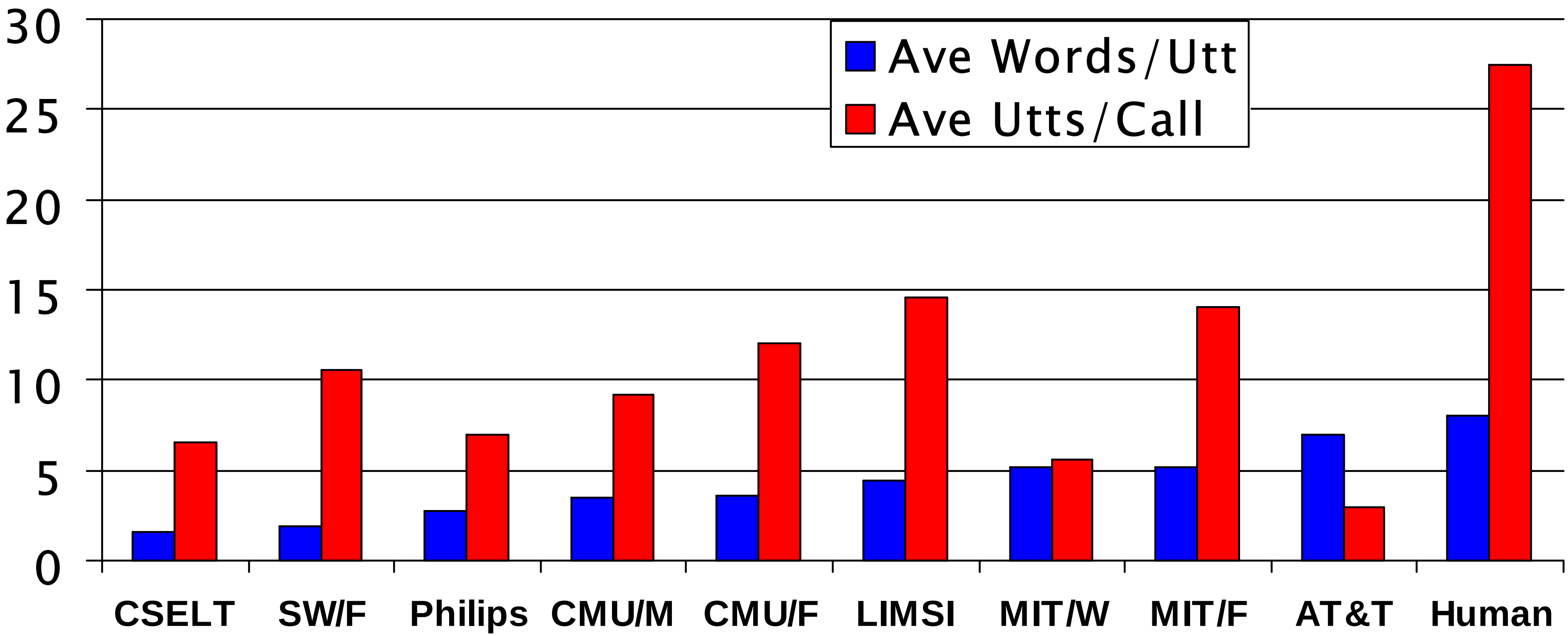
ASR for Spoken-Dialogue Systems

- Introduction
- Speech recognition issues
  - Example using SUMMIT system for weather information
- Reducing computation
- Model aggregation
- Committee-based classifiers

Example Dialogue-based Systems



- **Vocabularies typically have 1000s of words**
- **Widely deployed systems tend to be more conservative**
- **Directed dialogues have fewer words per utterance**
- **Word averages lowered by more confirmations**
- **Human-human conversations use more words**

Telephone-based, Conversational, ASR

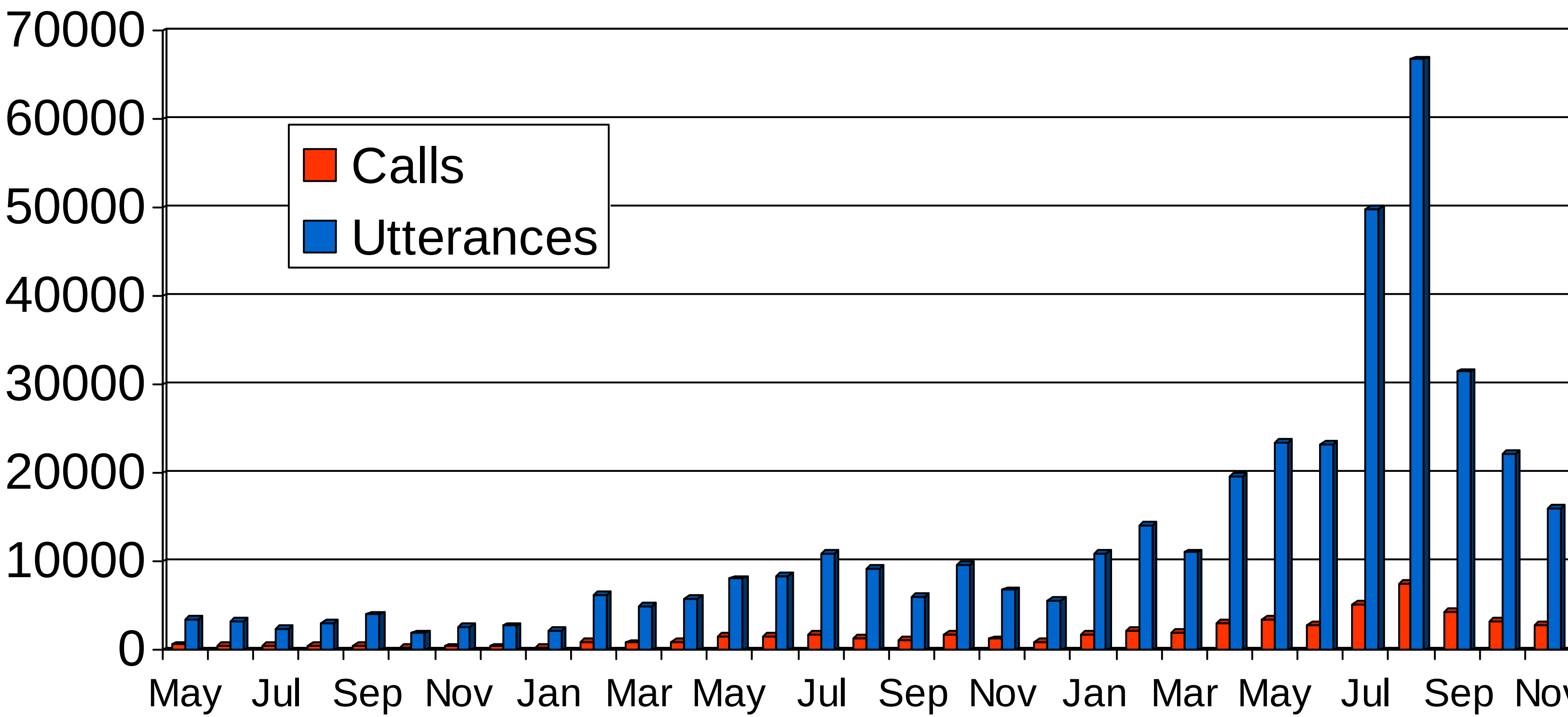
- Telephone bandwidths with variable handsets
- Noisy background conditions
- Novice users with small number of interactions
  - Men, women, children
  - Native and non-native speakers
  - Genuine queries, browsers, hackers
- Spontaneous speech effects
  - e.g., filled pauses, partial words, non-speech artifacts
- Out-of-vocabulary words and out-of-domain queries
- Full vocabulary needed for complete understanding
  - Word and phrase spotting are not primary strategies
  - Mixed-initiative dialog provides little constraint to recognizer
- Real-time decoding

