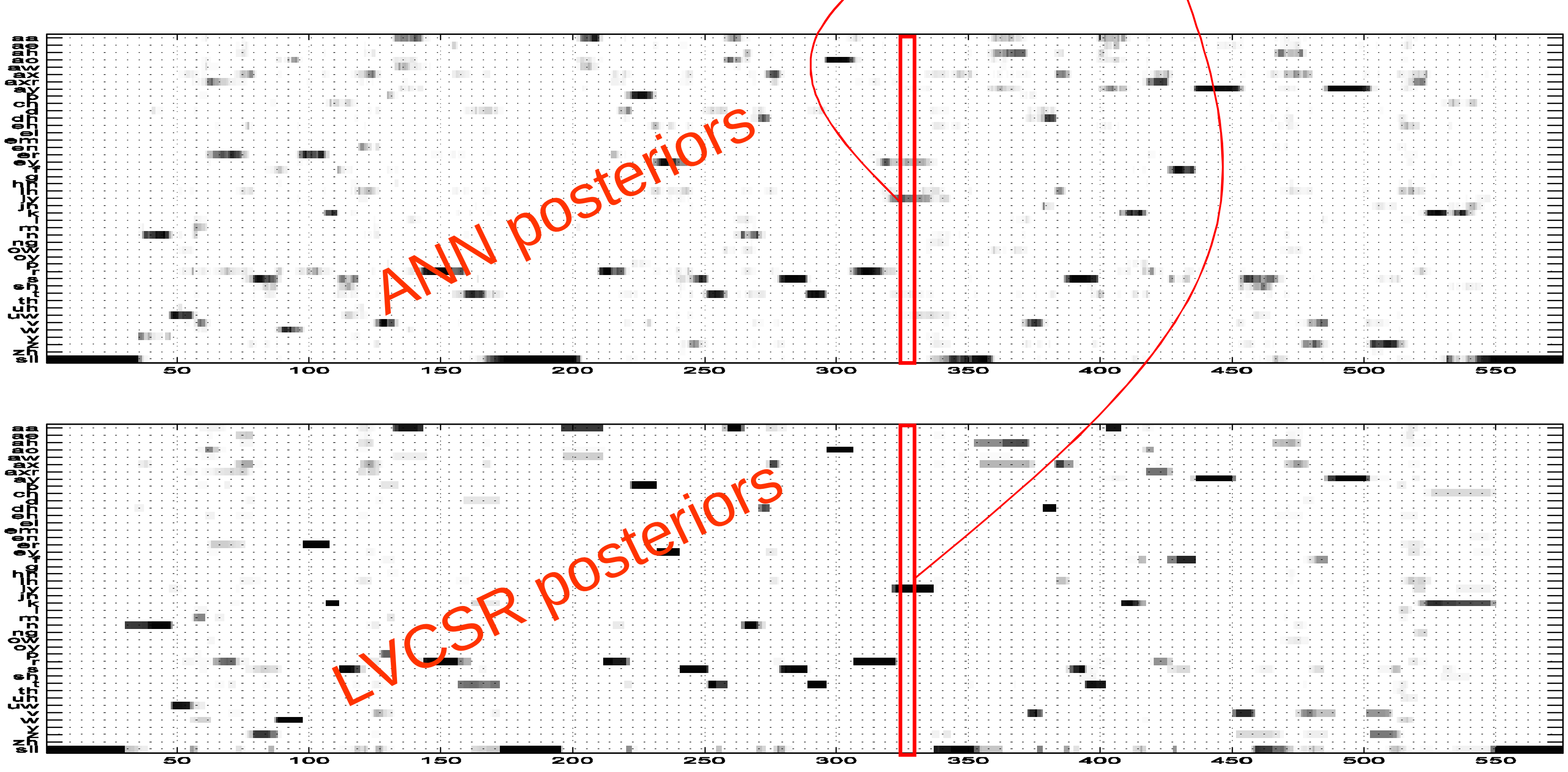
Posteriogram form strongly or weakly constrained recognizer

Compute frame entropy

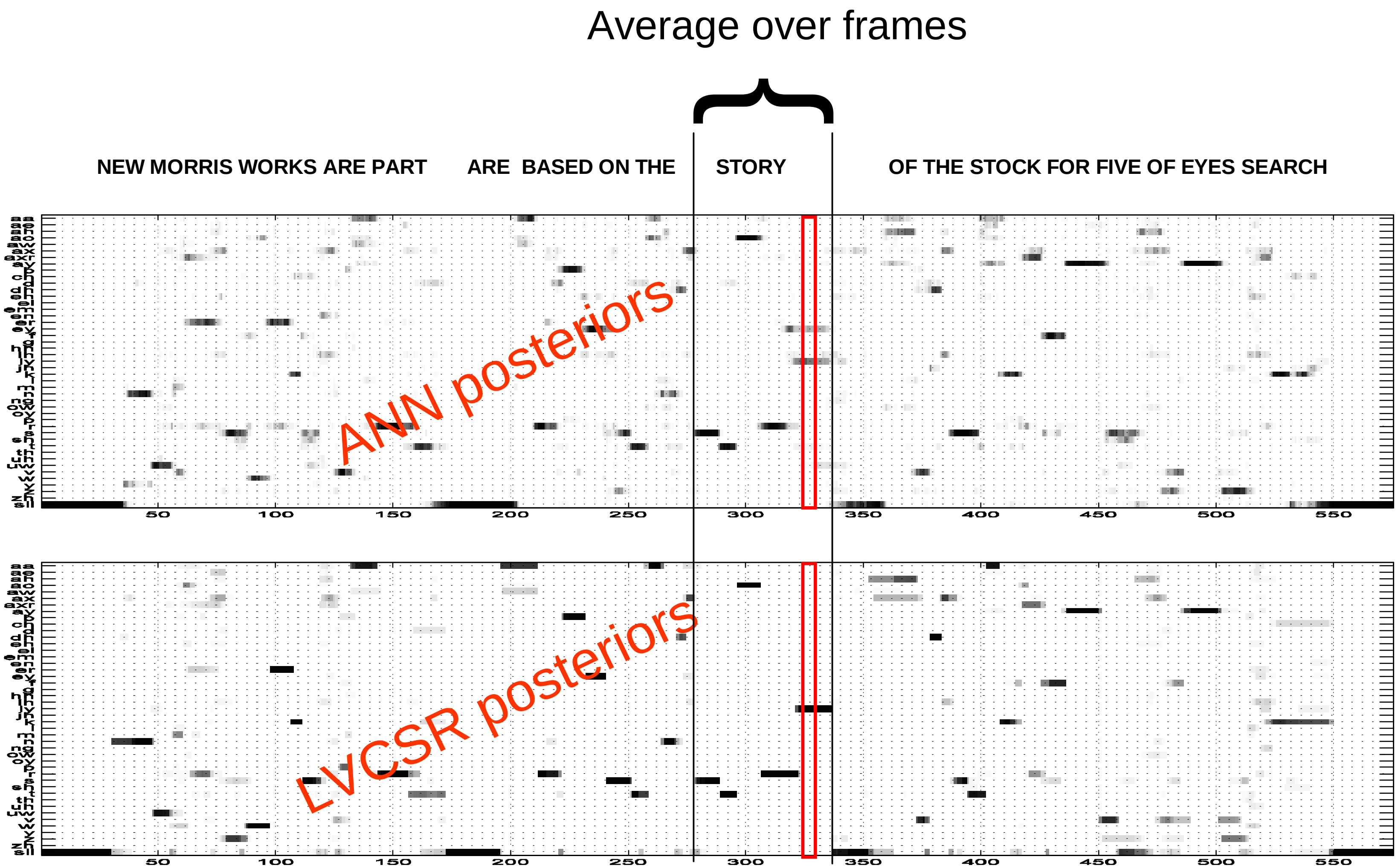


Finding Inconsistencies in LVCSR

Frame-by-frame divergence between ● and ●



Computing Word Level Scores



Computing Word Level Scores

- Maximum over hypothesized word boundary
- Variance over hypothesized word boundary
- Average over hypothesized word boundary
- Averaging over hypothesized phonemes / by number of phonemes
- Phones on word boundary troublesome
- Arithmetic vs. geometric average

Explored confidence measures

Only LVCSR
Only weakly constraint recognizer
Comparison of LVCSR weakly constraint recognizer

		entropy	Posterior	Divergence
LVCSR	word lattice		fWER, Cmax	
	phone posteriors			
Phone recognizer posteriors				
NN phone posterior estimator				

Few additional standard measures:

- Confusion networks based measures
- Acoustic stability
- Word acoustic and LM score
- Word duration

Ranking Measures for OOV Detection and Error

Ranking measures according to performance on:
- Detecting words overlapped with OOV words

Average rank of measures in a field

		entropy	posterior	Divergence
LVCSR	word lattice	5.0	12.5	
	phone posteriors	8.8	16.0	
Phone recognizer posteriors		29.0	21.0	19.0
NN phone posterior estimator		30.0	24.0	26.0

•other scores: 24.8

Ranking Measures for OOV Detection and Error

Ranking measures according to performance on:
- Detecting words overlapped with OOV words
- Detecting misrecognized words

Average rank of measures in a field

		entropy	posterior	Divergence
LVCSR	word lattice	5.0 7.0	12.5 12.0	
	phone posteriors	8.8 12.0	16.0 16.0	
Phone recognizer posteriors		29.0 28.0	21.0 24.0	19.0 22.5
NN phone posterior estimator		30.0 31.0	24.0 26.0	26.0 25.3

•other scores: 24.8 23.5

Summary for individual scores:

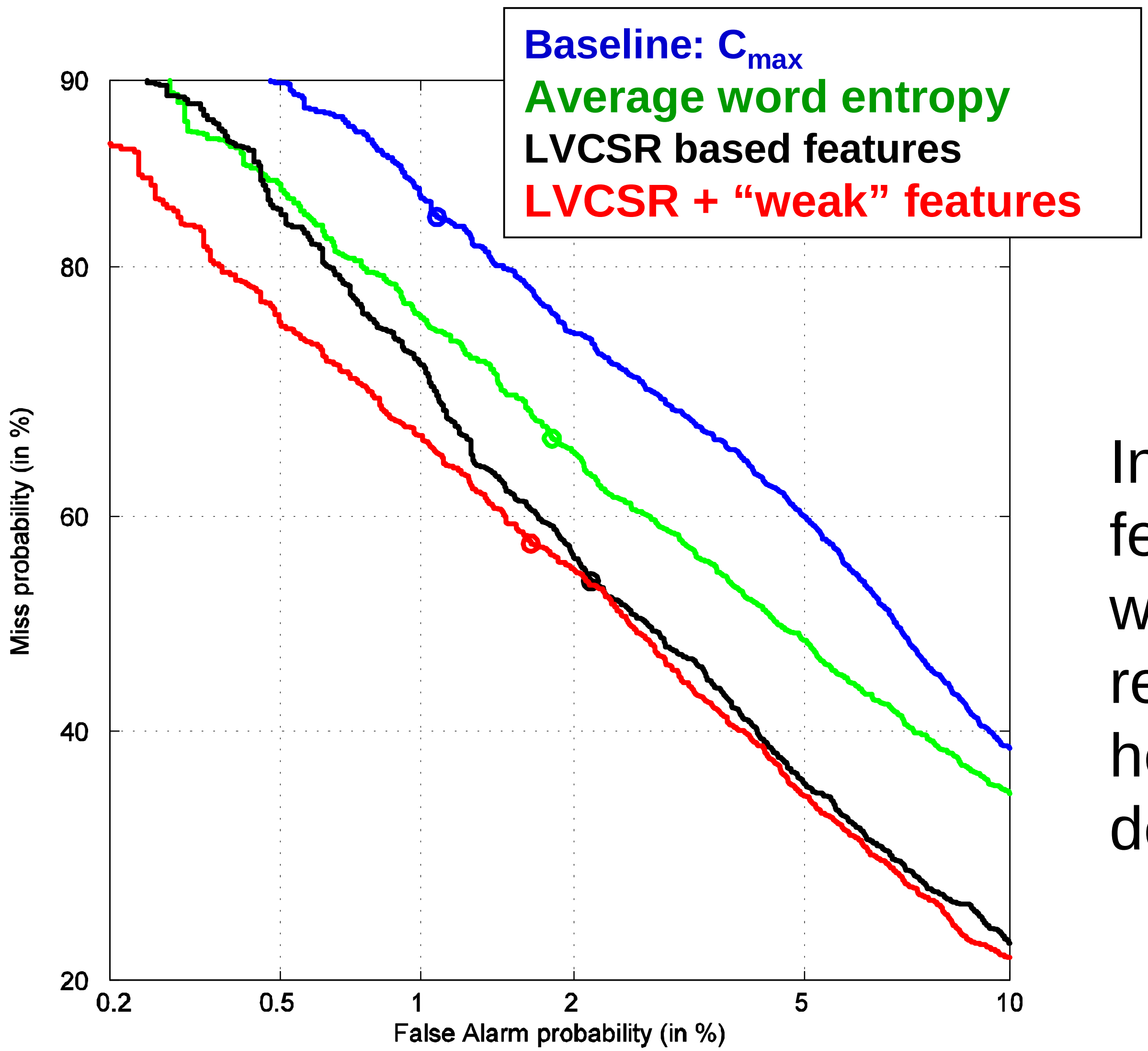
•LVCSR based entropy scores perform generally better that posterior scores especially for OOV detection

•Scores derived using only LVCSR are preferred individual scores

Combining Features

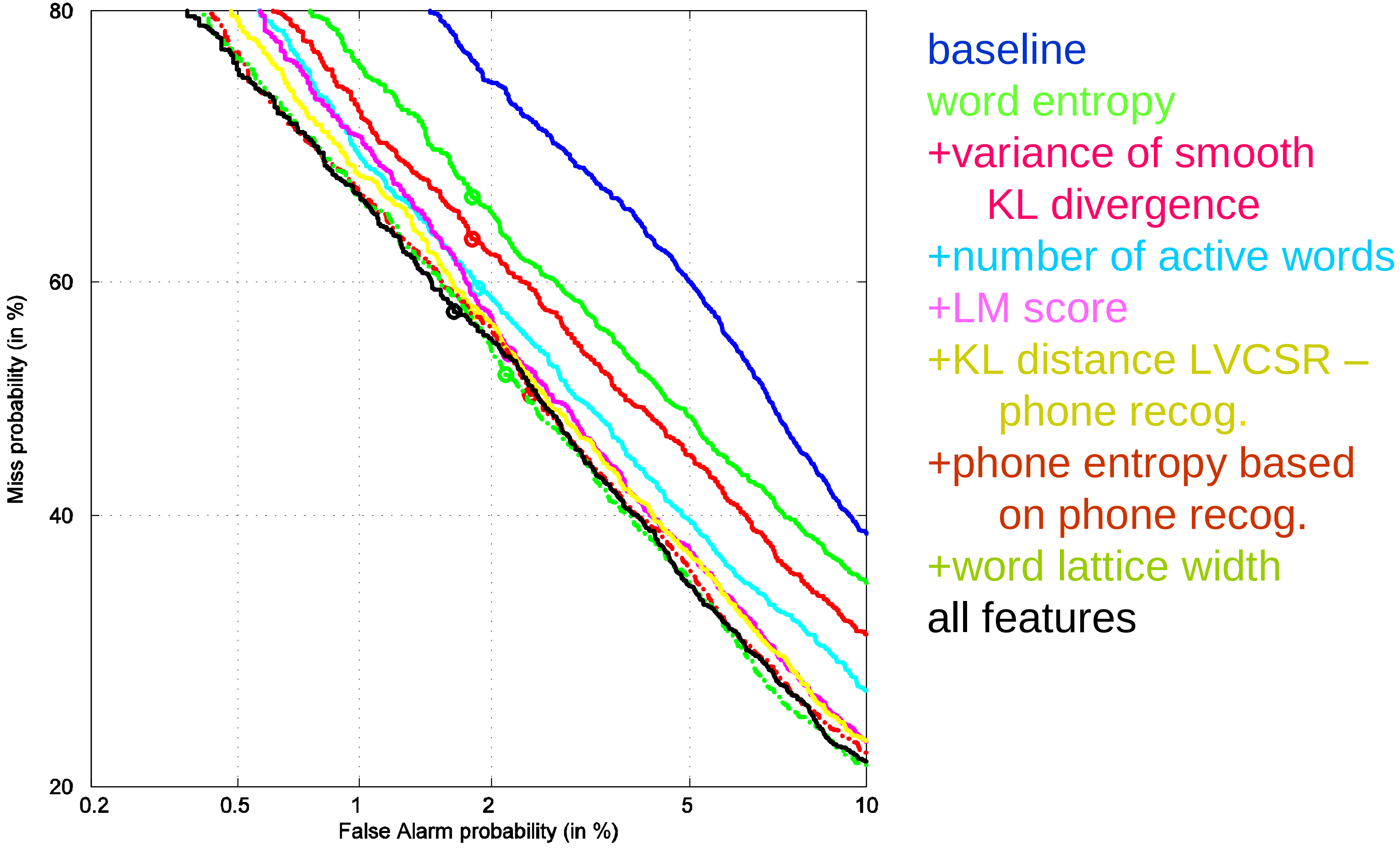
- Maximum Entropy classifier
- Trained for:
  - Detecting misrecognized words overlapping with OOV
  - Detecting misrecognized words

Combined System for OOV Detection

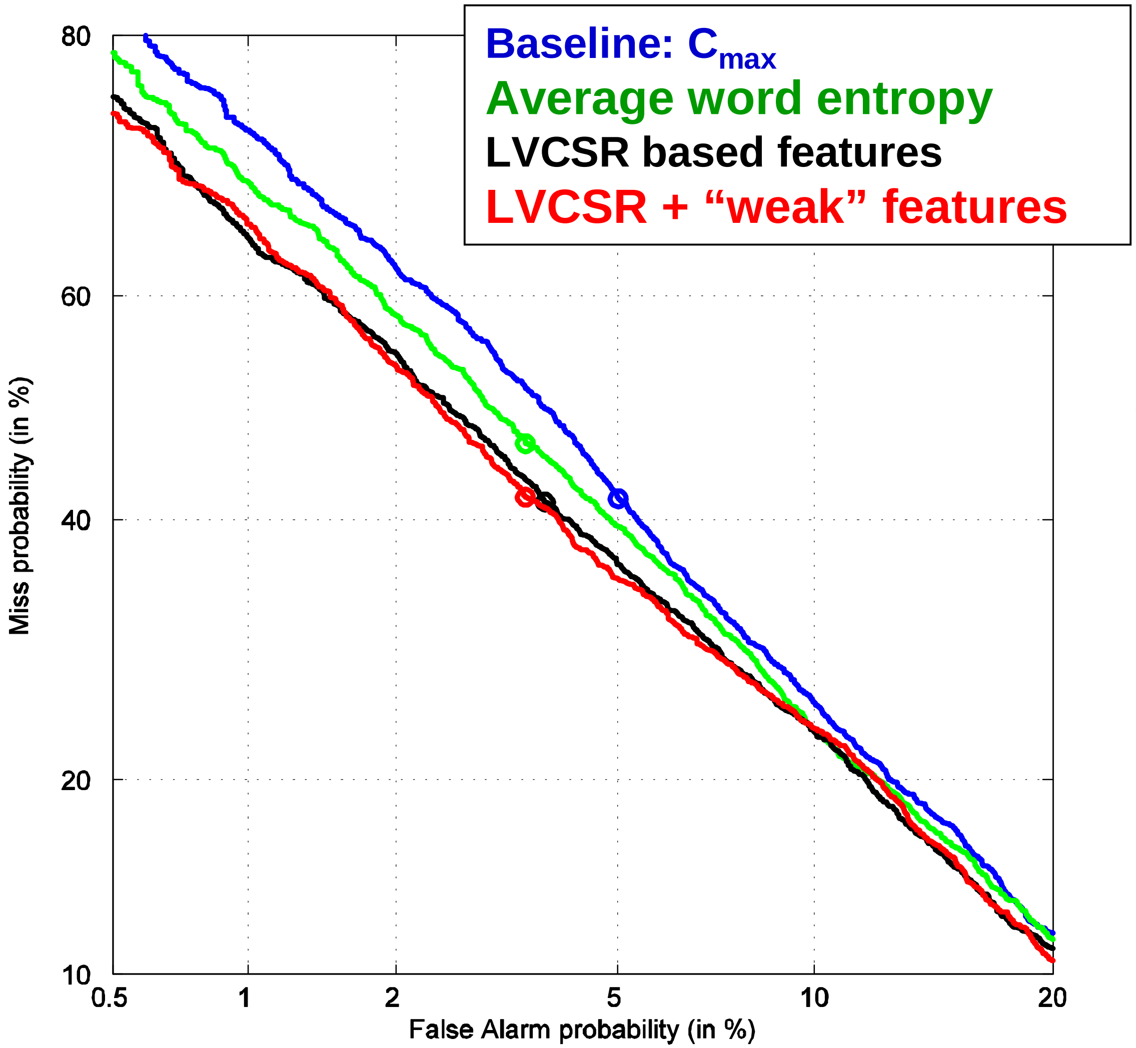


In combination, features based on weakly constrained recognizer help to improve OOV detection

OOV Detection: Search for Features



Combined System for Error Detection



In combination, features from weakly constrained recognizer don’t improve Error Detection => only useful for discrimination between OOV and other misrecognitions

9:50-10:10

Estimation of Effectiveness of Features for Classification Based On Mutual Information Measure

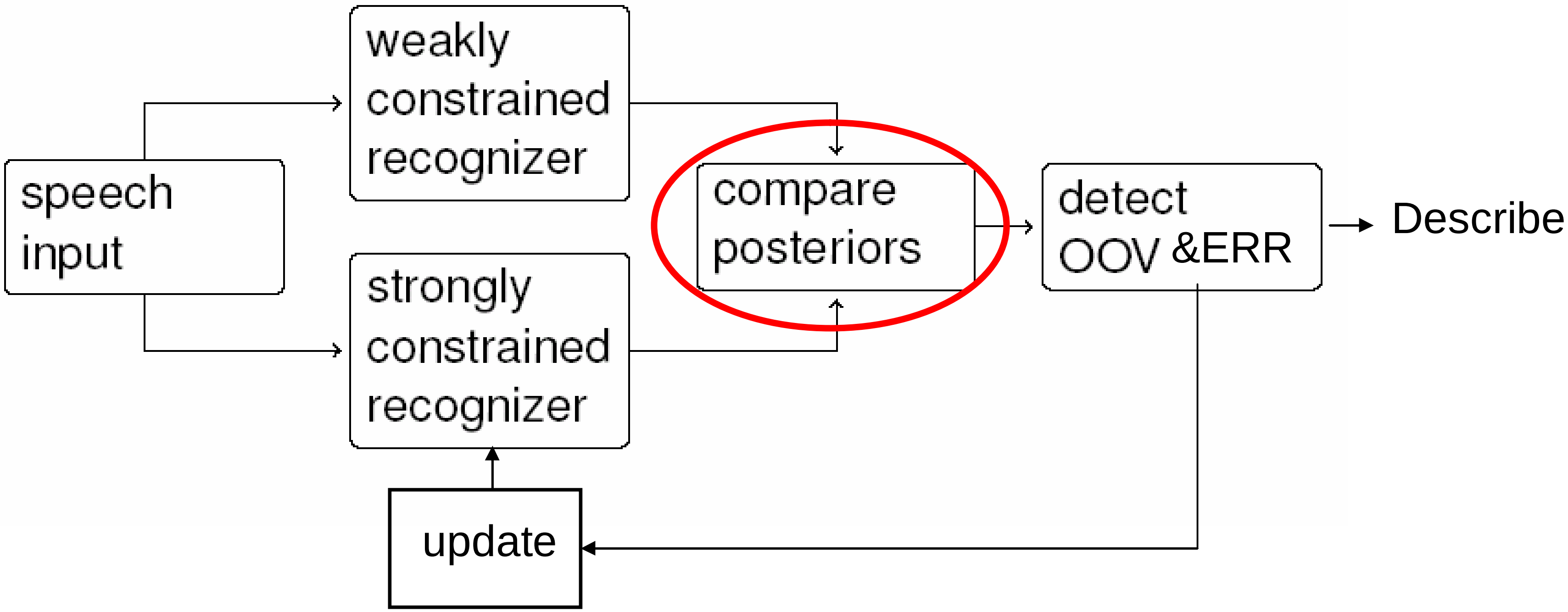
Aryia Rostow and Sanjeev Khudanpur

Where is the information for detecting errors?

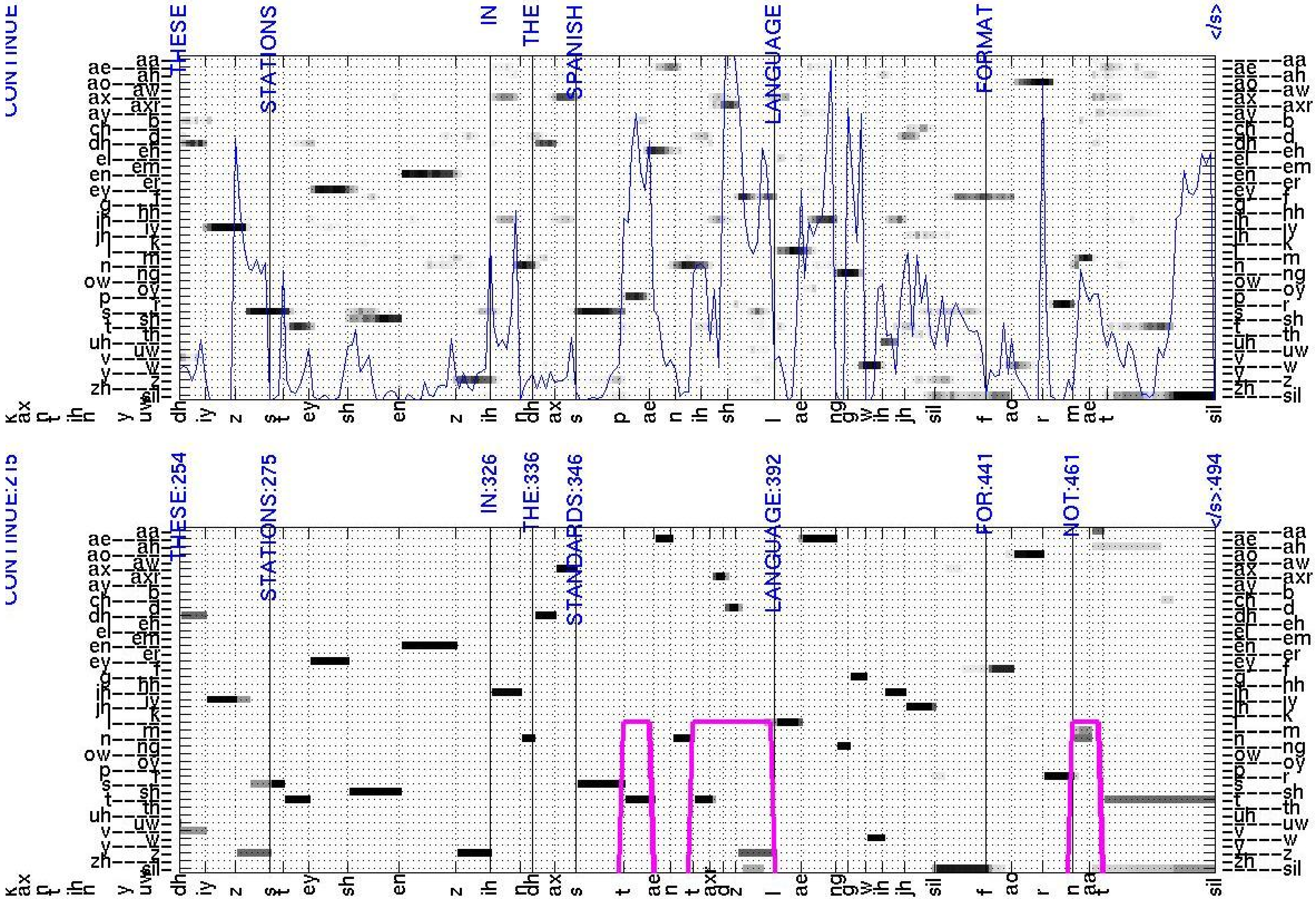
Sanjeev Khudanpur

Ariya Rastrow

Basic Idea of Project



- Is there any inconsistency
- In the case of inconsistency which one we should trust in
- Even if consistent is there a way to know if they are correct


CONTINUE
THESE
STATIONS
IN
THE
SPANISH
LANGUAGE
FORMAT
</s>
CONTINUE:215
THESE:254
STATIONS:275
IN:326
THE:336
STANDARDS:346
LANGUAGE:392
FOR:441
NOT:461
</s>:494

Mutual Information

$$I(X;Y) = h(X) - h(X \mid Y)$$
$$0 \leq I(X;Y) \leq H(Y)$$

- How can we compute the mutual-information numerically?
- Is it adequate for our task (of detecting errors/OOV)?

Numerical Issues for Computing Mutual-Information

$$h(X) = \int f(x)\log\big(f(x)\big)dx$$

$$\approx \sum f(x_i)\log\big(f(x_i)\big)\Delta(x_i)$$

Approximating integral with summation

$$= \sum p(i)\log\left(\frac{p(i)}{\Delta(i)}\right)$$

Histogram estimation of probability

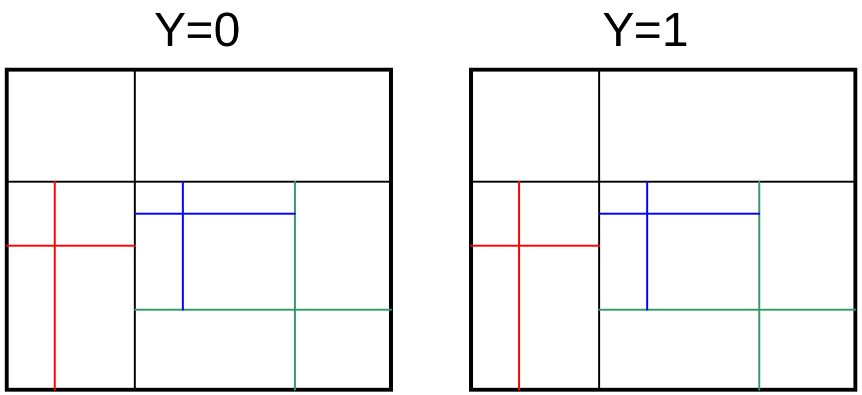
$$\approx \sum \hat{p}(i)\log\left(\frac{\hat{p}(i)}{\Delta(i)}\right)$$

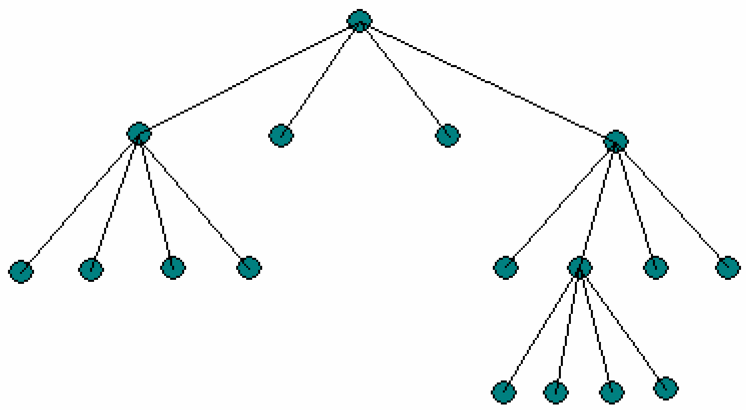
Estimation of the Information by an Adaptive Partitioning

$$I(X1,X2;Y) = D^{R}(X1,X2;Y) + D_{R}(X1,X2;Y)$$

$$\lim_{k\to\infty} D^{R^{(k)}}(X1,X2;Y) = I(X1,X2;Y)$$

$$\lim_{k\to\infty} D_{R^{(k)}}(X1,X2;Y) = 0$$



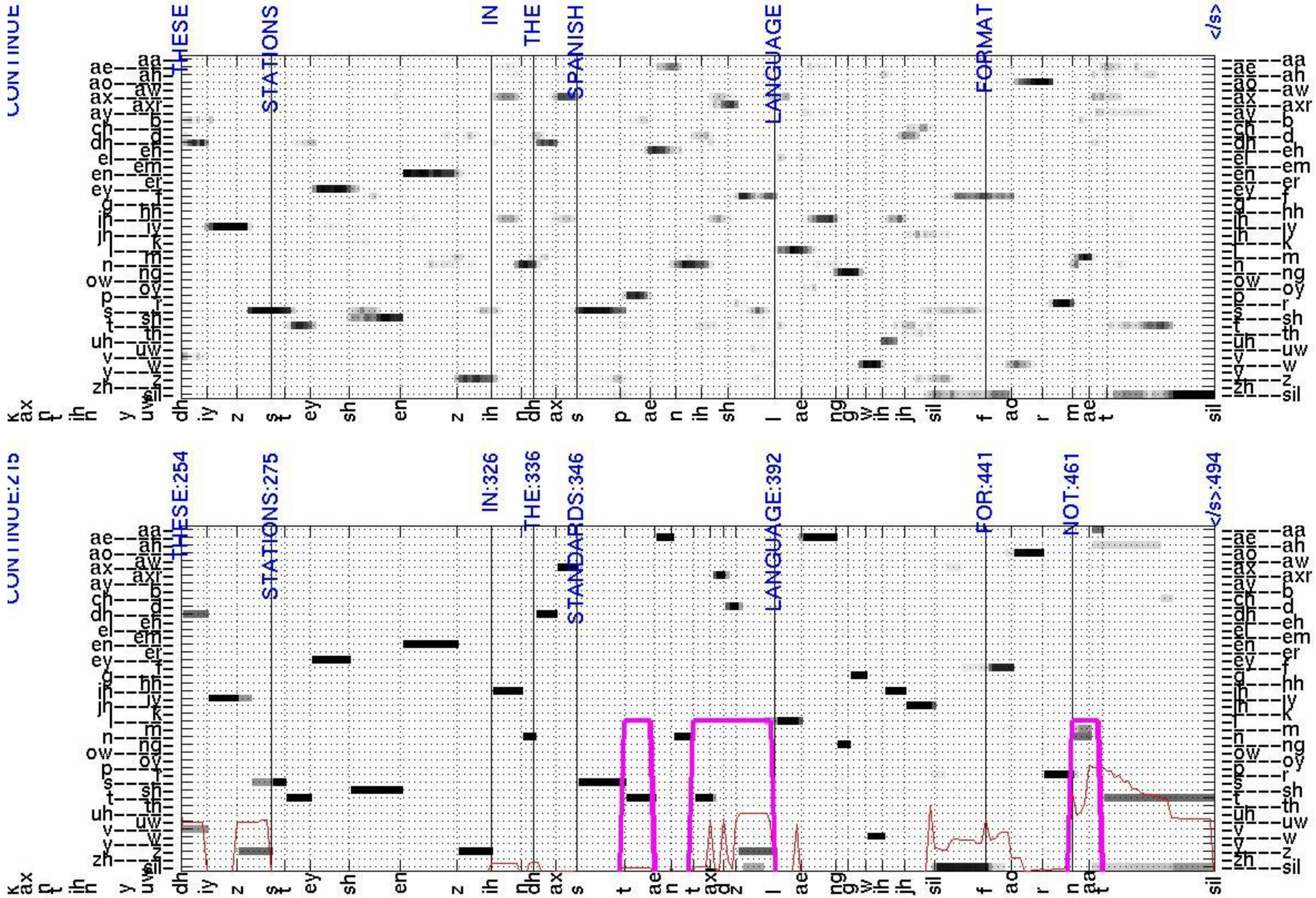


"Estimation of the Information by an Adaptive Partitioning of the Observation Space" Georges A. Darbellay and Igor Vajda

Statistics of data set

- Computing Mutual Information on test data
- 75081 correct phones,4308 incorrect phones
- à p(Y=0) = 0.946 and p(Y=1)=0.054