Data Collection Issues

- **System development is chicken & egg problem**
- **Data collection has evolved considerably**
  - **Wizard-based $\rightarrow$ system-based data collection**
  - **Laboratory deployment $\rightarrow$ public deployment**
  - **100s of users $\rightarrow$ thousands $\rightarrow$ millions**
- **Data from real users solving real problems accelerates technology development**
  - **Significantly different from laboratory environment**
  - **Highlights weaknesses, allows continuous evaluation**
  - **But, requires systems providing real information!**
- **Expanding corpora requires unsupervised training or adaptation to unlabelled data**

Data Collection (Weather Domain)

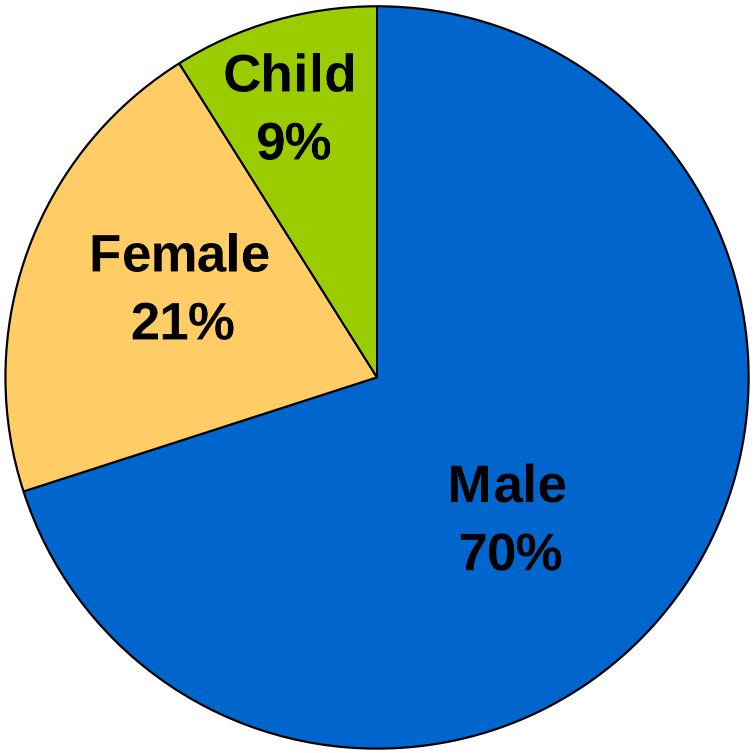
- **Initial collection of 3,500 read utterances and 1,000 wizard utterances**