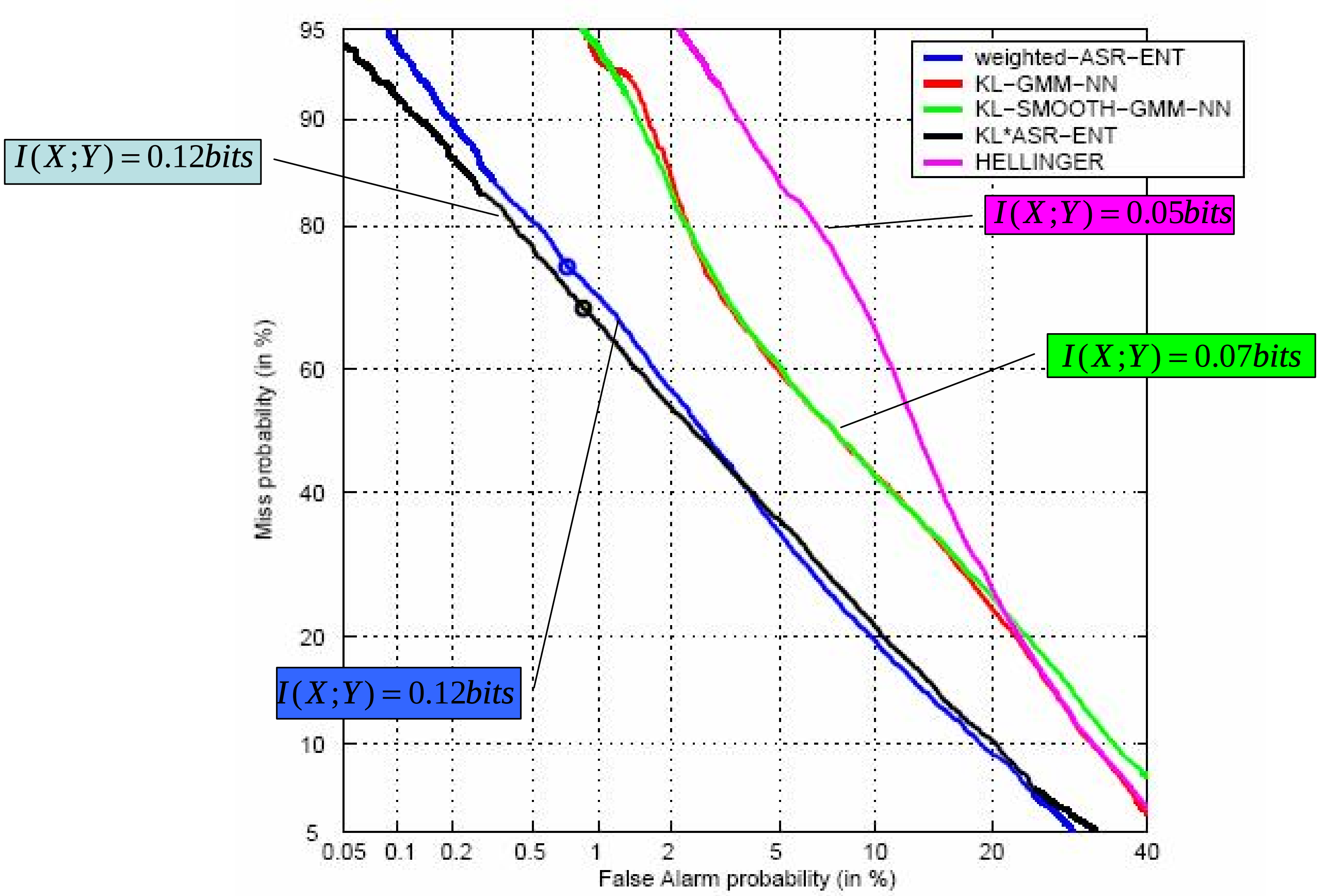
$$\Rightarrow H(Y) = 0.305 \quad bits$$



$$I(X;Y) = 0.12\,bits$$

$$H(Y) = 0.305\ bits$$

DET curves on phone level



Mutual Information using more than one divergence

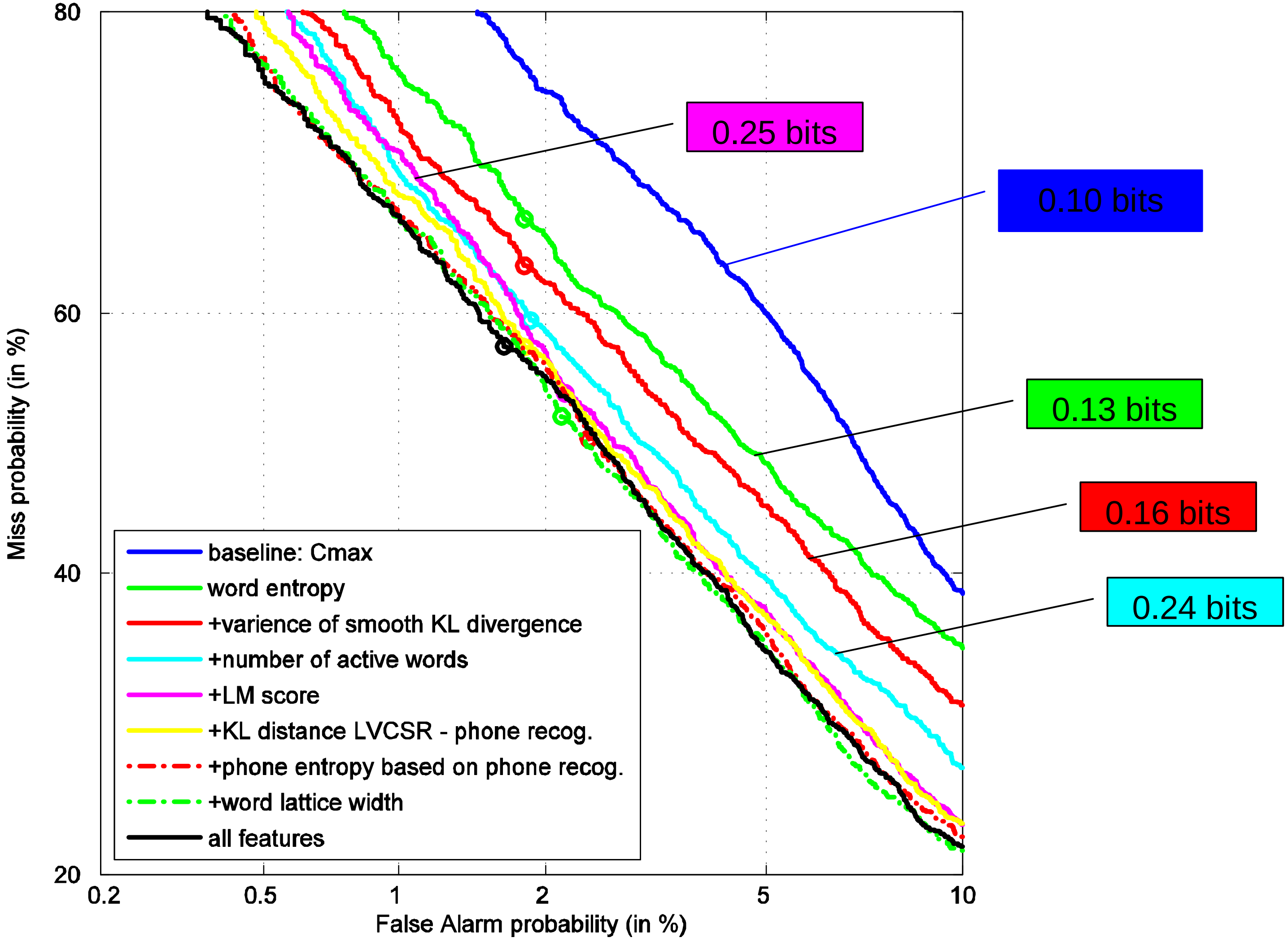
Mutual Information

$$I(X1, X2; Y) = h(X1, X2) - h(X1, X2 \mid Y)$$

$$0 \leq I(X1, X2; Y) \leq H(Y)$$

$$I(X1, X2; Y) = I(X1; Y) + I(X2; Y \mid X1) \mathrel{\substack{\geq \\ \leq}} I(X1; Y) + I(X2; Y)$$

OOV Detection: Search for Features

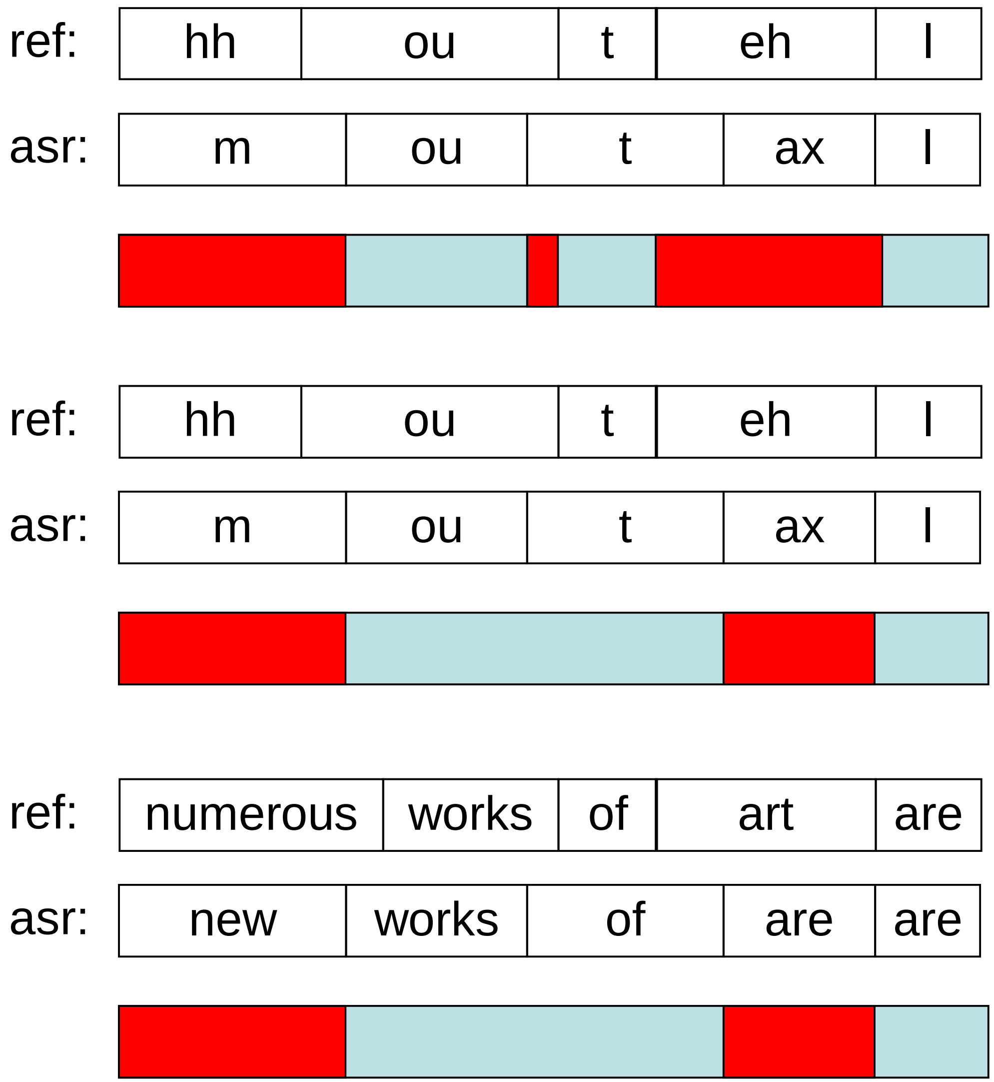


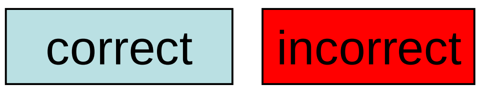
Comparison of Strongly and Weakly Constrained Phoneme Posterior Streams Using Trained Artificial Neural Net

Pavel Matejka

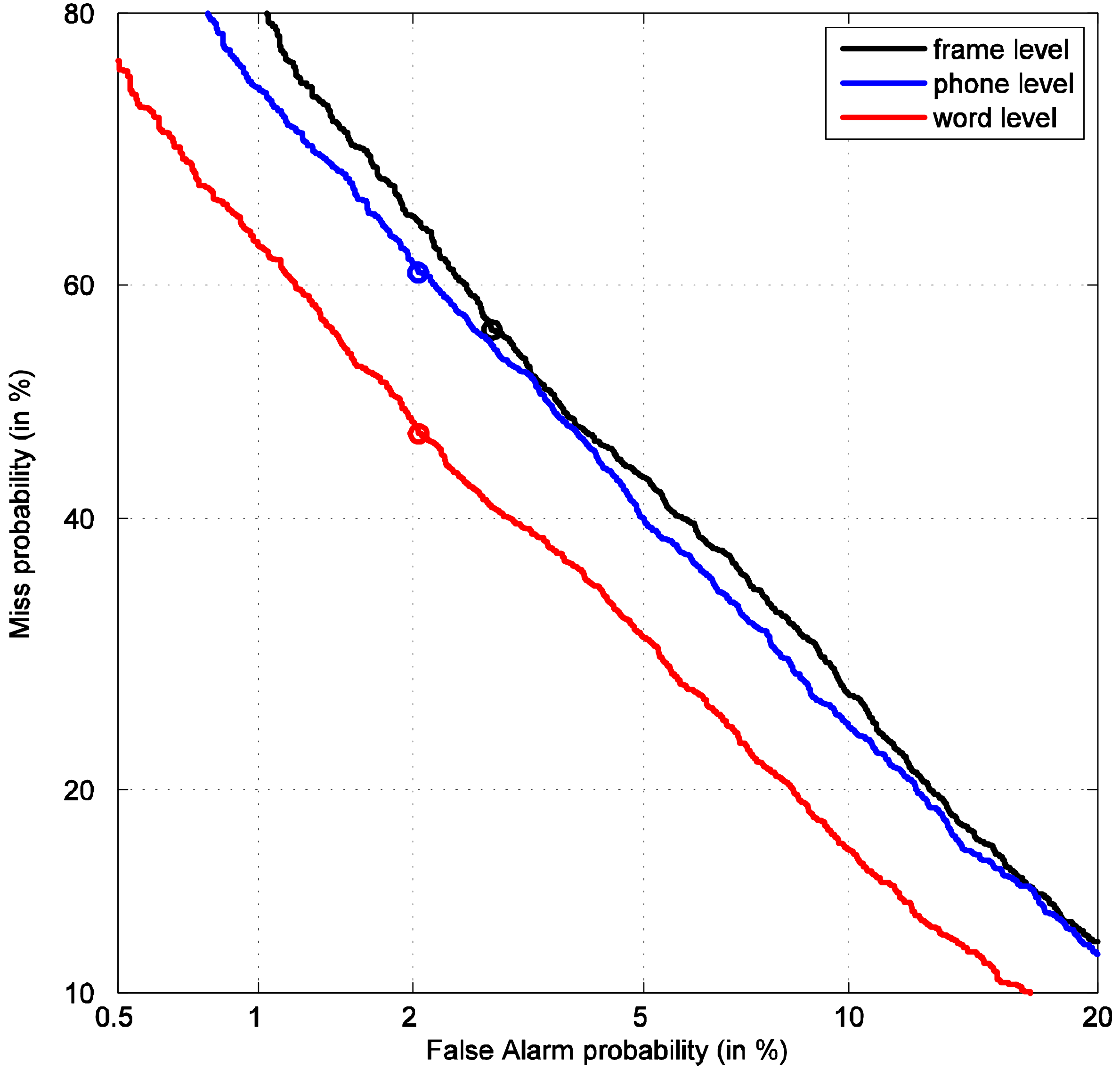
Training the Neural Network

- Frame level
- Phoneme level
- Word level

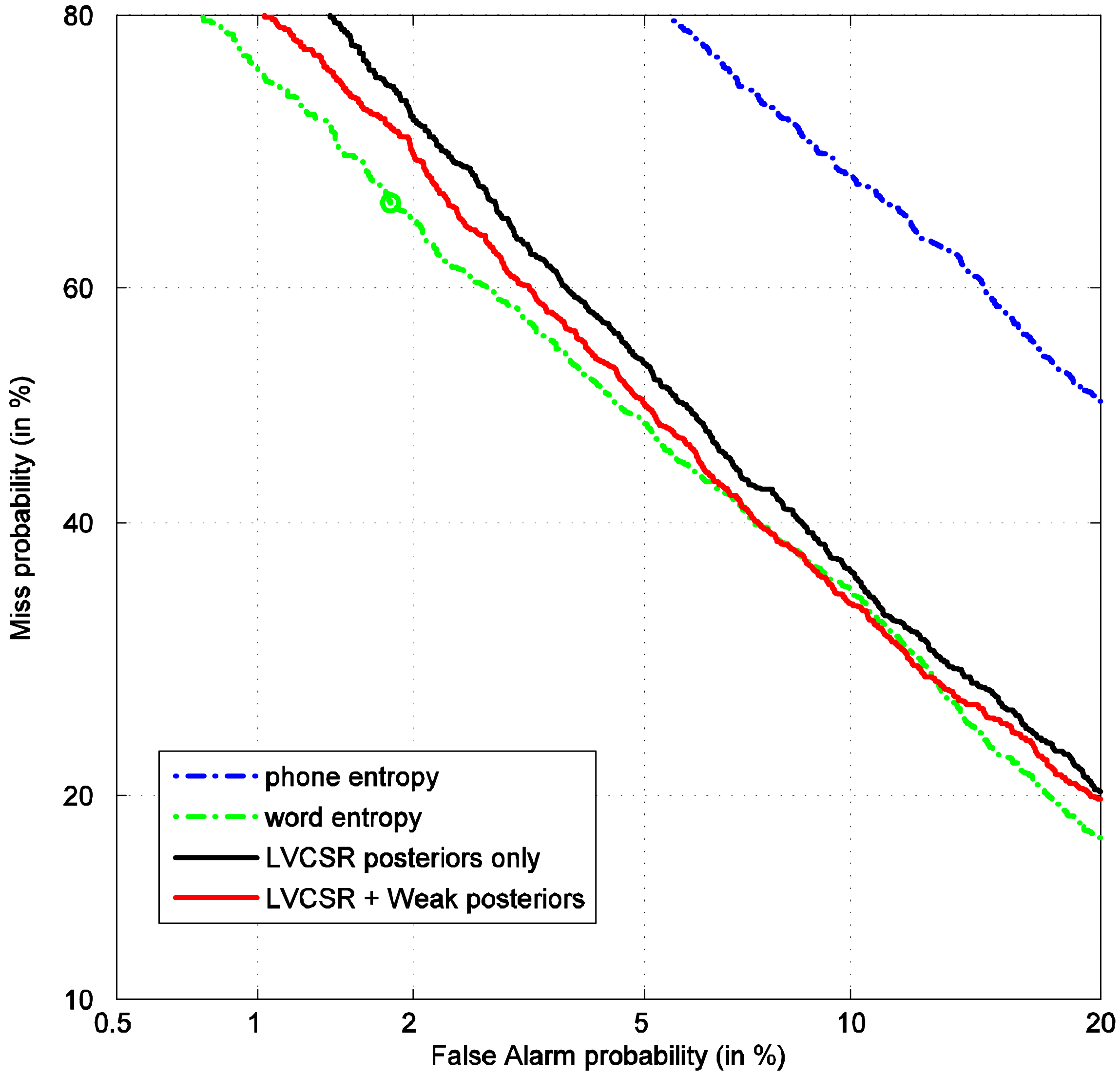




Training on the Frame, Phone, and Word Level

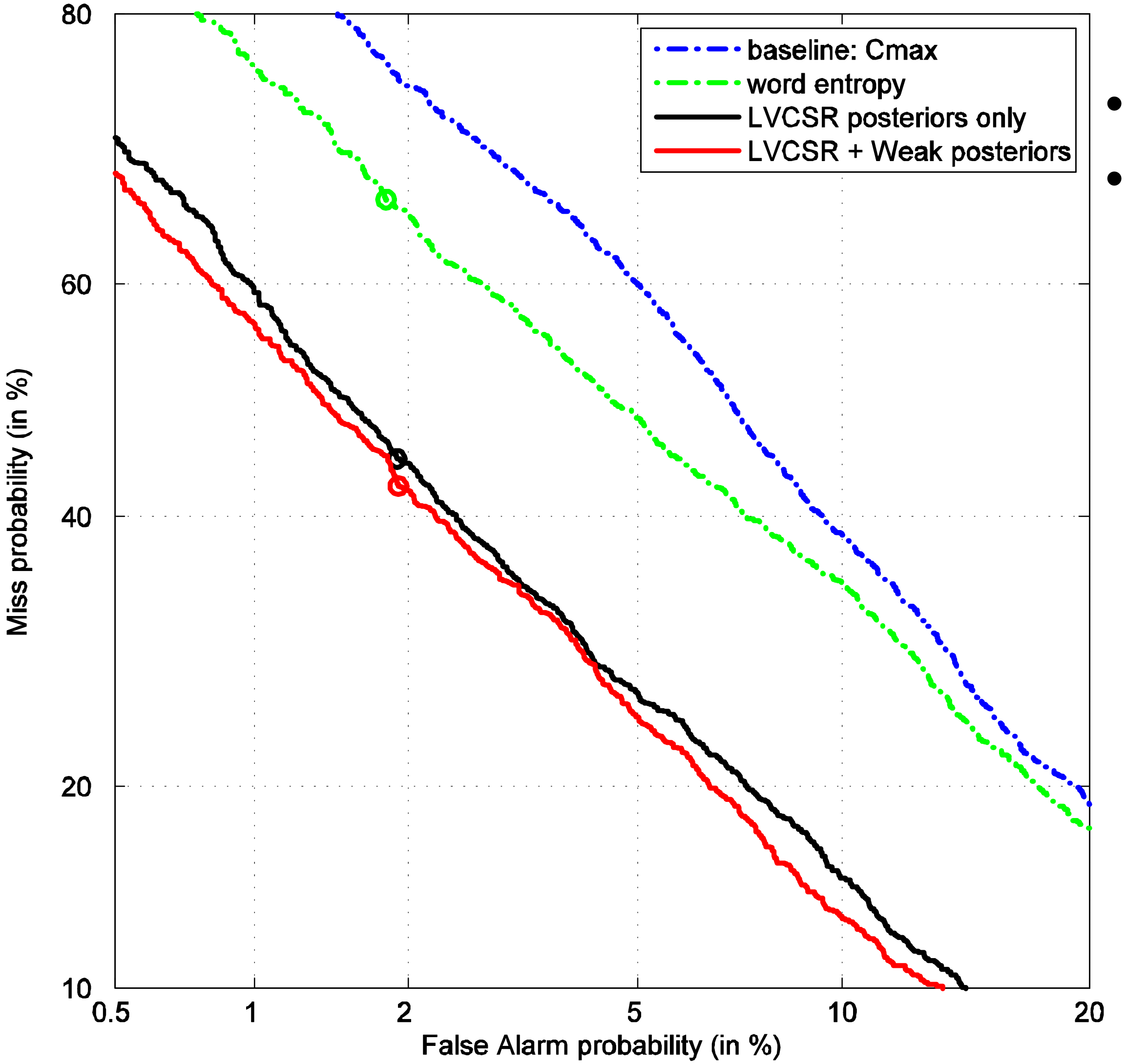


Context of 0 Frames (1 Frame Only)