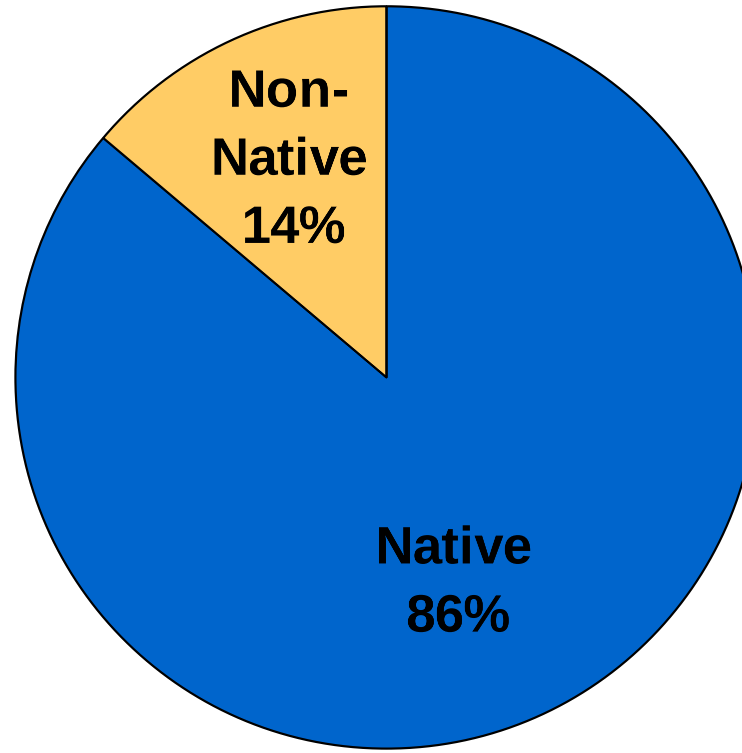


- **Over 756K utterances from 112K calls since May, 1997**

Weather Corpus Characteristics

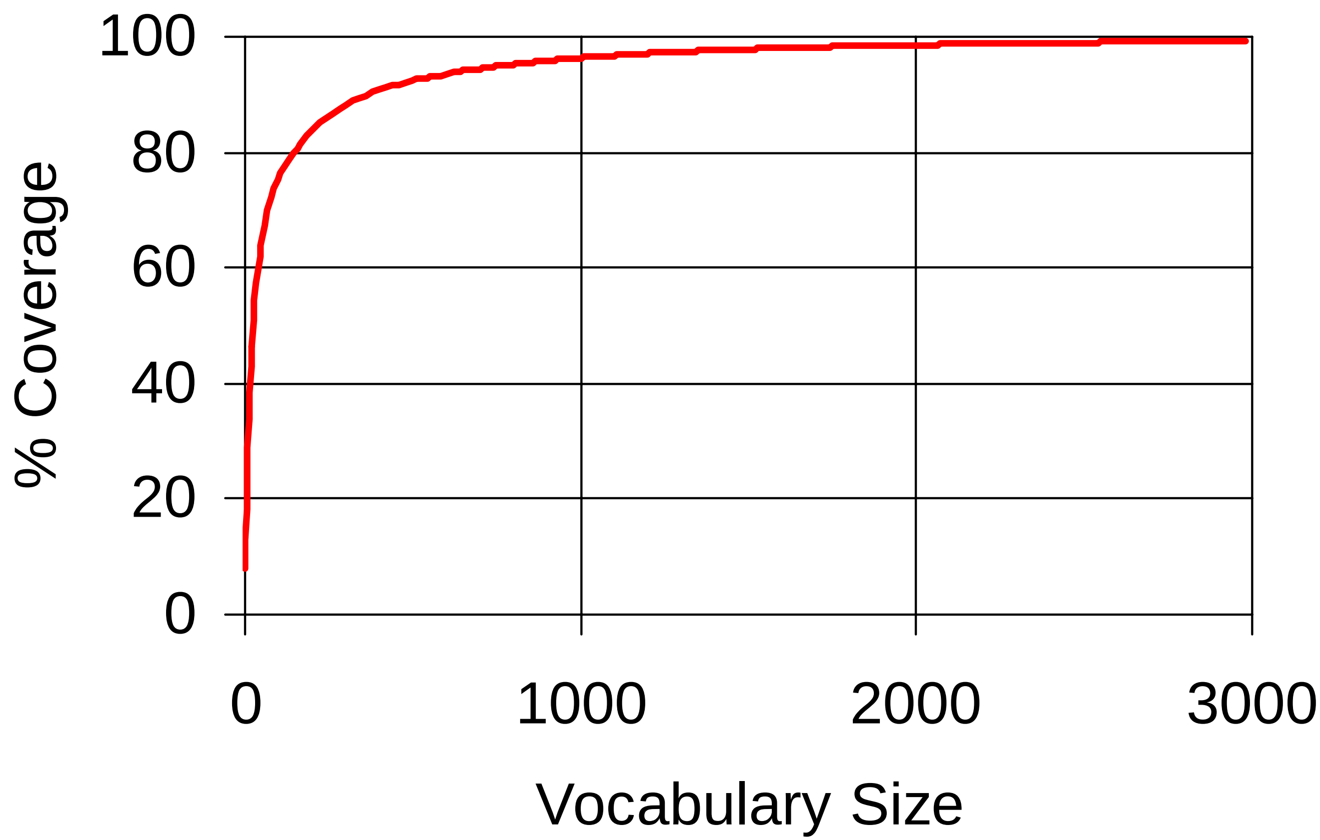
- **Corpus dominated by American male speakers**





- **Approximately 11% of data contained significant noises**
- **Over 6% of data contained spontaneous speech effects**
- **At least 5% of data from speakerphones**

Vocabulary Selection



- **Constrained domains naturally limit vocabulary sizes**
- **2000 word vocabulary gives good coverage for weather**
- **~2% out-of-vocabulary rate on test sets**

Vocabulary

- Current vocabulary consists of nearly 2000 words
- Based on system capabilities and user queries

Type	Size	Examples
Geography	933	boston, alberta, france, africa
Weather	217	temperature, snow, sunny, smog
Basic	815	i, what, january, tomorrow

- Incorporation of common reduced words & word pairs

Type	Examples
Reduction	give_me, going_to, want_to, what_is, i_would
Compound	clear_up, heat_wave, pollen_count

- Lexicon based on syllabified LDC PRONLEX dictionary

Example Vocabulary File

<>*
<pause1>
<pause2>
<uh>
<um>
<unknown>*
a
a_m
am
don+t
new_york_city
sixty
today
today+s

- Sorted alphabetically
- Utterance start & end marker
- Pauses at utterance start & end
- Filled pause models
- *'d items have no acoustic realization
- Out-of-vocabulary word model
- <>'d words don't count as errors
- Underbars distinguish letter sequences from actual words
- \+ symbol conventionally used for '
- Lower case is a common convention
- Numbers tend to be spelled out
- Each word form has separate entry

Example Baseform File

<pause1> : - +
<pause2> : - +
<uh> : ah_fp
<um> : ah_fp m
a_m : ey & eh m
either : (iy , ay) th er
laptop : l ae pd t aa pd
new_york : n uw & y ao r kd
northwest : n ao r th w eh s td
trenton : tr r eh n tq en
winter : w ih nt er

previous symbol can repeat

special filled pause vowel

alternate pronunciations

word break allowing pause

Editing Generated Baseforms

- **Automatically generated baseform file should be manually checked for the following problems:**
  - **Missing pronunciation variants that are needed**
  - **Unwanted pronunciation variants that are present**
  - **Vocabulary words missing in PRONLEX**

```
going_to     : g ow ix ng & t uw
reading      : ( r iy df ix ng , r eh df ix ng )
woburn       : <???>
```

↓

```
going_to     : g ( ow ix ng & t uw , ah n ax )
reading      : r eh df ix ng
woburn       : w ( ow , uw ) b er n
```

Applying Phonological Rules

- *Phonemic* baseforms are canonical representation
- Baseforms may have multiple acoustic realizations
- Acoustic realizations are *phones* or *phonetic units*
- Example:

batter : b ae tf er

This can be realized phonetically as:

bcl b ae tcl t er ← **Standard /t/**

or as:

bcl b ae dx er ← **Flapped /t/**

Example Phonological Rules

- **Example rule for /t/ deletion ("destination"):**

{s} t {ax ix} => [tcl t];

Left Context | Phoneme | Right Context | Phonetic Realization

- **Example rule for palatalization of /s/ ("miss you"):**

{} s {y} => s | sh;

Language Modelling

- Class bi- and trigrams used to produce 10-best outputs
- Training data augmented with city and state constraints
- Relative entropy measure used to help select classes

raining, snowing	humidity, temperature
cold, hot, warm	advisories, warnings
extended, general	conditions, forecast, report

- 200 word classes reduced perplexities and error rates

Type	Perplexity	% Word Error Rate
word bigram	18.4	16.0
+ word trigram	17.8	15.5
class bigram	17.6	15.6
+ class trigram	16.1	14.9

Defining *N*-gram Word Classes

CITY ==> boston
CITY ==> chicago
CITY ==> seattle

Class definitions have class name on left and word on right

<U>_DIGIT ==> one
<U>_DIGIT ==> two
<U>_DIGIT ==> three

Class names with "<U>_" forces all words to be equally likely

DAY ==> today | tomorrow

Alternate words in class can be placed on same line with "|" separator

The Training Sentence File

- **An *n*-gram model is estimated from training data**
- **Training file contains one utterance per line**
- **Words in training file must have same case and form as words in vocabulary file**
- **Training file uses the following conventions:**
  - **Each clean utterance begins with <pause1> and ends with <pause2>**
  - **Compound word underbars are typically removed before training**
  - **Underbars automatically re-inserted during training based on compound words present in vocabulary file**
- **Special artifact units may be used for noises and other significant non-speech events:**
  - **<clipped1>, <clipped2>, <hangup>, <cough>, <laugh>**

Example Training Sentence File

<pause1> when is the next flight to chicago <pause2>

<pause> to san <partial> san francisco <pause2>

<pause1> <um> boston <pause2>

<clipped1> it be in time <pause2>

<pause1> good bye <hangup>

<pause1> united flight two oh four <pause2>

<pause1> <cough> excuse me <laugh> <pause2>

partial word, e.g., san die(go)