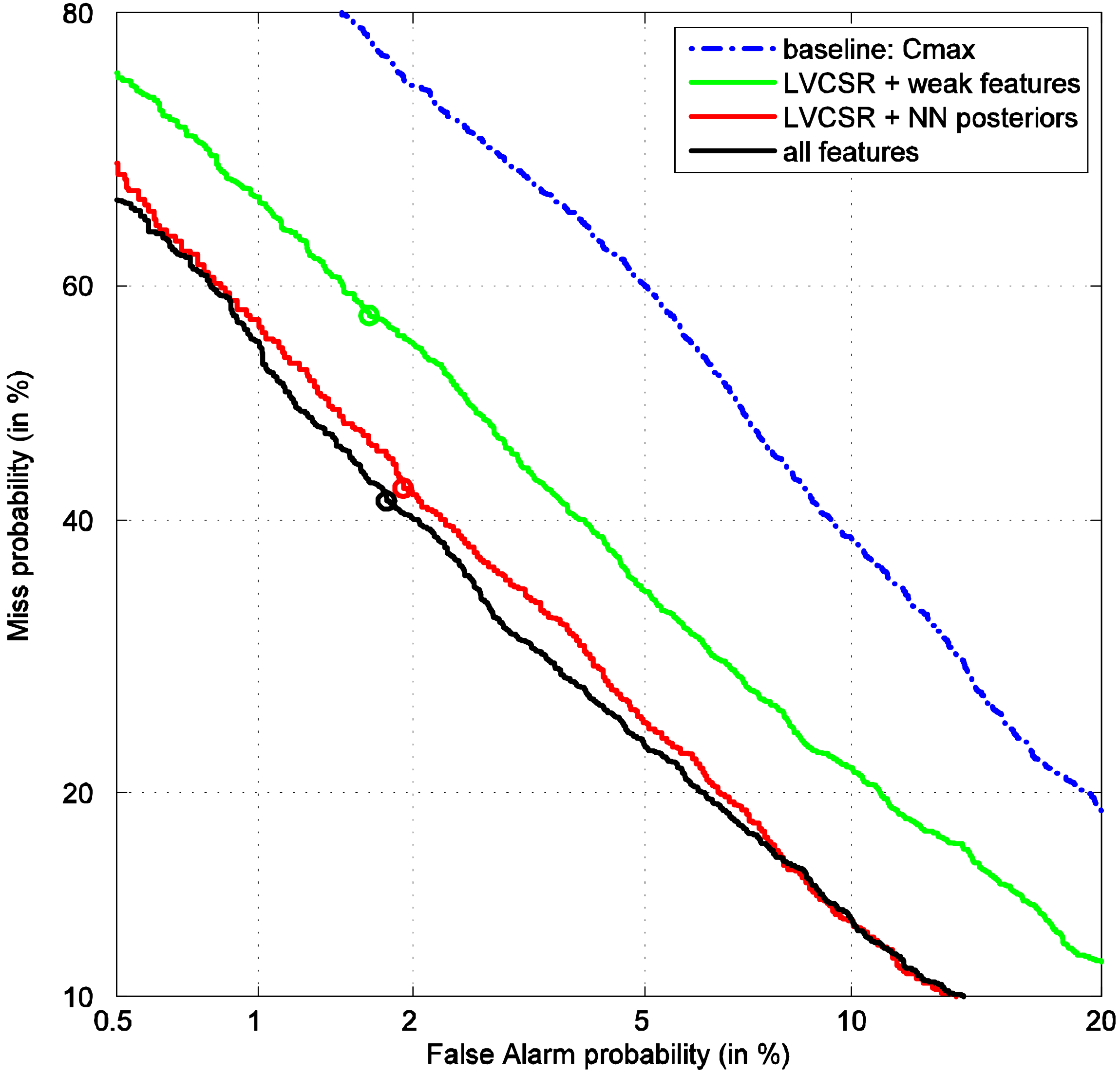
- Performance competitive with word entropy from an LVCSR lattice

Context of 1 Frames (3 Frames Total)



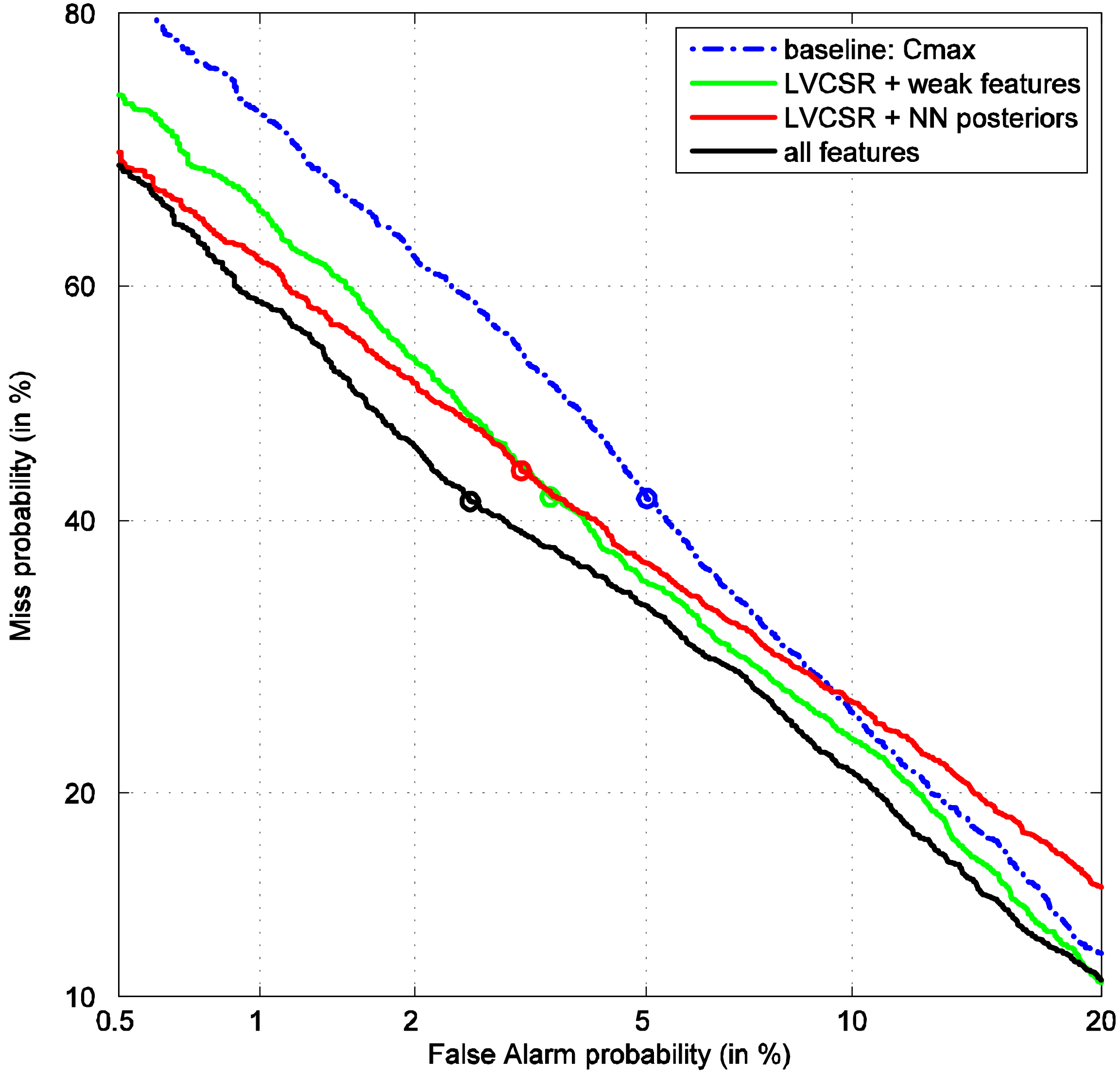
- Best Single System
- Uses downsampled input
  - e.g. 1st and 13th frame together with center frame
  - 130 ms window

Comparison to the rest of the features



- Significantly better than baseline and LVCSR + weak features

Performance for Detection of Errors



- Same performance as all LVCSR + weak features merged together

Neural Network Based Detection Conclusions

- Best single performing system
- Substantial improvement over state of the art OOV detection
- Can also be used to detect errors comparable to word entropy
- Takes advantage of the entropy in the strong recognizer posterior distribution, however adding the weak recognizer's posterior distribution consistently improves performance in OOV and Error detection

Normalization of scores using counts of phonemes

- Optimal threshold for word detection/rejection can be seen as linear combinations of contributions of different phonemes

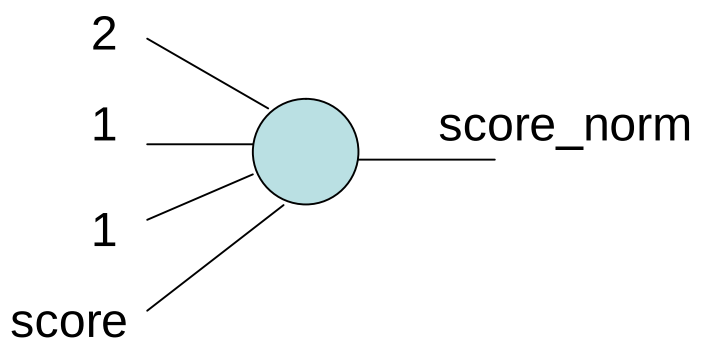
- For example word “total”:

$$2\,t + 1\,ow + 1\,el + c = opt\_thr$$

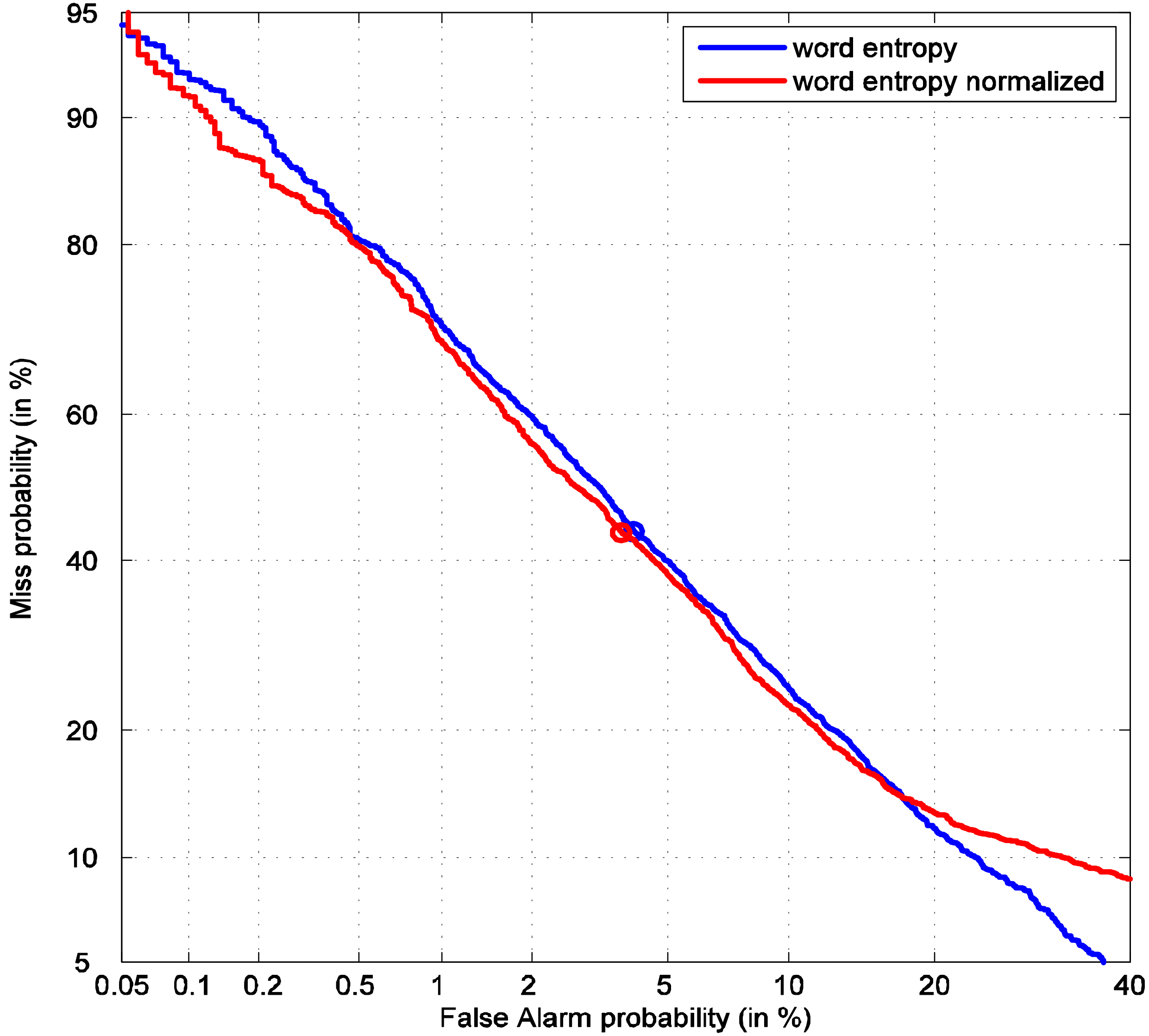
- The word based threshold can be subtracted from scores:

$$score\_norm = score - (2\,t + 1\,ow + 1\,el + c)$$

- Then a SVM can be trained for detection / rejection task to get variables



Normalization of scores using counts of phonemes



10 minute BREAK for Schmoozing

Everybody

Phoneme Recognizers and Transducers in Detection of Out-Of-Vocabulary Words and Recognizer Confidence Estimation

Sally Isaacoff, Jon Nedel, Chris White and Geoff Zweig

Using Phone-to-Word Transduction and Phone-Level Comparisons for Detection of ASR Errors and OOVs

JHU Summer Workshop 2007

"Whazwrong" Team

Sally Isaacoff, Jon Nedel, Chris White, Geoffrey Zweig

Preview

- Phone-to-Word Transduction Method (Geoff Zweig)
- Transduction from Universal Phones to Words for Universal Word Decoding
- Transduction for Word Confidence Estimation and OOV Detection

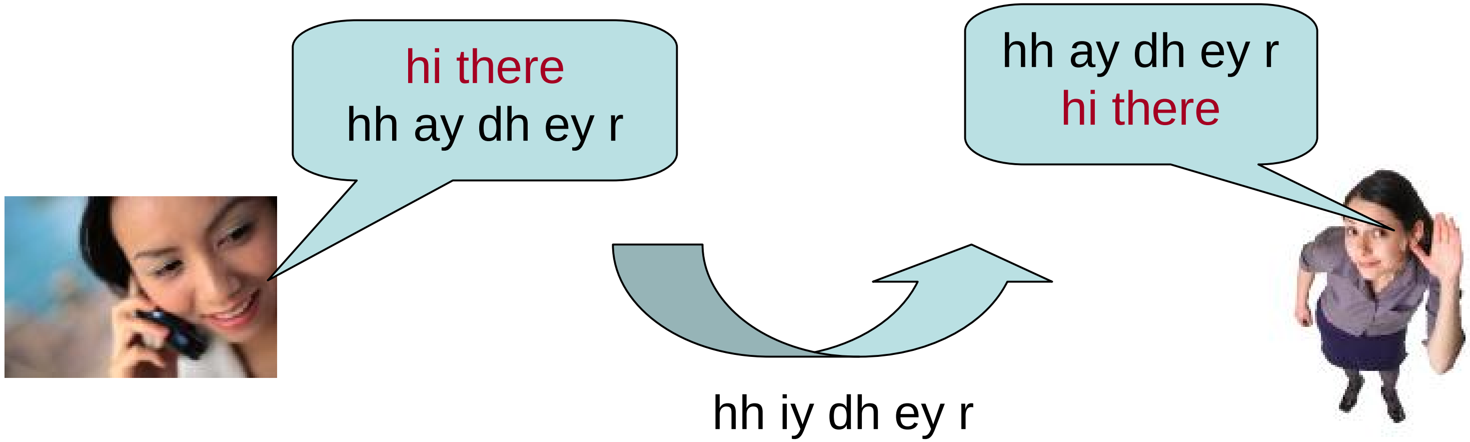
Phone-to-Word Transduction Method for ASR (Geoff Zweig)

- Break the decoding process into two steps:
  1. Get the phones
  2. Turn the phones into words

- Apply this to decoding with a Universal phone set and acoustic model:
  – If successful, we can use 1 acoustic model across languages and still recover words!

- Apply this to OOV detection:
  – Take the phone string from a weakly-constrained recognizer
  – Transduce it to words
  – See where the words differ from the strongly-constrained recognizer

The Transduction Picture



Phone sequence corrupted by noise;
LM, error model allow recovery

Transducer Formulation

W: intended words (unknown)
l_i: intended phones/letters (unknown)
l_c: corrupted phones/letters (observed)
Want to find the likeliest word and phone sequence underlying the observations

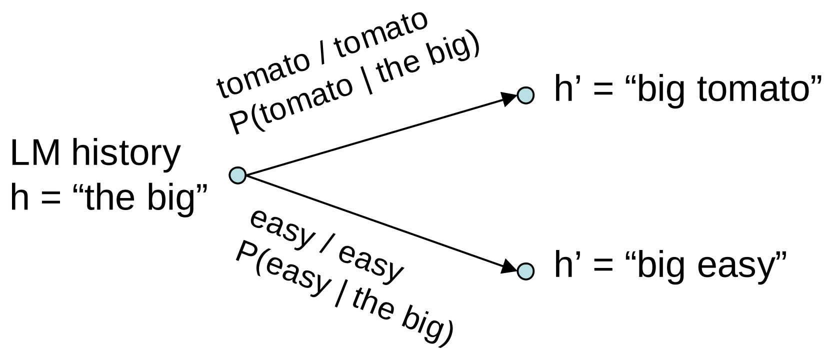
$$\arg\max_{w,l_i} P(w, l_i \mid l_c) = \arg\max_{w,l_i} \frac{P(w, l_i) P(l_c \mid w, l_i)}{P(l_c)}$$

$$= \arg\max_{w,l_i} P(w) P(l_i \mid w) P(l_c \mid w, l_i)$$

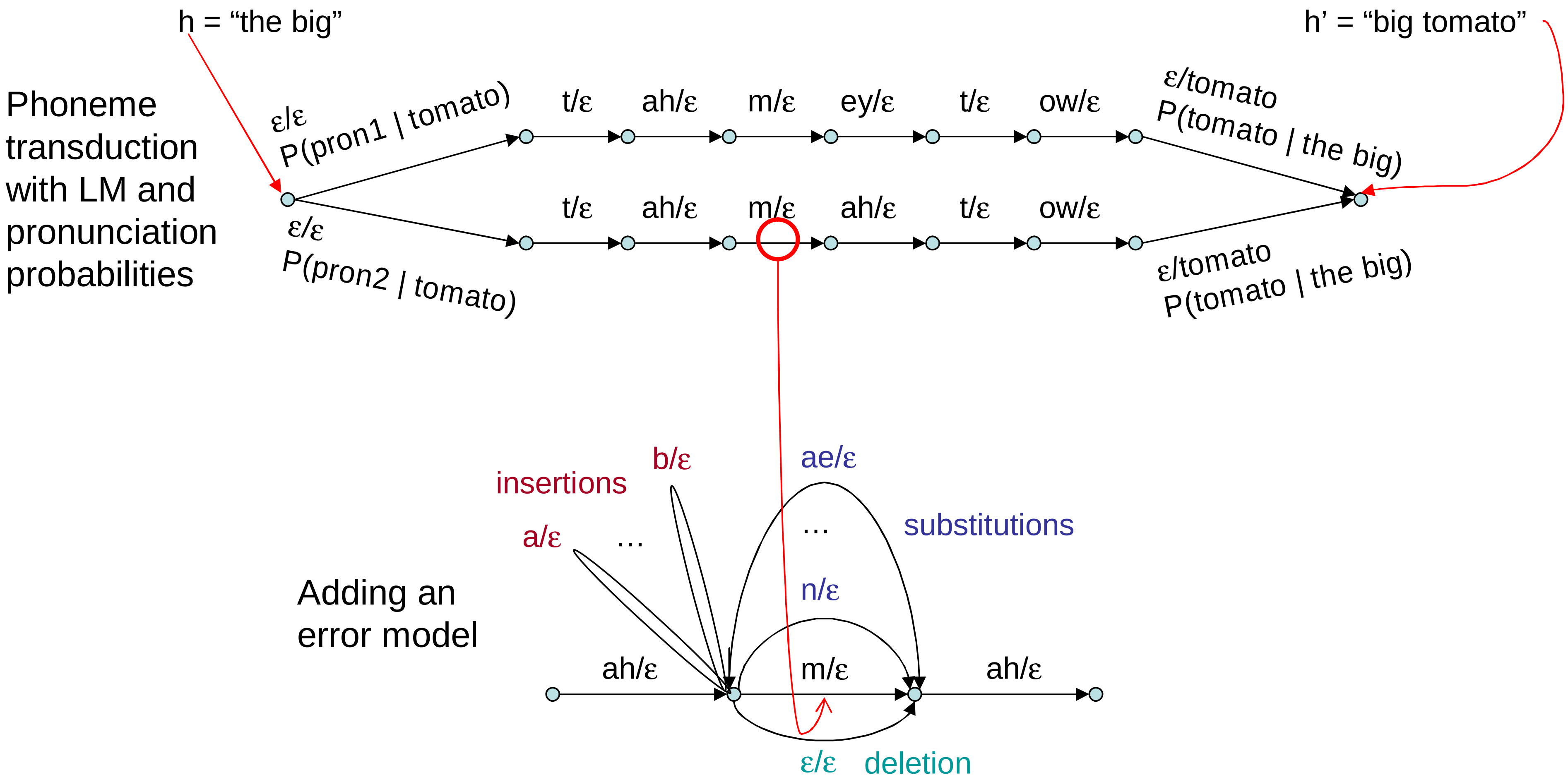
$$= \arg\max_{w,l_i} \underbrace{P(w)}_{\text{Language model}} \underbrace{P(l_i \mid w)}_{\text{Pronunciation probabilities}} \underbrace{P(l_c \mid l_i)}_{\text{Error model}}$$

Transducer Representation

Encoding a language model



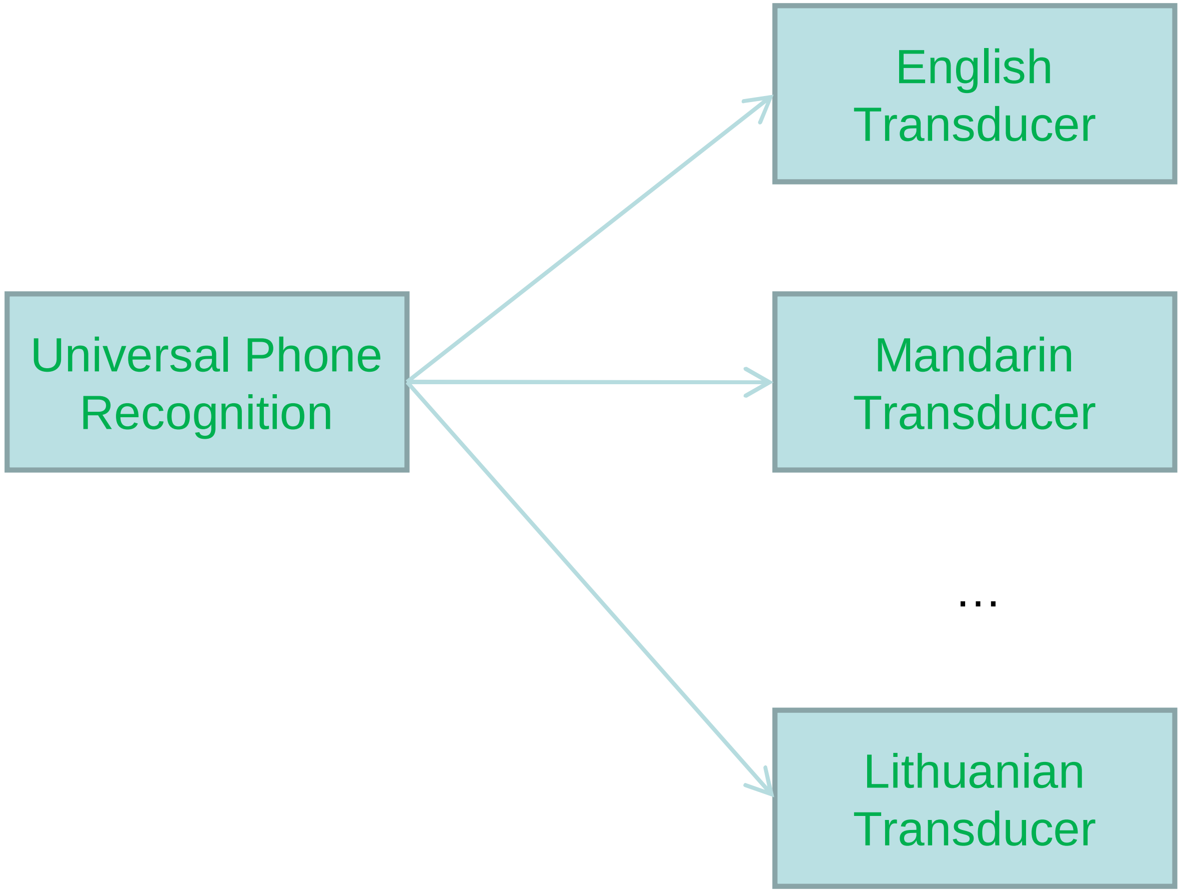
Transducer Representation



Transduction to Words from Universal Phone Decoding

- Currently, deploying speech recognition to a new language requires building a new acoustic model
- This is time-consuming, complicated, trouble-prone, expensive and brain-power intensive
- Is it possible to use a single phone-level acoustic model across languages?
- To deploy in a new language, could we just work on the purely symbolic phone-to-word transduction?

Desired System Structure



Universal Phone Recognition (UPR)

- System developed by Dr. Patrick Shone
- Uses an International Phonetic Alphabet (IPA) based universal phone set
- Core software: HTK, Version 3.3
- Trained with CallHome, Switchboard, OGI data

- Acoustic Models:
  - 3-state HMM with observables modeled by Gaussian Mixture Models, 17 mixtures per state
  - Basic units: monophones (faster), or right-context diphones w/ allophony (slower)

h ei p^{h} a l w ə z h æ p n ɪ n

⇩

- h $_{h}$[ei] $_{ei}$[p^{h}] $_{p^{h}}$[a] $_{a}$[l] $_{l}$[w] $_{w}$[ə] $_{ə}$[z] $_{z}$[h] $_{h}$[æ] $_{æ}$[p] $_{p}$[n] $_{n}$[ɪ] $_{ɪ}$[n]
  - Bigrams of phone sequences

- System can be run in language-specific mode or in universal mode

The CallHome Database

- English, Mandarin, Egyptian Arabic, German, Spanish, Japanese
- Native speakers paid to call home and chat
- ~10 hours of audio in each language
- Transcriptions provided
- Lexicons provided

CallHome database vital stats

	# Training words	# Test words	Lex. Size (prons)	OOV rate
Egyptian	149k	33k	57k	1.6%
German	165	43	315	1.1
English	167	43	99	1.8
Mandarin	159	42	44	3.1
Spanish	145	38	45	2.6
Japanese	154	39	80	18.5

Japanese is unusable due to high OOV rate

Results: Phone-to-Word Transduction from UPR Output on CallHome

	PER – Universal AM (monophone)	WER – Universal AM (monophone)	PER – Language specific AM (diphone)	WER – Language specific AM (diphone)	WER – reference phones
Egyptian	86.9	99.0	60.8	82.8	0.6
German	86.1	94.9	63.0	85.1	0.9
English	82.8	94.8	56.4	76.6	2.3
Spanish	88.8	97.0	56.7	79.7	5.0
Mandarin*	85.3	98.3	62.8	79.0	8.9

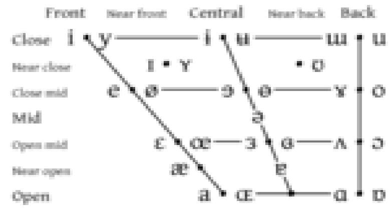
* Mandarin results present CER rather than WER.

Mandarin WER is high because UPR has no tone information

Conclusions: UPR-Based Transduction

- A 2-step process is possible
  - Shown effective across 5 CallHome languages
- The UPR phone set is quite reasonable
  - Shown effective across 5 CallHome languages
- The universal-mode UPR AM needs work
  - 80% PER too high
- The bottleneck is getting the phones right
  - 3–5% WER if the phones are right

Lexicons: LDC → IPA



- For each CallHome language, the Linguistic Data Consortium (LDC) developed a pronunciation lexicon
- Phone set and lexicon conventions are language specific
- Similar sounds in different languages have different LDC symbols:

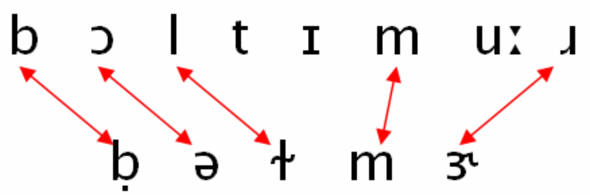
International Phonetic Alphabet (IPA): ʧ

LDC:	English “C”	Spanish “c”	German “tS”
	“chair”	“chico”	“Deutsch”

- Developed an LDC → IPA mapping for each CallHome language
- IPA-based Lexicons for the 6 CallHome languages are now available
  - Enables comparison of phone recognition accuracy across languages
  - Can be used to train multi-lingual phone recognizers with LDC data (*e.g.* Brno can now train a state-of-the-art multi-lingual phone recognizer after the workshop)

Linguistic Similarity Metric

- Vowel features:
  - height, backness, pulmonic, nasal, length, rounding, rhotic, palatalized
- Consonant features:
  - pulmonic, length, aspiration, alternate airstream, rhotic, nasal, plosive, click, fricative, approximant, lateral, apical, rounding, labial, coronal, dorsal, palatalized, velarized, radical, glottal
- Idea:
  - Develop a linguistically-motivated “distance” between any two IPA phones (*e.g.* Hamming distance in Withgott & Chen phone space)
- Applications:
  - Alignment of hypothesized phones to reference phones based on linguistic similarity, rather than a simple exact string match
    - “Baltimore”



  - Better-trained error models for phone-to-word transduction?
  - Error model for phone-to-word transduction in languages with no available training data

Transduction for OOV Detection and Confidence Estimation

- Intuition:
  - Transducer makes it possible to compare word-level output of strongly- and weakly-constrained recognizers
    - Even though the weakly-constrained recognizer has no word level output
  - If the two agree on a word, we are more confident it is right
  - If the two disagree, we can measure how much by comparison of phone streams

Transduction for OOV/Error Detection

REF: numerous works of art are based on the story

ASR: new morris works are part are based on the story

```
ASR:nUmqriswRksGpartarbYstanDJstqrI
HMM:nUmRswRksJvQrtQrbYstanDJstqrI
```

Transduction for OOV/Error Detection

REF: numerous works of art are based on the story

ASR: new morris works are part are based on the story

`ASR:nUmqriswRksGpartarbYstanDJstqrI`
`HMM:nUmRswRksJvQrtQrbYstanDJstqrI`

TD: new works that are based on the story

Transduction for OOV/Error Detection

REF: numerous works of art are based on the story

out of vocabulary

ASR: new morris works are part are based on the story

ASR error

ASR:nUmqriswRksGpartarbYstanDJstqrI

HMM:nUmRswRksJvQrtQrbYstanDJstqrI

TD: new works that are based on the story

Transduction for OOV/Error Detection

REF: numerous works of art are based on the story

```
ASR: new MORRIS works ARE PART are based on the story
 TD: new ****** works *** THAT are based on the story
```

Transduction for OOV/Error Detection

REF: numerous works of art are based on the story

```
 ASR: new MORRIS works ARE PART are based on the story
  TD: new ****** works *** THAT are based on the story
```

```
ASR: nUmqriswRksGpartarbYstanDJstqrI
 TD: nUwRksDAtarbYstanDJstqrI
HMM: nUmRswRksJvQrtQrbYstanDJstqrI
```

Transduction for OOV/Error Detection

REF: numerous works of art are based on the story

```
ASR: new MORRIS works ARE PART are based on the story
 TD: new ****** works *** THAT are based on the story
```

```
ASR: nUmqriswRksGpartarbYstanDJstqrI
 TD: nU-----wRksD-A-tarbYstanDJstqrI
HMM: nUmR--swRksJvQrtQrbYstanDJstqrI
```

Multi-string alignment

Transduction for OOV/Error Detection

REF: numerous works of art are based on the story

```
 ASR: new MORRIS works ARE PART are based on the story
  TD: new ****** works *** THAT are based on the story


 ASR: nUmqriswRksGpartarbYstanDJstqrI
  TD: nU-----wRksD-A-tarbYstanDJstqrI
 HMM: nUmR--swRksJvQrtQrbYstanDJstqrI


ASR: nU   mqris   wRks   G   part
 TD: nU   -----   wRks   D   -A-t
HMM: nU   mR--s   wRks   J   vQrt
```

ASR word boundary

Transduction for OOV/Error Detection

REF: numerous works of art are based on the story

```
ASR: new MORRIS works ARE PART are based on the story
 TD: new ****** works *** THAT are based on the story
```

```
ASR: nUmqriswRksGpartarbYstanDJstqrI
 TD: nU-----wRksD-A-tarbYstanDJstqrI
GMM: nUmR--swRksJvQrtQrbYstanDJstqrI
```

deletion+=1

delete p cost: 1.346

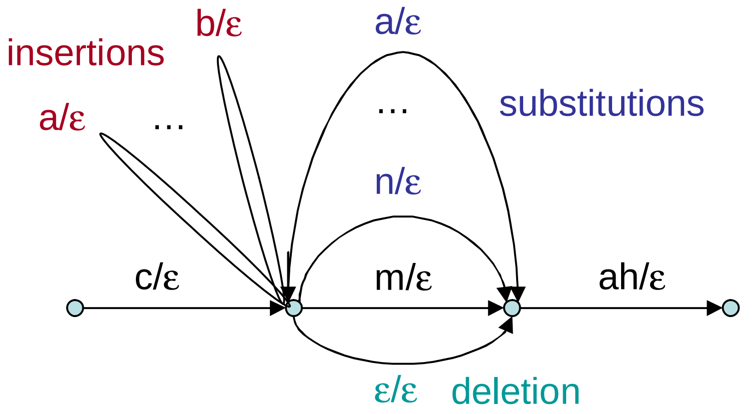
```
ASR: nU   mqris   wRks   G   part
 TD: nU   -----   wRks   D   -A-t
GMM: nU   mR--s   wRks   J   vQrt
```

sub J and D cost: 2.535

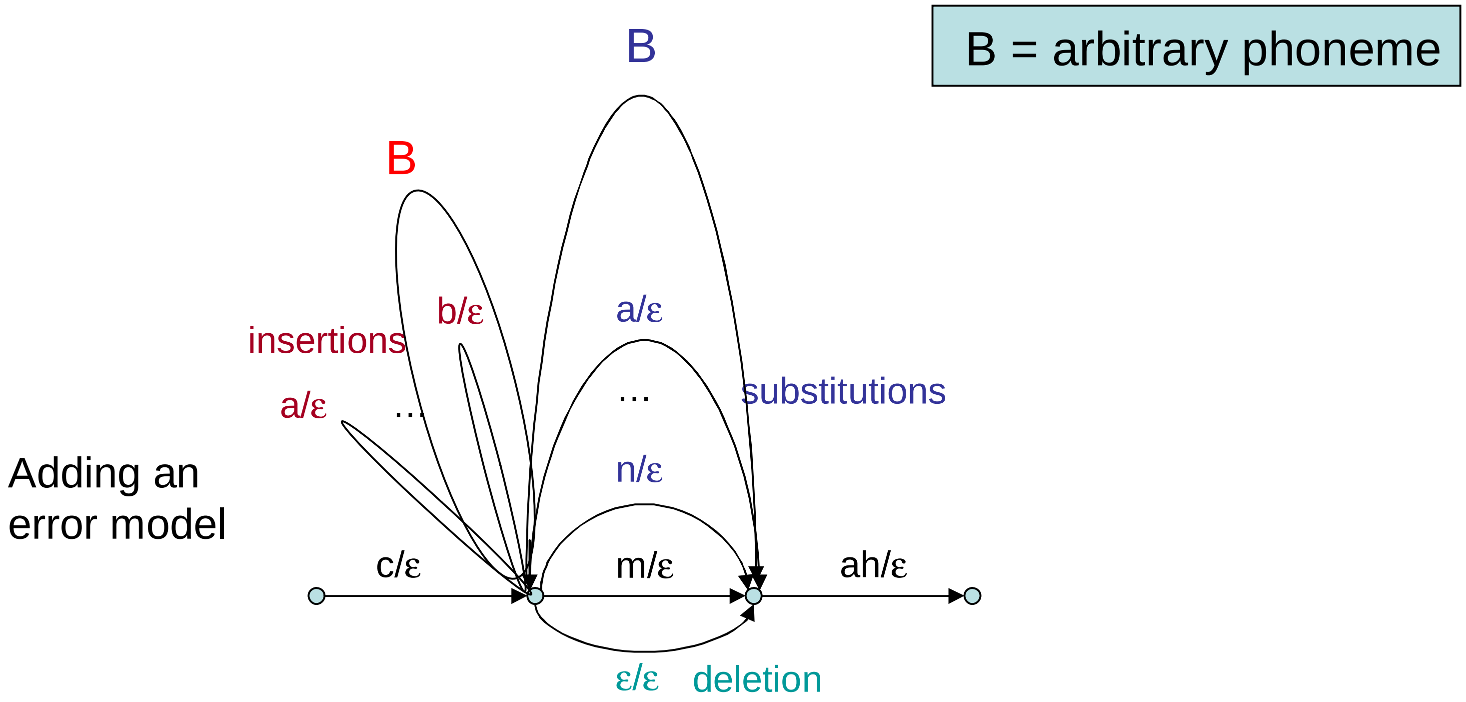
substitution+=1

Transducer Representation: Error Model

Adding an
error model



Adding a new phoneme

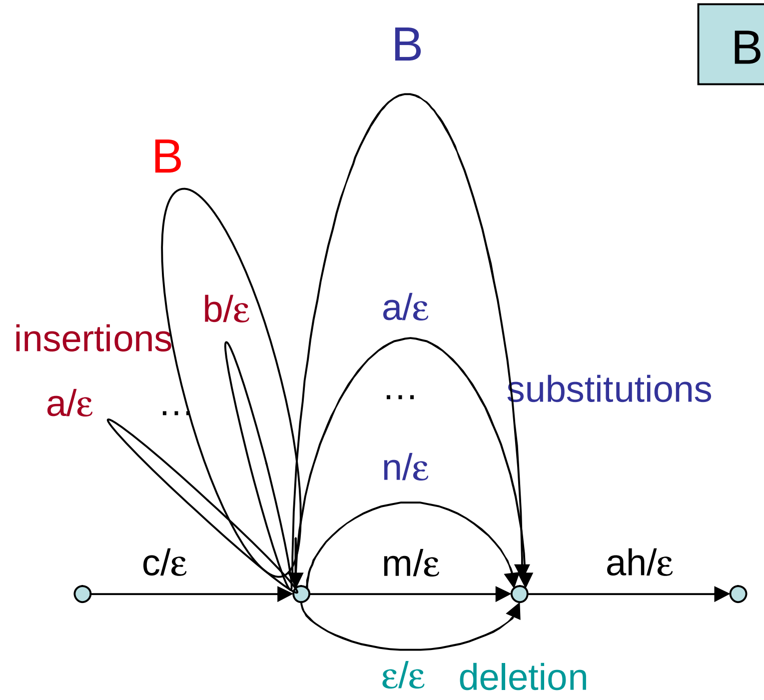


Augmenting the Dictionary

UNKNOWN(1) B
UNKNOWN(2) B B
UNKNOWN(3) B B B
UNKNOWN(4) B B B B
UNKNOWN(5) B B B B B
UNKNOWN(6) B B B B B B
.
.