clipped word, e.g., ~(w)ill it

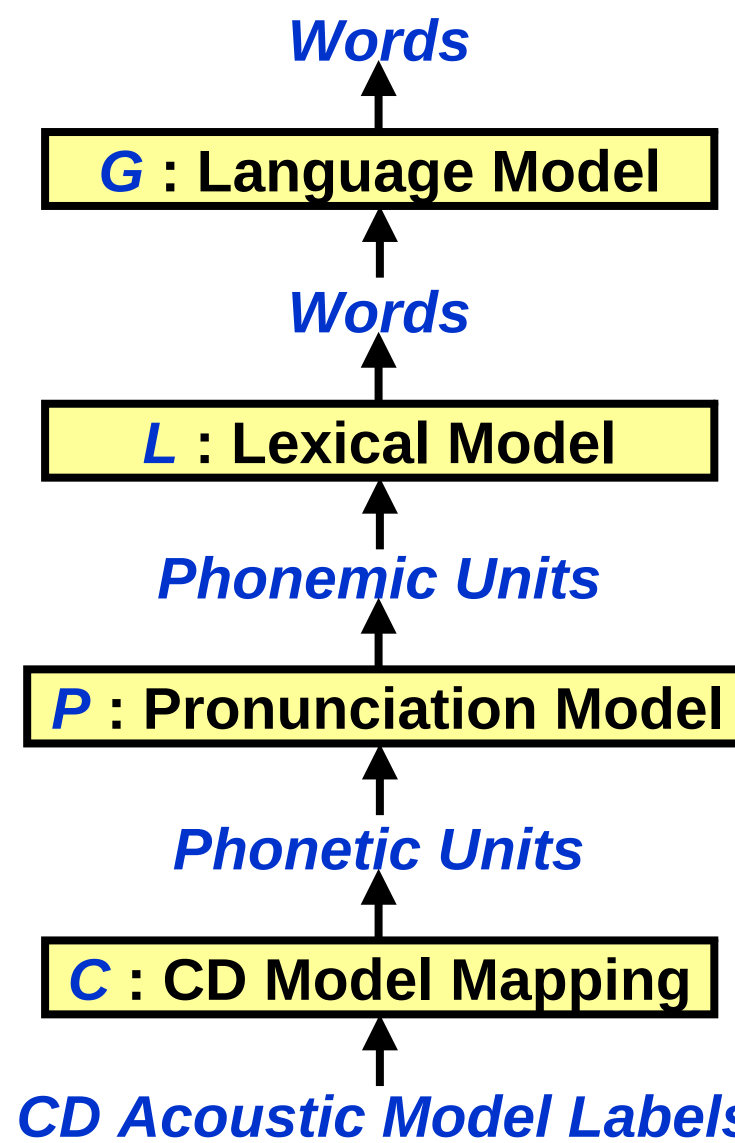
all significant sounds are transcribed

Composing FST Lexical Networks

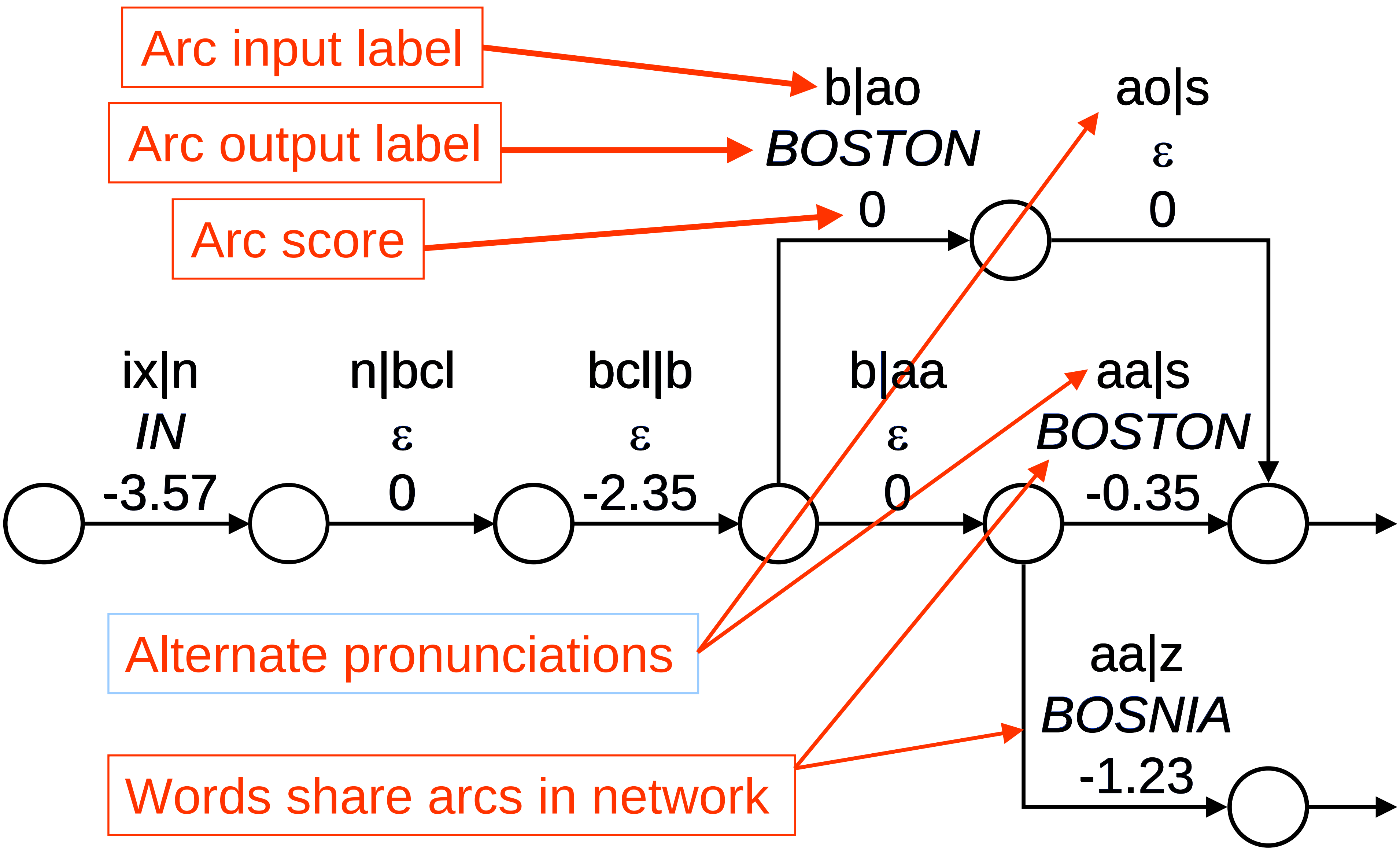
- **Four basic FST networks are composed to form full search network.**
  - ***G*** **: Language model**
  - ***L*** **: Lexical model**
  - ***P*** **: Pronunciation model**
  - ***C*** **: Context-dependent acoustic model mapping**

- **Mathematical composed using the expression:**

$$C \circ P \circ L \circ G$$



FST Example



Arc input label
Arc output label
Arc score
b|ao
BOSTON
0
ao|s
ε
0
ix|n
IN
-3.57
n|bcl
ε
0
bcl|b
ε
-2.35
b|aa
ε
0
aa|s
BOSTON
-0.35
aa|z
BOSNIA
-1.23
Alternate pronunciations
Words share arcs in network

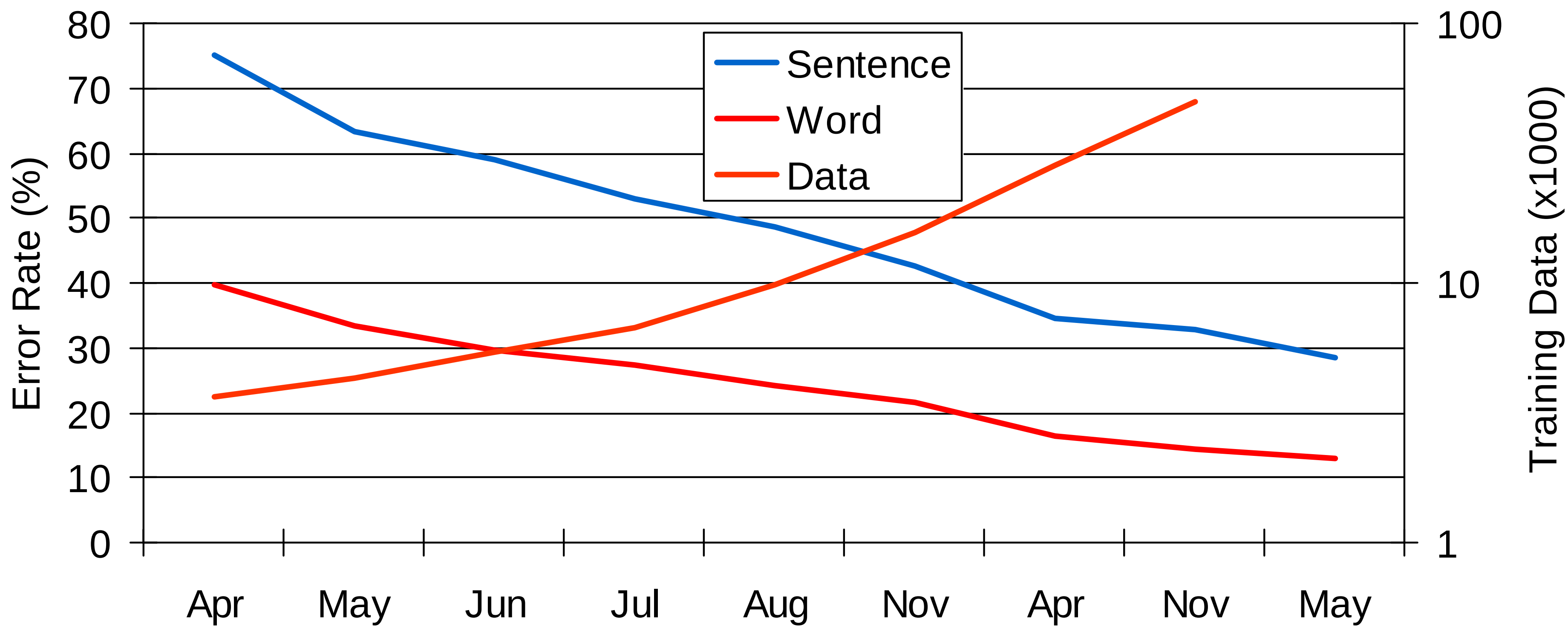
Acoustic Models

- **Models can be built for segments and boundaries**
  - **Best accuracy can be achieved when both are used**
  - **Current *real-time* recognition uses only boundary models**
- **Boundary labels combined into classes**
  - **Classes determined using decision tree clustering**
  - **One Gaussian mixture model trained per class**
  - **112 dimension feature vector reduced to 50 dimensions via PCA**
  - **1 Gaussian component for every 50 training tokens (based on # dims)**
- **Models trained on over 100 hours of spontaneous telephone speech collected from several domains**

Search Details

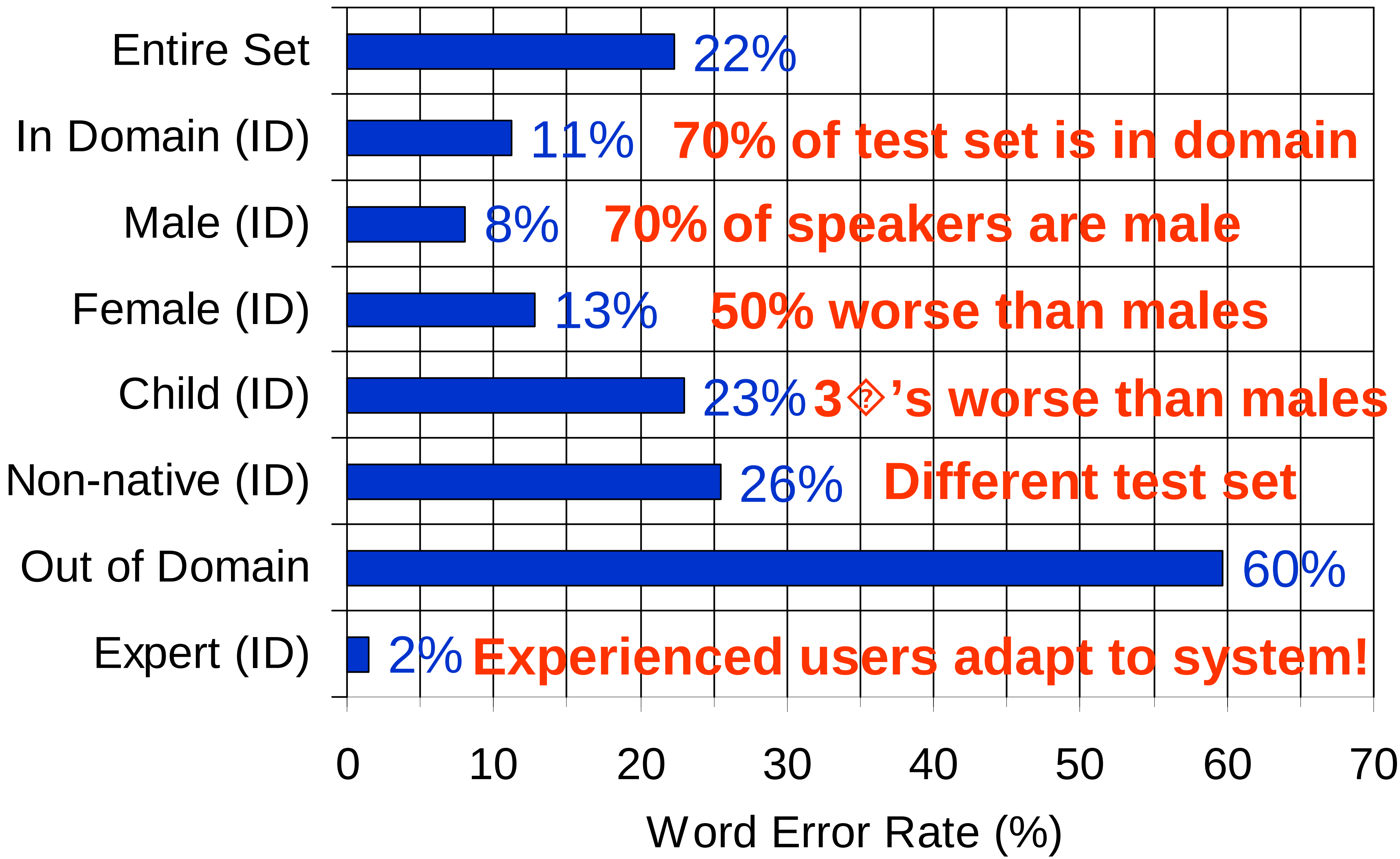
- **Search uses forward and backward passes:**
  - **Forward Viterbi search using bigram**
  - **Backwards A* search using bigram to create a word graph**
  - **Rescore word graph with trigram (i.e., subtract bigram scores)**
  - **Backwards A* search using trigram to create *N*-best outputs**
- **Search relies on two types of pruning:**
  - **Pruning based on relative likelihood score**
  - **Pruning based maximum number of hypotheses**
  - **Pruning provides tradeoff between speed and accuracy**
- **Search can control tradeoff between insertions and deletions**
  - **Language model biased towards short sentences**
  - **Word transition weight (wtw) heuristic adjusted to remove bias**