B = arbitrary phoneme

Adding an
error model



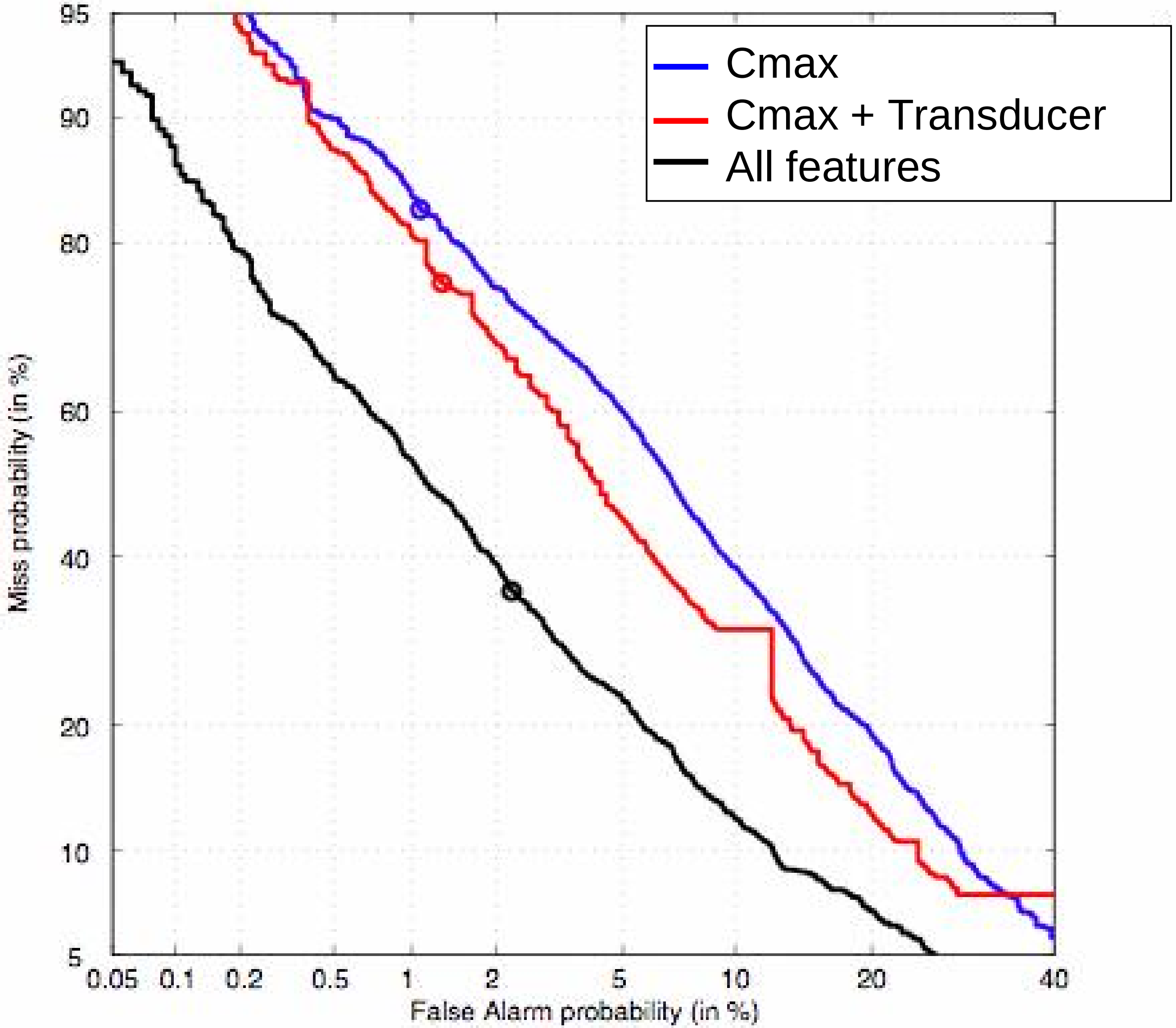
Transduction for OOV/Error Detection

REF: numerous works of art are based on the story

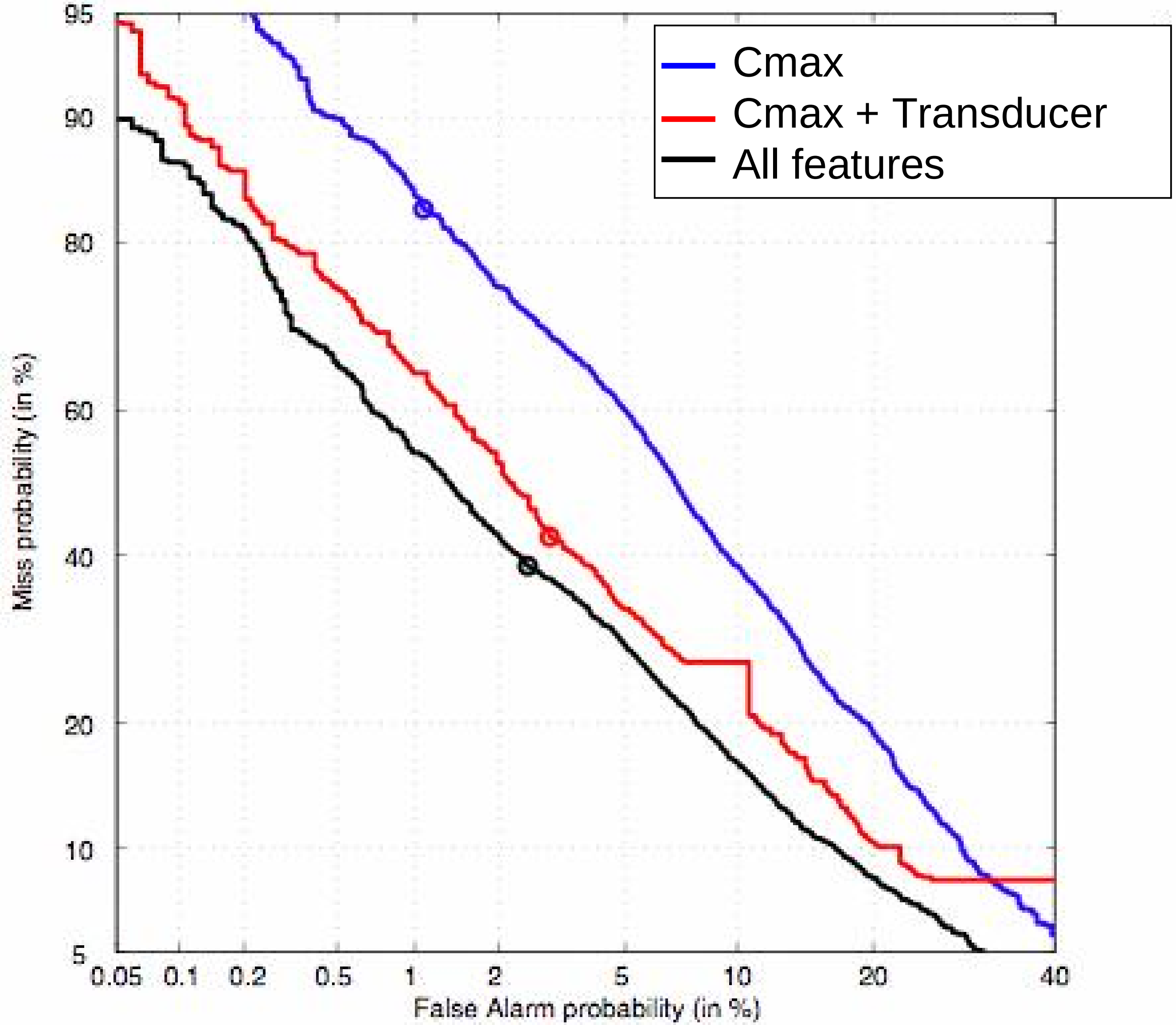
```
ASR: new MORRIS   works ** ARE PART     are based on the story
 TD: new UNKNOWN  works OF OUR UNKNOWN are based on the story
```

```
 ASR: nUmqriswRksGpart--arbYstanDJstqrI
  TD: nUBB---wRksJvQrBBBarbYstanDJstqrI
 HMM: nUmR--swRksJvQrt--QrbYstanDJstqrI
```

OOV Detection DET Curve



Error Detection DET Curve



What if we had a perfect phonetic transcription?

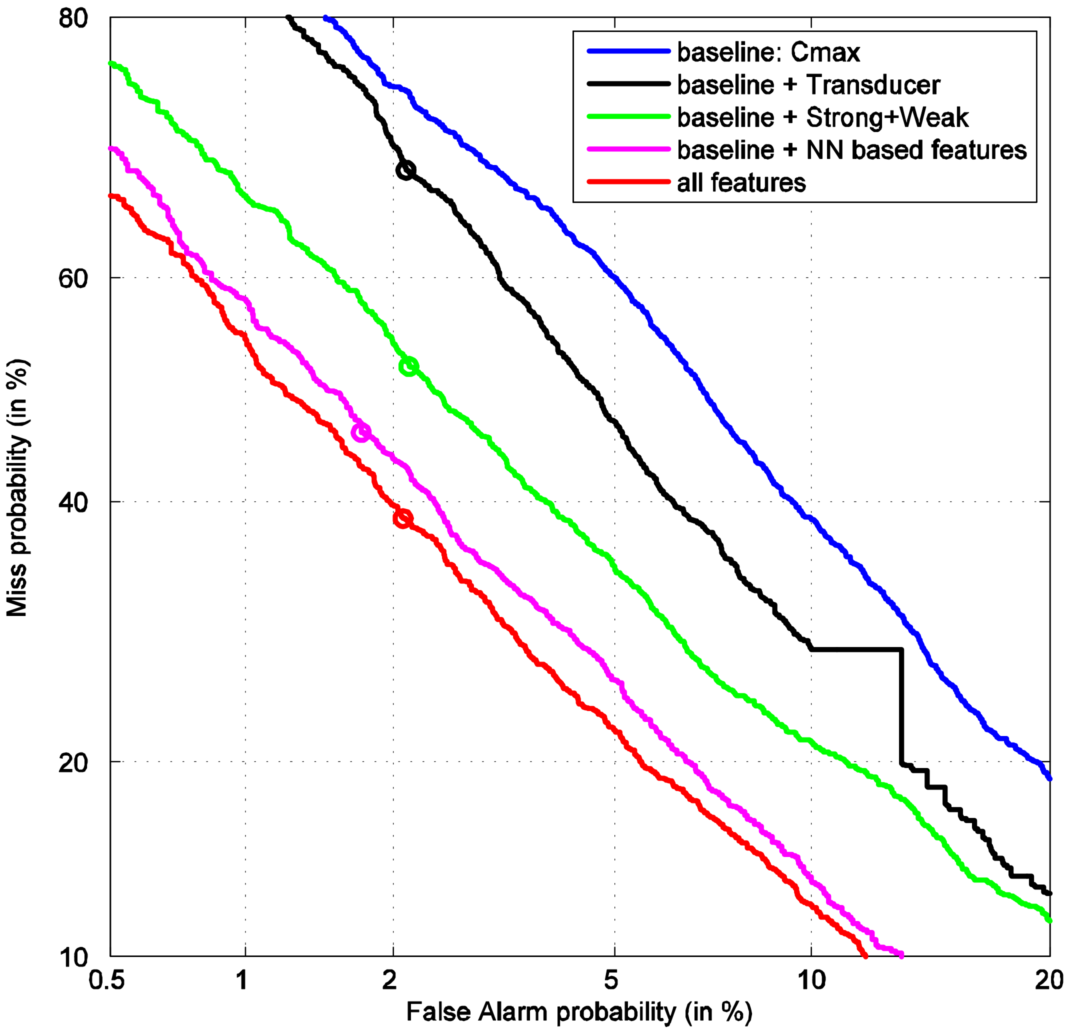
- Using a 20K dictionary and reference phones: 1.9% WER (~1% OOV)
- Using our 5K dictionary and reference phones: 8.3% WER (~4% OOV)

- Best system commits 7.73% detection errors (1511/19556)
- Transducer based detector with 20K dict and reference phones: 0.79%
- Transducer based detector with 5K dict and reference phones: 2.36%

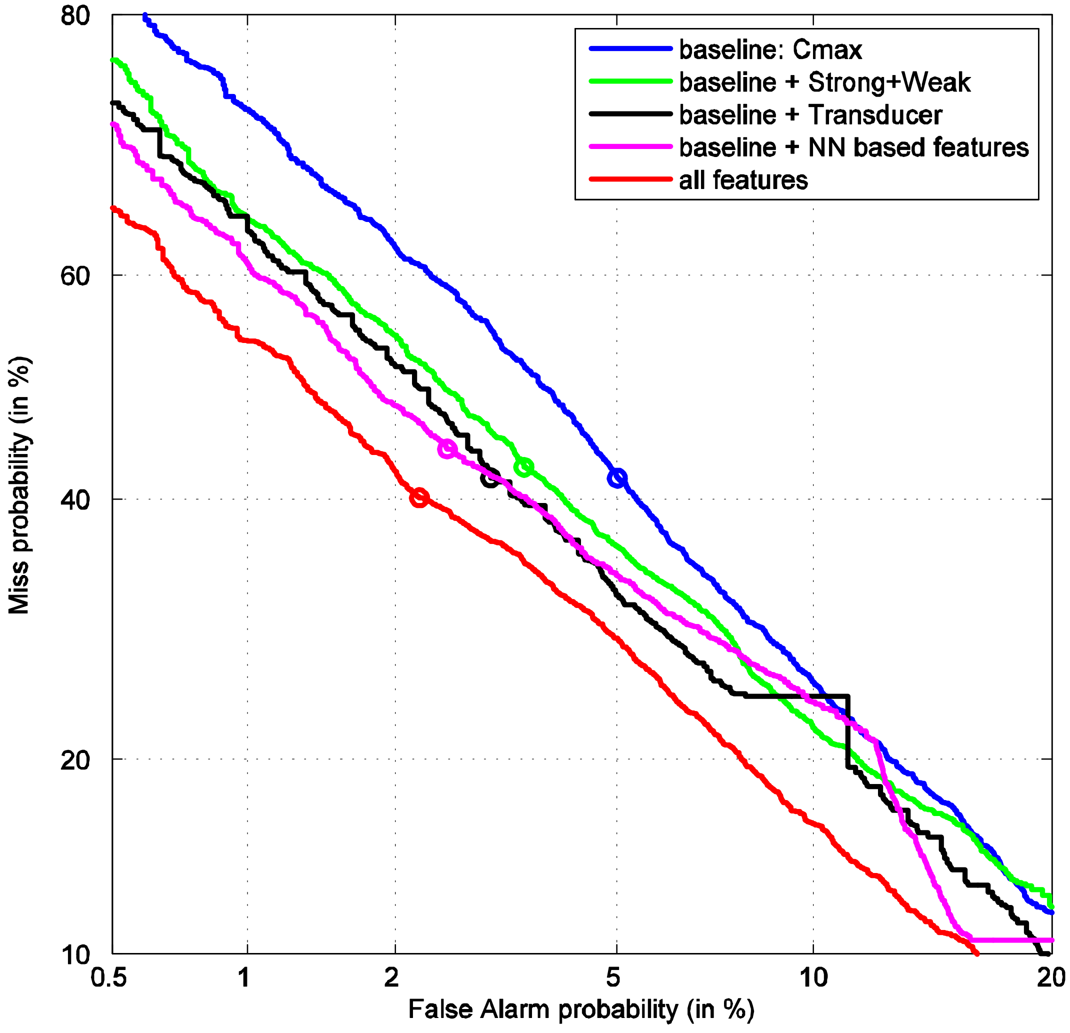
Conclusions: Transduction for OOV Detection and Confidence Estimation

- Transduced word + phone alignment can detect ASR inconsistency
  - Error and OOV detection
- Error model can allow the ‘UNKNOWN’ word to be predicted explicitly
  - Could keep track of repeated unknown phone sequences
- Lower phone error = Better detection
  - Less sensitive to OOV rate (size of dictionary) than WER

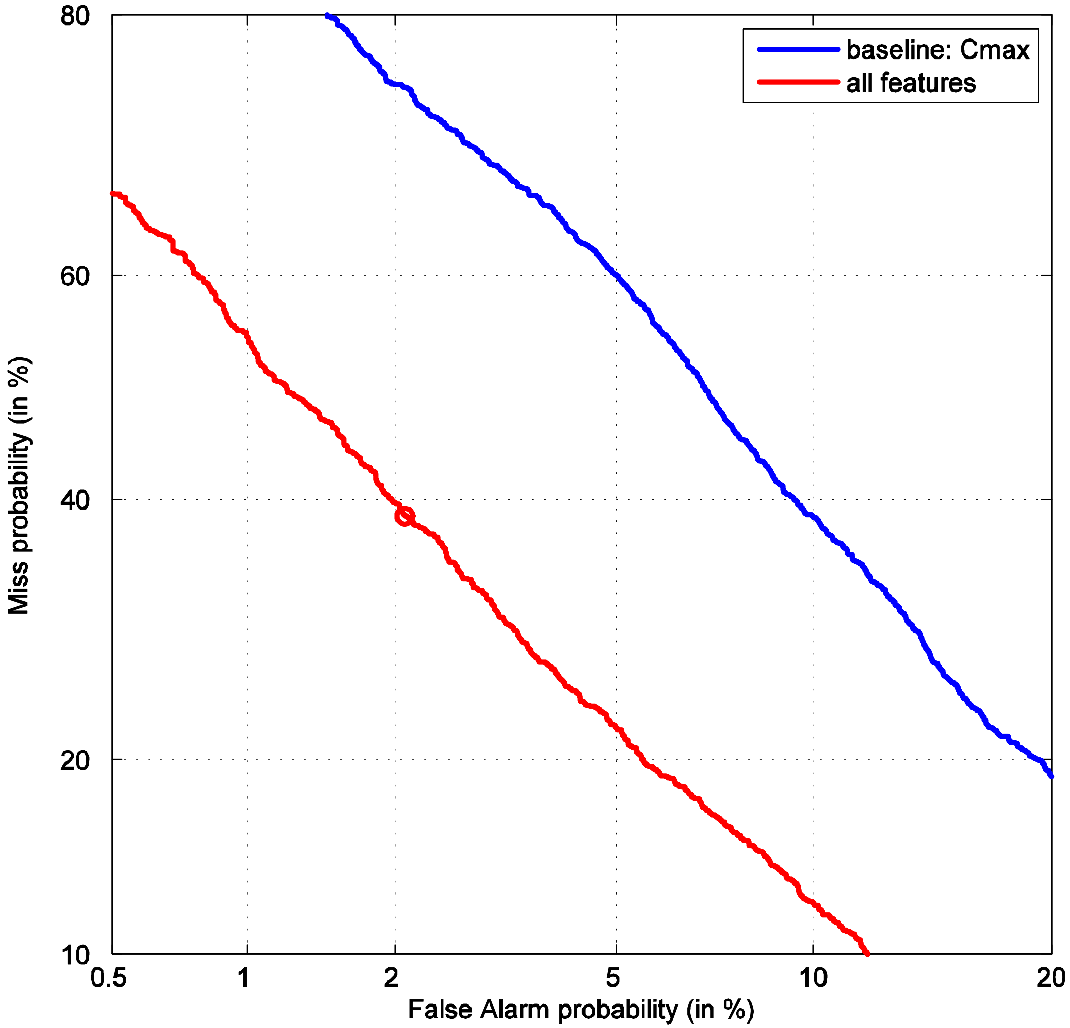
Detection of Out-Of-Vocabulary words (OOV)



Performance for Detection of Errors



Detection of Out-Of-Vocabulary words (OOV)



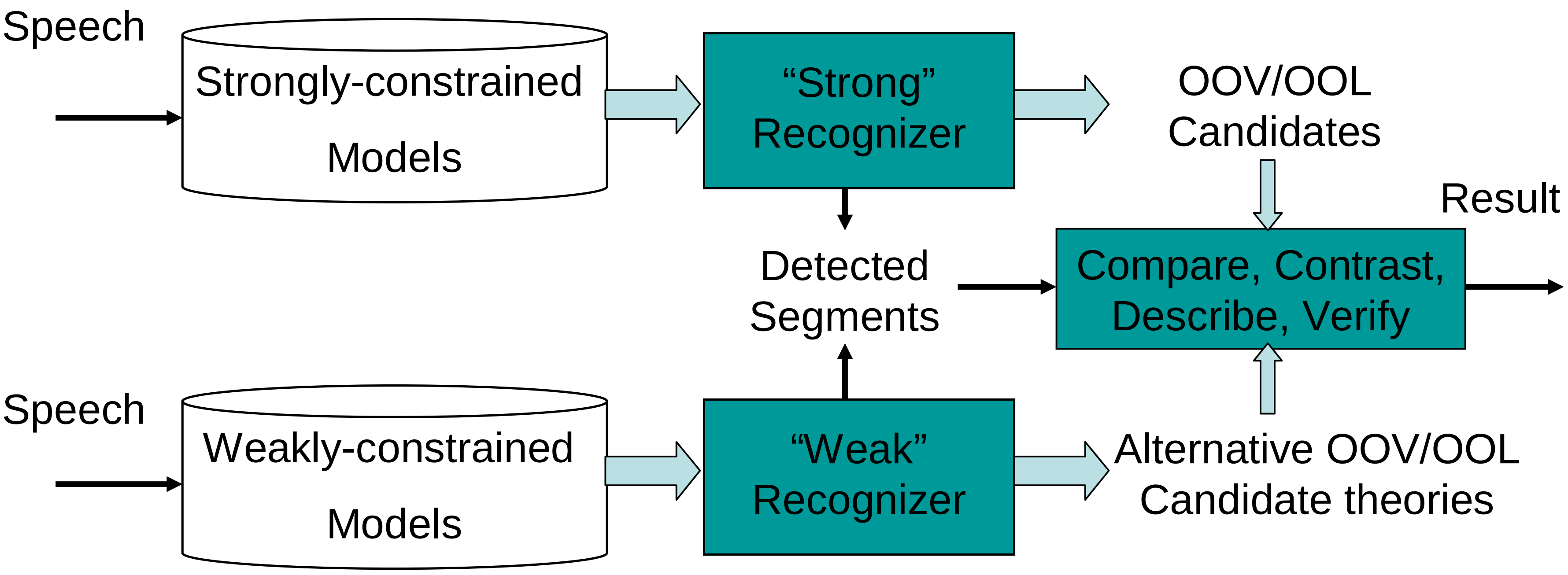
Detection of English Words in Mandarin Utterances and Improved Detection through Universal Phone Models