Recognition Experiments

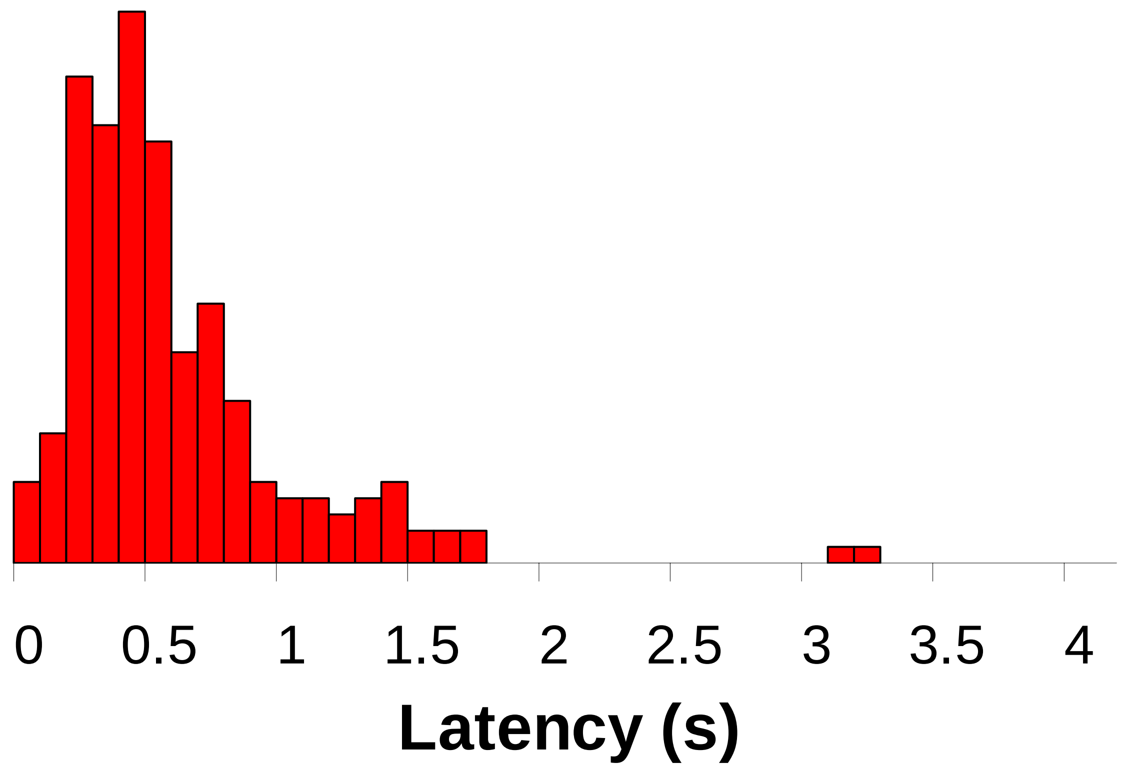


- **Collecting real data improves performance:**
  - **Enables increased complexity and improved robustness for acoustic and language models**
  - **Better match than laboratory recording conditions**

Error Analysis (2506 Utterance Test Set)