Rong Tong, Chin-Hui Lee, Haizhou Li, Marco Siniscalchi

Detection of English Words in Mandarin Utterances and Improved Detection through Universal Phone Models

The WS07 UPR Team
Sally, Chris, Rong, Marco, Jon, Haizhou, Chin

Multilingual Phone Recognition for Model Inconsistency Detection & Recovery (Especially for Out-Of-Language Segments)



Language-specific phone recognition is an important tool for OOL description

Summary of Work Done at WS07

- OOL word detection in spontaneous speech (Rong)
  - -In contrast with OOV in read speech, OOL word detection is hard
  - -Good phone models help a great deal
- Universal phone recognition (Haizhou)
  - -Cross-mapping of LDC, IPA, and I^2R phone symbols
  - -Language-universal, -specific, -adaptive phone modeling
  - -Mandarin phone and syllable recognition comparison
- Phone recognition with attribute detection (Marco)
  - -Acoustic phonetic features and speech attributes are fundamental
  - -Attribute to phone mapping for language-specific phone modeling
  - -Multilingual phone recognition (English, Mandarin, Spanish, Hindi)

Issues with OOL Word Detection

- OOL word detection (vs. OOV detection)
  - In-language (IL) and out-of-language (OOL) pronunciations
  - IL LVCSR-based detection with corresponding word boundaries
- Alternative theory hypothesization (strong/weak)
  - Free-phone or free-syllable (for syllabic languages) loop
  - Fixed vs. variable word boundaries
- OOL word verification (“Yes” or “No”)
  - Tests based on null and alternative hypotheses with same models
  - Confidence measures based on different acoustic models
- OOL word description and recognition
  - Description with recognized phone or syllable sequence
  - Recognition with a supplementary vocabulary using accented IL vs. non-accented OOL acoustic models

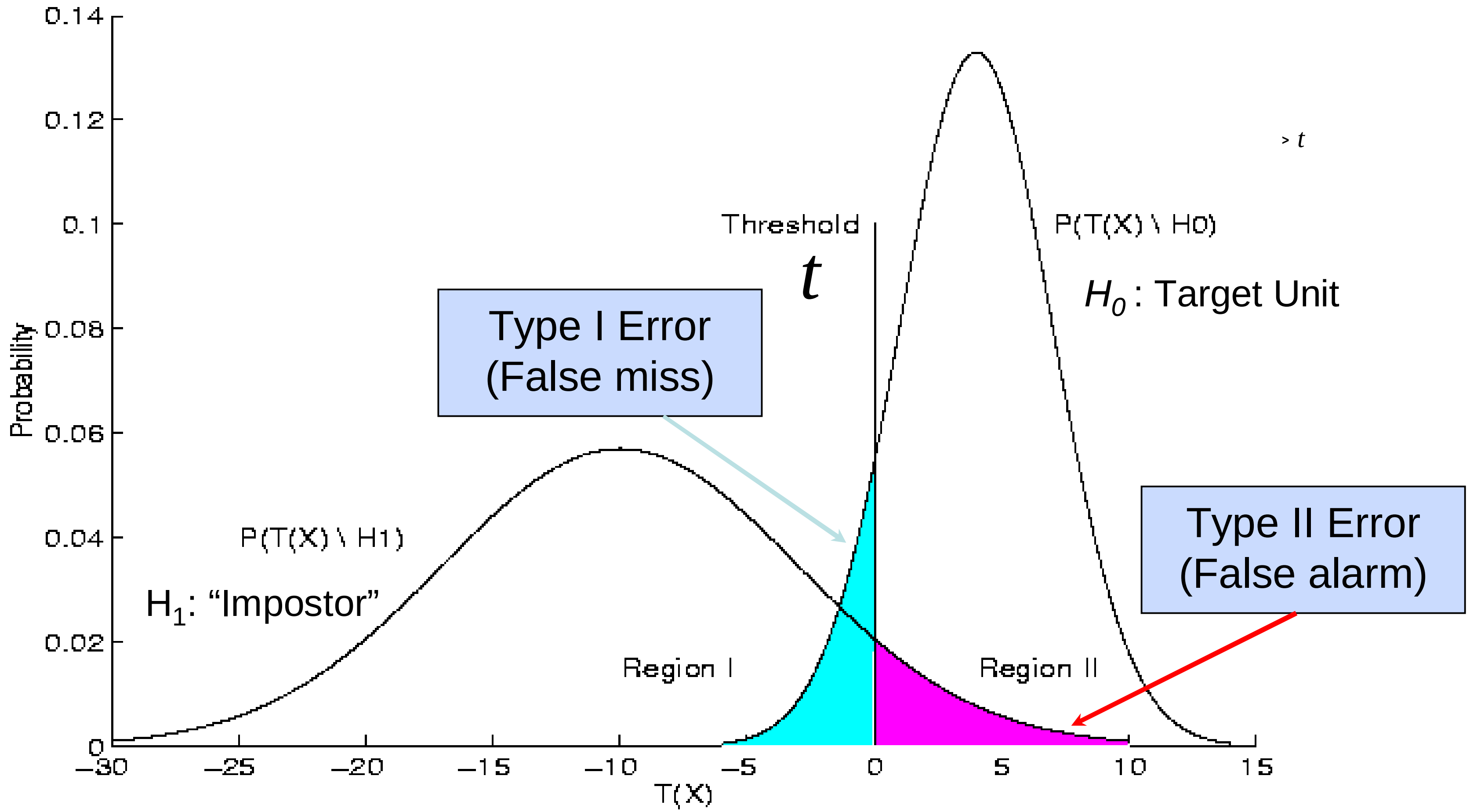
Pattern Verification as Hypothesis Testing

- Design two complementary hypotheses (obtained from a set of “strong” and “weak” recognizers)
  - The *null* hypothesis H_0: detected word is YES (IL)
  - The *alternative* hypothesis H_1: detected word is NO (OOL)
- Plot distributions with date either from H_0 or H_1
- Make decision with a likelihood ratio test (LRT)

If $T = f_0(X \mid \hat{q}_0) / f_1(X \mid \hat{q}_1) > t$, answer YES; otherwise NO

$f_0(.)$ is the model chosen for H_0, $f_1(.)$ for H_1. $\hat{q}_o$ and $\hat{q}_1$ are parameters

Competing Distributions of Test Statistic *T(X)* (Computed for Each Unit of Interest)



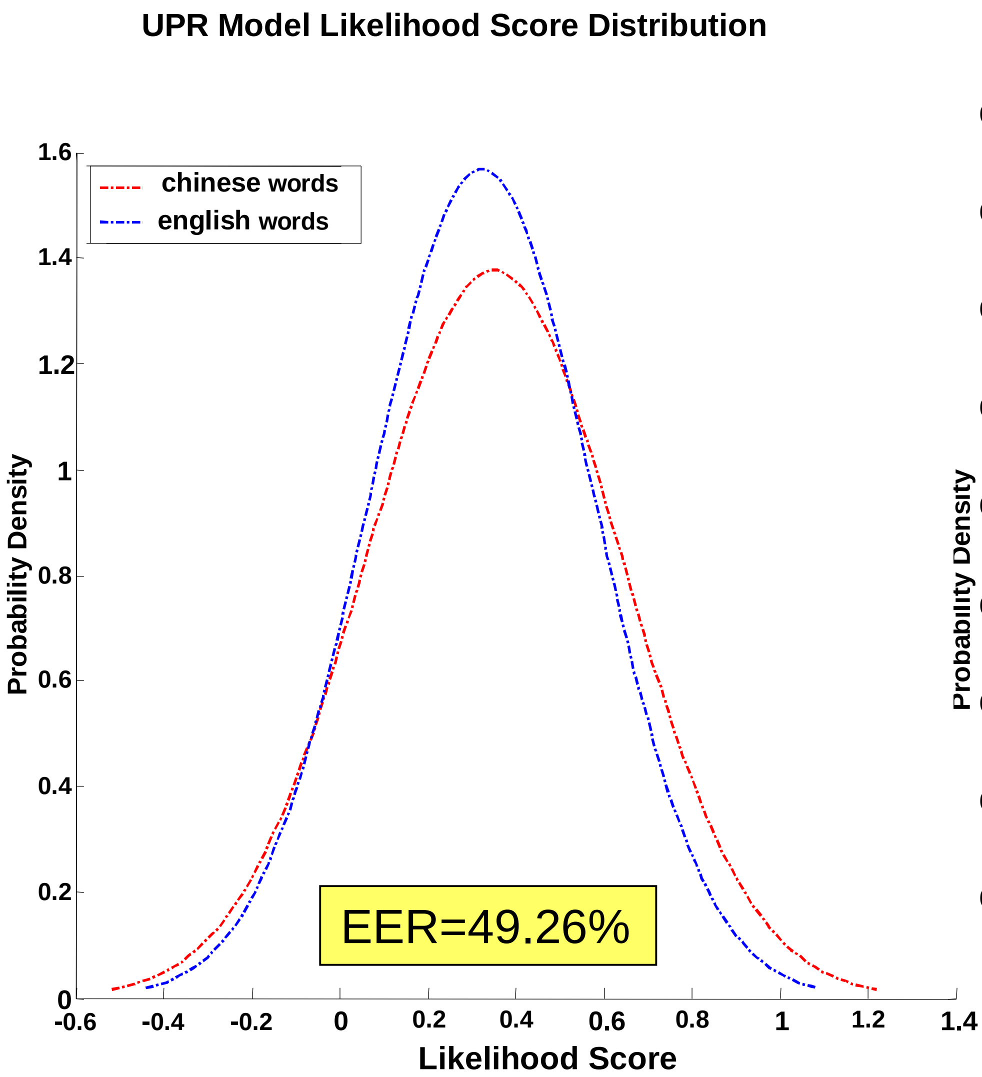
OOL Word Detection

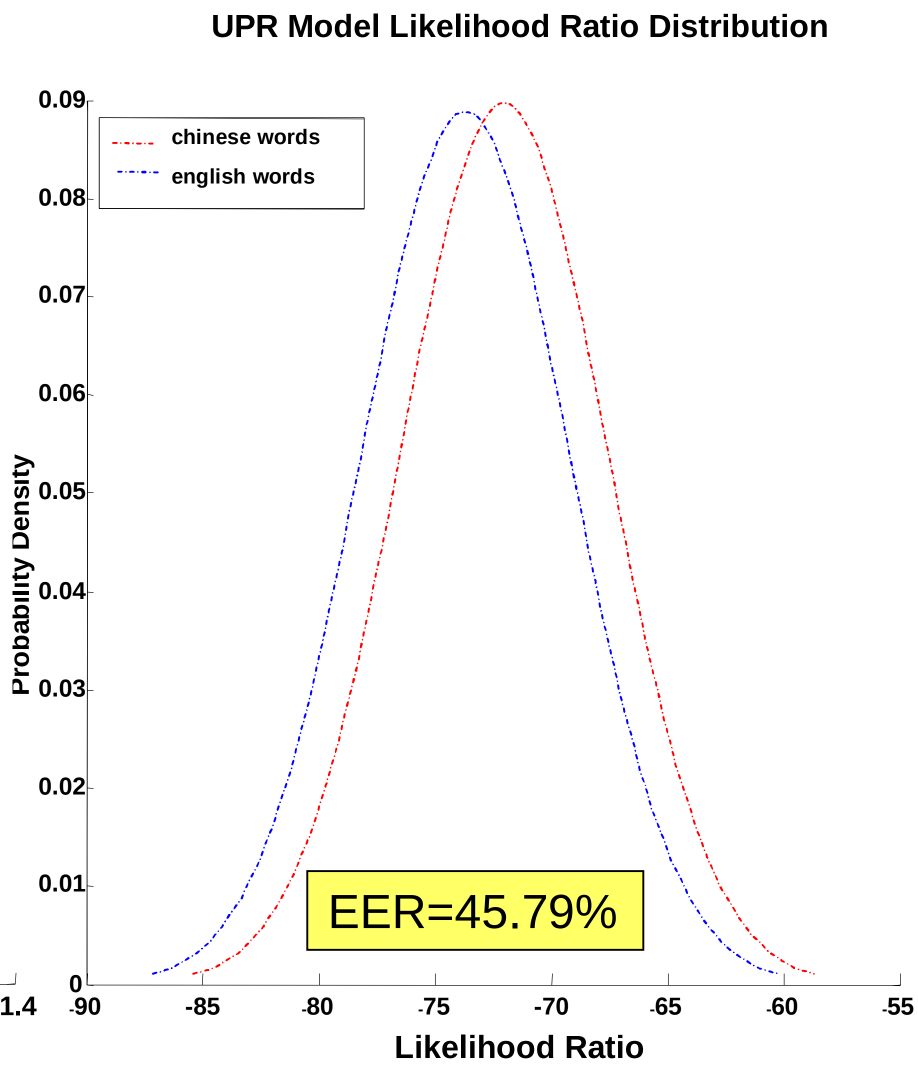
- 931 utterances were selected from 2001 HUB5 Mandarin and CallHome Mandarin data
- Each utterance has at least one English word
- Word boundaries are obtained by performing a forced alignment with the I^2R Mandarin system assuming that the OOL segment is replaced by a randomly-selected IL word
- Likelihood ratio tests for three phone models
  - H_0: decode with bigram language model
  - H_1: decode with phone loop grammar

System/ EER	Likelihood Score (%)	Likelihood Ratio (%)
I2R Mandarin	44.41	26.54
DoD UPR	49.26	45.79
DoD Mandarin	48.03	43.25

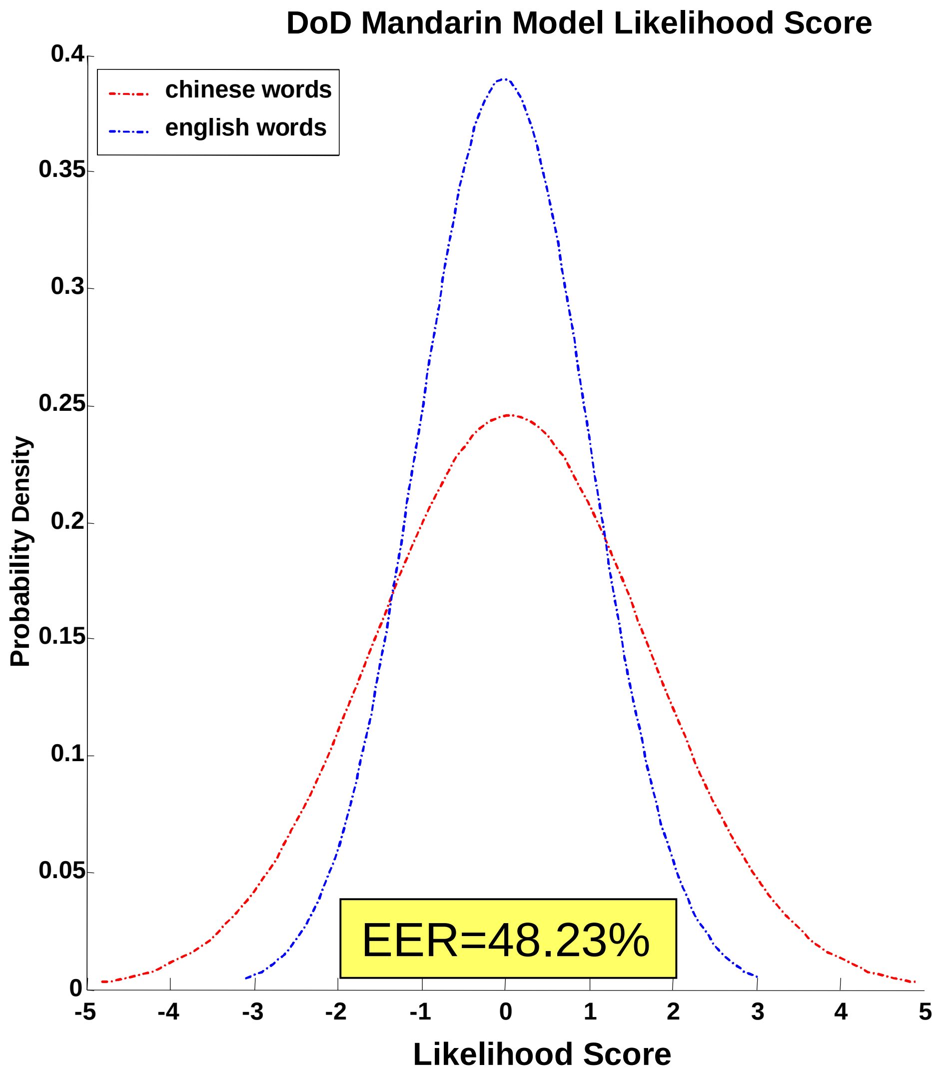
OOL Detection Equal Error Rate

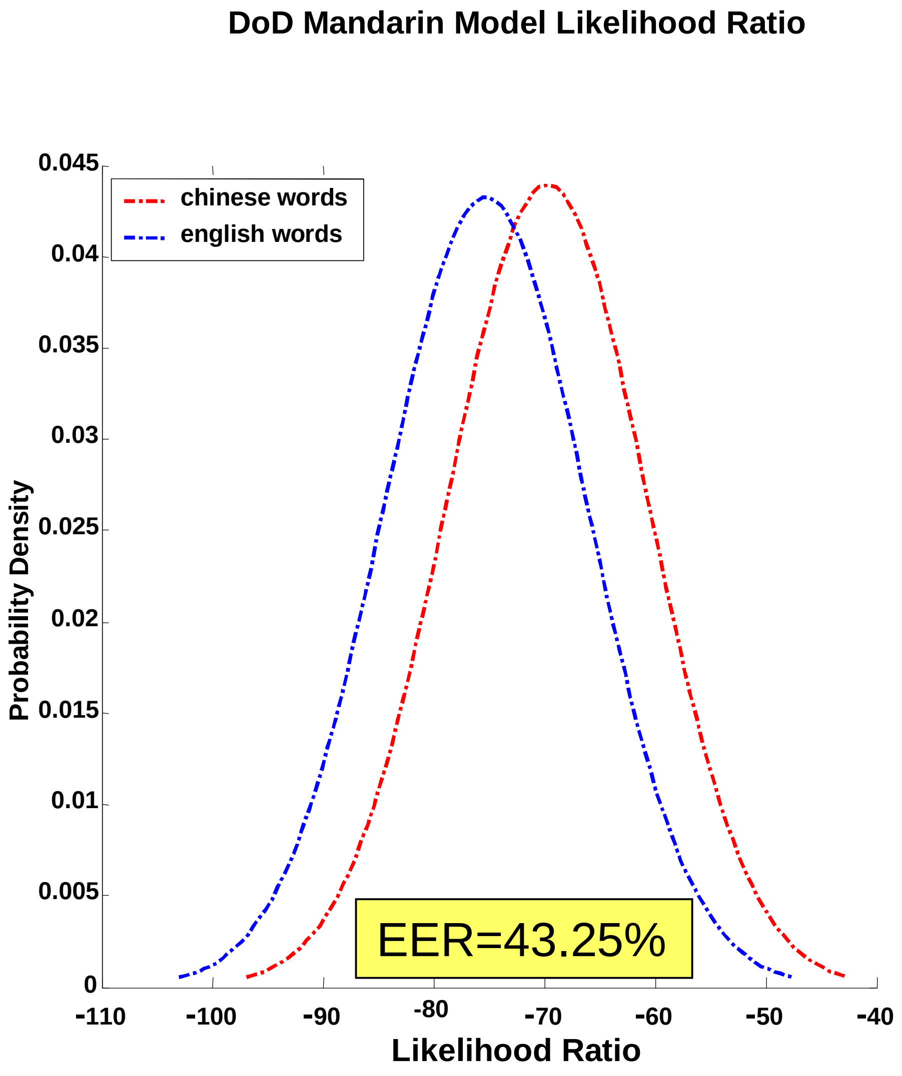
Characterization of IL/OOL Segments - UPR



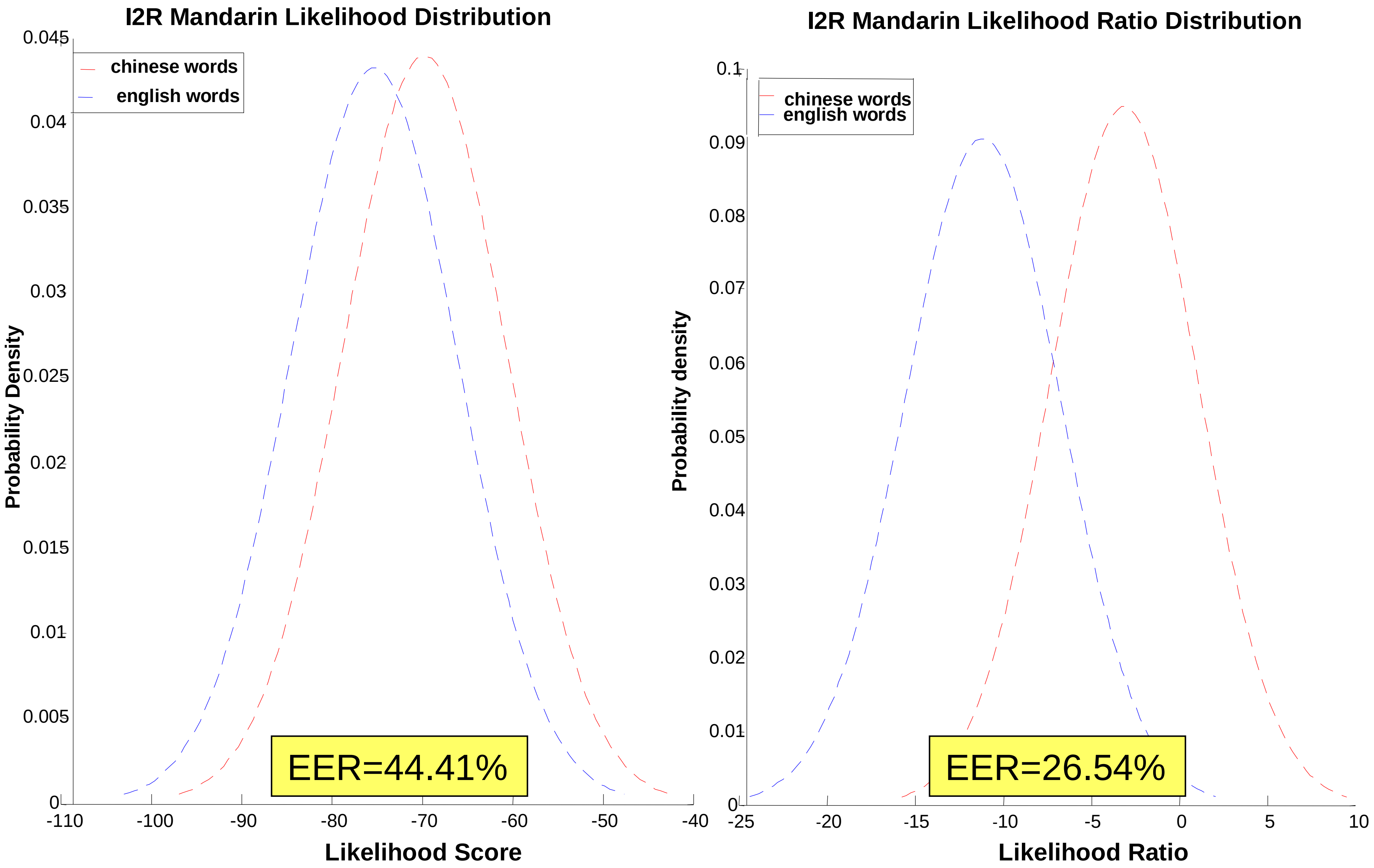


Characterization of IL/OOL Segments – DoD-M



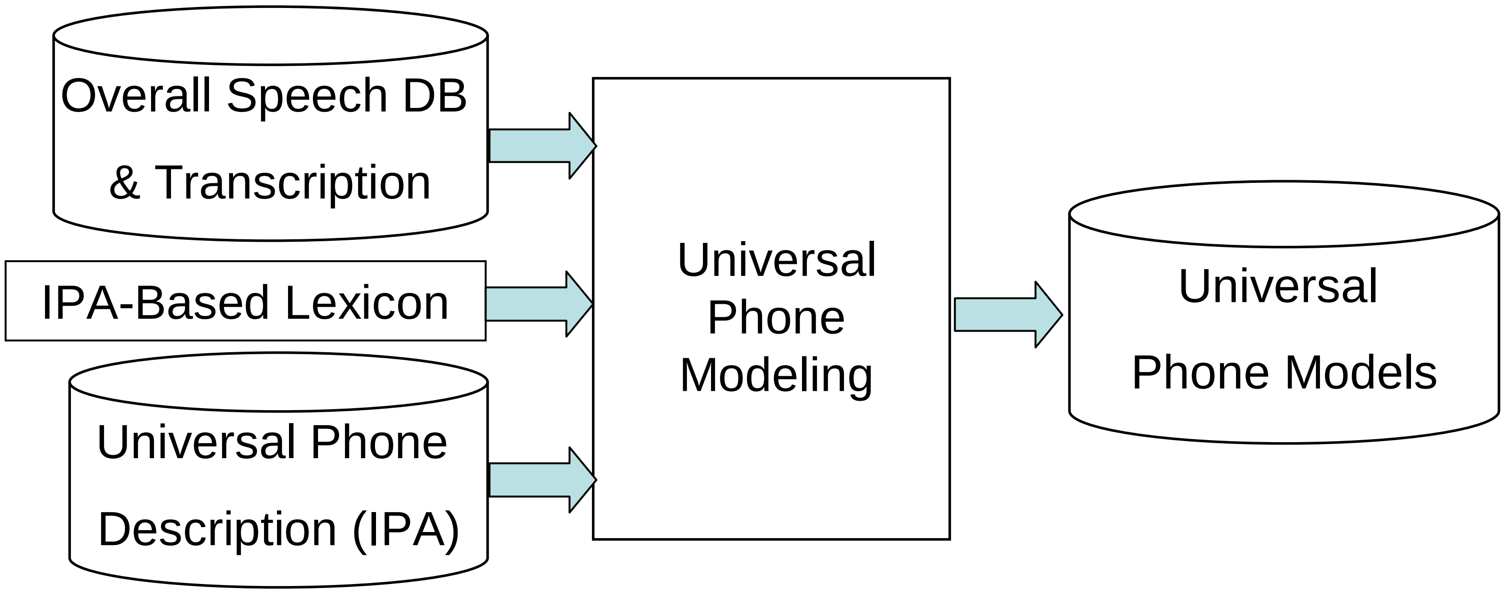


Characterization of IL/OOL Segments – I^{2}R-M



Building Universal Phone Models

- Share data among all languages for acoustic modeling
- Use existing universal phone definition, e.g. IPA
- Expand current ASR capabilities with UPM description
- Examine language-independence & language-dependency



Building Language-Specific Phone Models

- <u>I²R acoustic model</u>
- - 16 mixtures/state
- - 4118 states

- <u>Training corpus 37.25 hrs</u>
- - CallHome
- - CallFriend
- - SAT with a global CMLLR

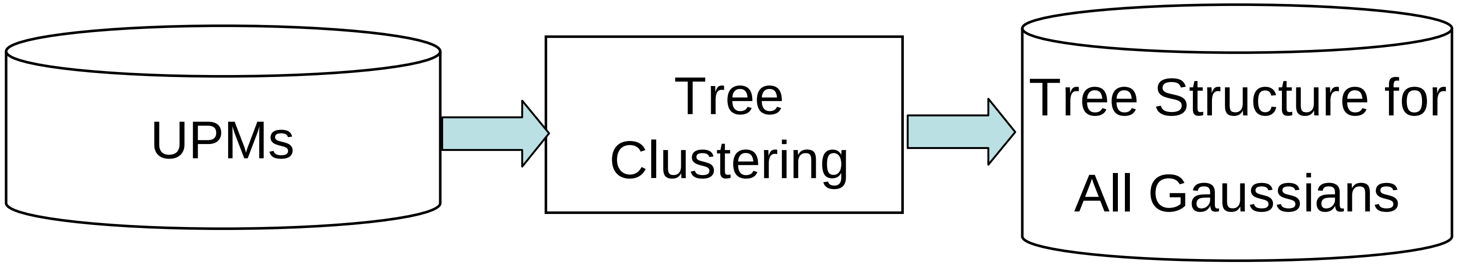
- <u>Lexical resources (team)</u>
- - I²R phonetic lexicon
- - IPA to I²R mapping
- - IPA to LDC mapping

Performance evaluations (%) on Eval 2001

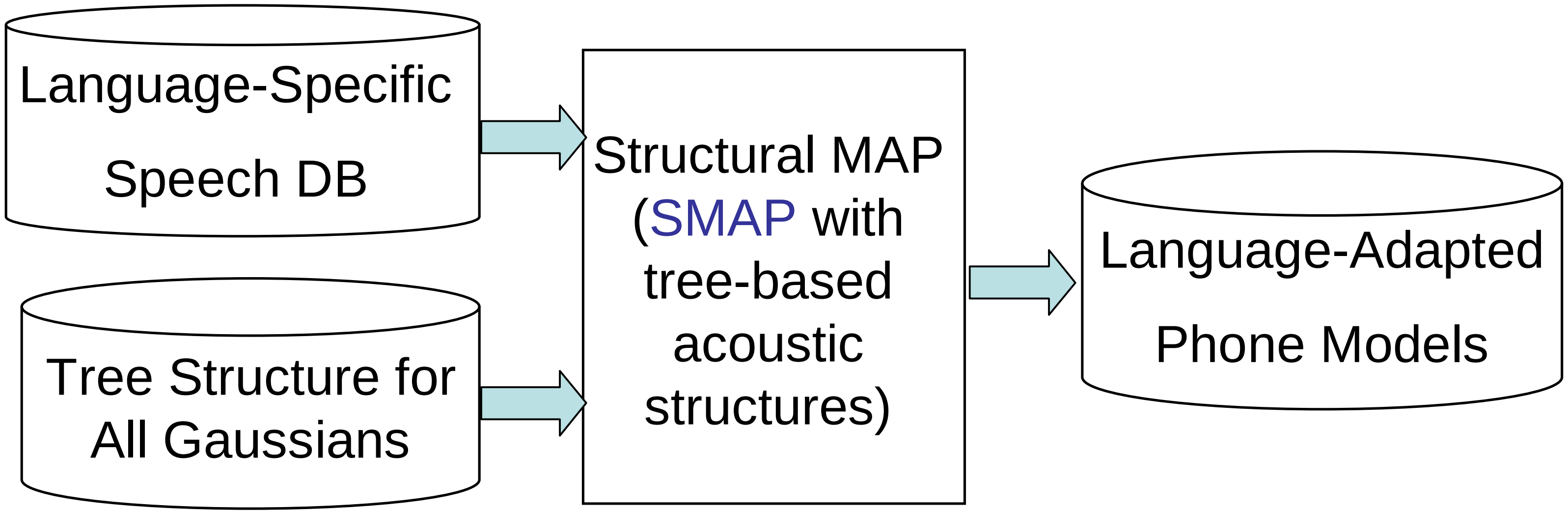
System	Phone Error Rate	Syllable Error Rate
I²R (triphone)	48.24	59.54
DoD Mandarin – monolingual (monophone)	64.11	73.95
DoD Mandarin – monolingual (diphone)	59.83	71.08
DoD UPR (IPA monophone)	70.26	76.73
DoD UPR (IPA diphone)	73.41	78.84

Building Language-Adapted Phone Models

(a) Tree for hierarchical structures modeling of the acoustic space

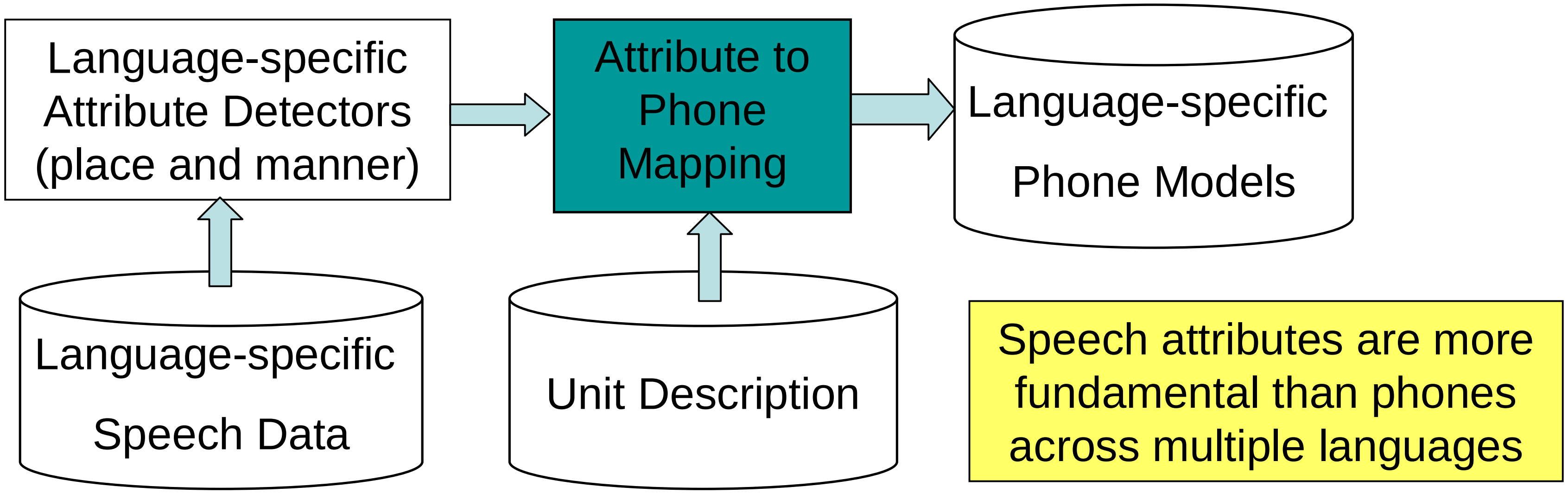


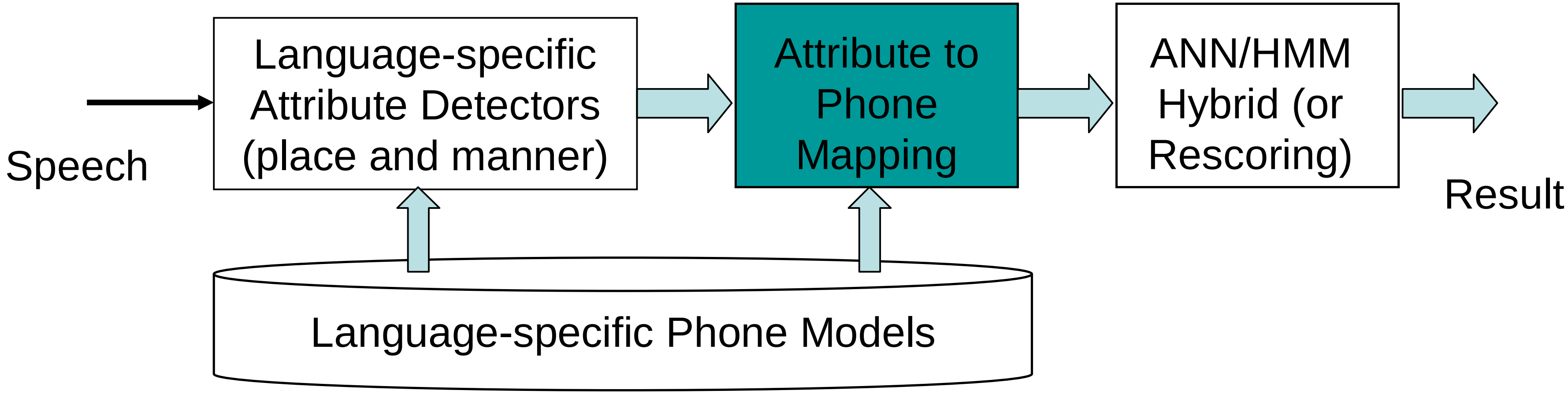
(b) Structural maximum a posteriori (SMAP, Shinoda/Lee, T-SAP 2001) adaptation of UPMs to improve accuracy over language-specific models



Sharing all cross-language structures and speech data to build models

Detector-Based Phone Models & Recognition





Speech Attribute Detectors

- Attributes are fundamental across many languages
  - A collection of 21 dedicated detectors, one for each articulatory attribute (either “present” or “absent”), more attributes seem to give better performance
    - Anterior, back, continuant, coronal, dental, fricative, glottal, approximant, high, labial, low, mid, nasal, retroflex, round, silence, stop, tense, velar, voiced, vowel
  - Detector design with existing or new techniques
    - HMM, ANN and others
    - Long-temporal energy based features and others
- Language-specific phone models
  - Hybrid ANN/HMM model with 3 states per model
  - English, Mandarin, Spanish and Hindi

Mandarin-Specific Attribute to 44-Phone Mapping

	Attribute	Phonemes
	FRICATIVE	ts tsh tsr tsrcl tscl c ch chcl f s sh shr hh v
	NASAL	m n ng
Manner	STOP	p ph t tH k kh
	VOWEL	ih iy iyw a aa ae ah ai aw eh er ey oe ow ox u uw
	APPROXIMANT	l w y l r
	HIGH	iyw iy uw ch ih sh shr
	CORONAL	l n s t
	DENTAL	t tH s
	GLOTTAL	hh
Place	LABIAL	p ph f w m v
	LOW	aa ae ao aw
	MID	eh ah ax oe ox er ao ow ey
	RETROFLEX	tsr tshr shr er r oe
	VELAR	k kh ng
	SILENCE	pause

Detector-Based Multilingual Phone Recognition

- Phone recognition for four languages
  - -English, Mandarin, Spanish, Hindi (four out of six)
  - -OGI Stories Multilingual Corpus
- Phone error rate (PER) comparable with the best

Language	ENG	SPA	HIN	MAN
Train [hr]	1.71	1.10	0.71	0.43
Test [hr]	0.42	0.26	0.17	0.11
Cross-V [hr]	0.16	0.10	0.07	0.03
# of Phones	39	38	46	44

Detector	46.68%	39.99%	45.55	49.19%
BUT*	45.26%	39.95%	45.74	49.93%

Future Work after WS07

- Compare and contrast for OOL word detection (GT)

  -Improve detection with enhanced detection and verification

  -Study code switching in mixed-language utterances

- Extend SMAP in the UPR framework (GT/I^{2}R)

  -Exploit correlation between phones in the multilingual context

  -Expand IPA to share data in training and adapting phone models

  -Continue to improve Mandarin and multilingual ASR

- Continue with language-universal UPR (GT/NTNU)

  -Build language-universal detectors for all acoustic phonetic features

  -Improve attribute to phone mapping for universal phone modeling

Proposal for Continuing Student Support:

Collaboration between JHU, BUT, OvGU Magdeburg

Led by Lukas Burget

Application of OOV and Error Detection Technology:

Language Identification

- Detection of accented English

LVCSR

- Augmenting dictionaries to cover OOV
- Integration into real-time LVCSR
- Applications to discriminative training and adaptation

Summary

Hynek Hermansky

Accomplishments

- Significant improvement in the state-of-the-art in detection of erroneously recognized words
  - differentiate between OOV induced and other errors
- Progress in recognition of domain-independent phones
- Towards alternative architectures in ASR
  - multiple parallel strongly and weakly constrained recognition paths
  - hierarchical phone recognition and sequence matching approach
- New ideas, new connections, new friendships

Thanks

- DoD: UPR support (Pat Schone)
- Microsoft Research: MaxEnt toolkit (Milind Mahajan)
- AMI Consortium (LVCSR system)
- Brno University of Technology (Martin Karafiat, Franta Grezl)
- Institute for Information Research Singapore (Mandarin LVCSR)
- Georgia Tech (detection-based system, SMAP package)

Wish We Had Six More Months J

- All these unfinished experiments…..
- Test techniques on more realistic (e.g. multilingual CallHome) task
- Apply in DARPA GALE, Spoken Term Detection, and Speaker and Language ID
- Implement detection of erroneously recognized words in BUT LVCSR system
- Describe detected OOVs for updating of lexicon
- Further progress towards Universal Phone Recognizer
- Plans for a special issue of Speech Communications on Dealing with Unexpected Acoustic Events in Machine Recognition of Speech
- and the list can go on……..

General Discussion About Why and How To Make Recognizers to Say “I Do Not Know”

Everybody