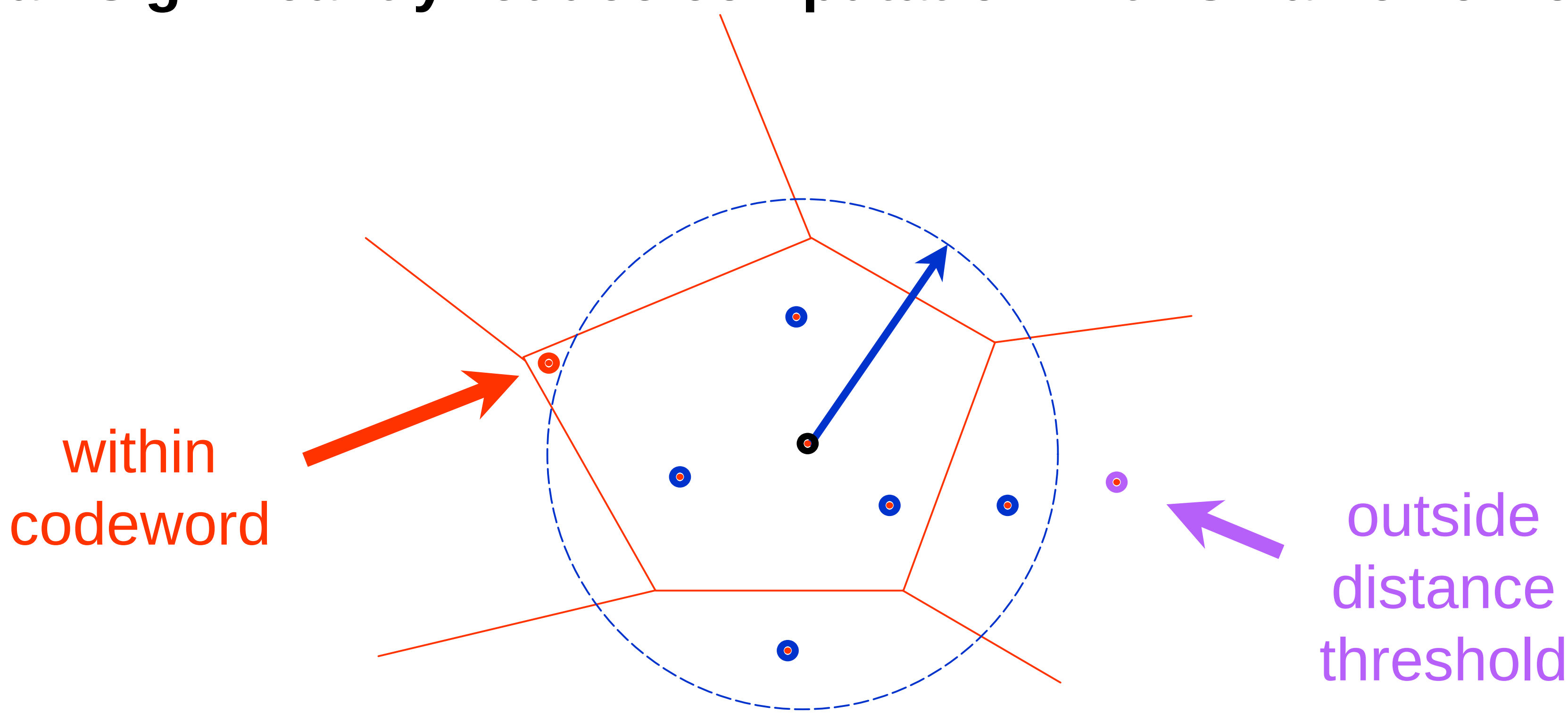
A* Search Latency



- **Average latency ≈ .62 seconds**
- **85% < 1 second ; 99% < 2 seconds**
- **Latency not dependent on utterance length**

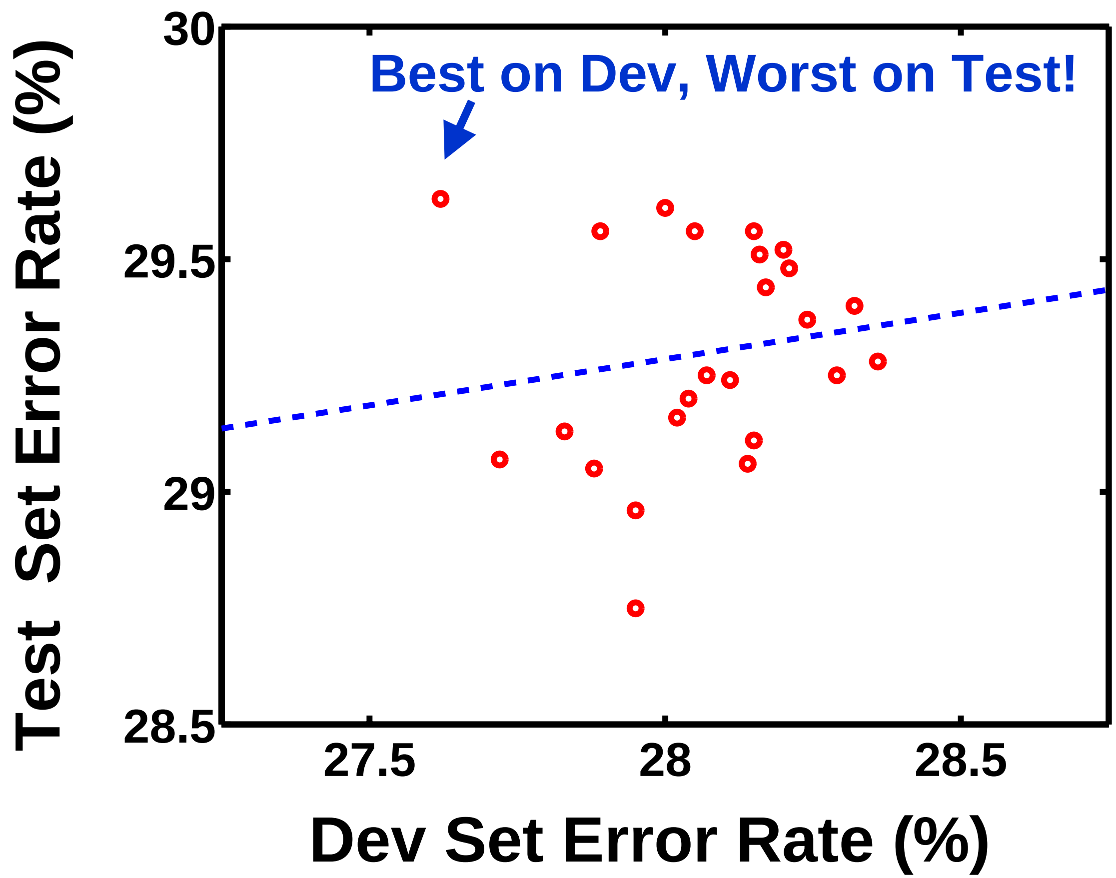
Gaussian Selection

- ~50% of total computation is evaluation of Gaussian densities
- Can use binary VQ to select mixture components to evaluate
- Component selection criteria for each VQ codeword:
  - Those within distance threshold
  - Those within codeword (i.e., every component used at least once)
  - At least one component/model per codeword (i.e., only if necessary)
- Can significantly reduce computation with small error loss



Model Aggregation

- *K*-means and EM algorithms converge to different local minima from different initialization points
- Performance on development data not necessarily a strong indicator of performance on test data
  - TIMIT phonetic recognition error for 24 training trials



Correlation
Coeff = 0.16

Aggregation Experiments

- Combining different training runs can improve performance
- Three experimental systems: phonetic classification, phonetic recognition (TIMIT), and word recognition (RM)
- Acoustic models:
  - Mixture Gaussian densities, randomly initialized *K*-means
  - 24 different training trials
- Measure average performance of M unique N-fold aggregated models (starting from 24 separate models)

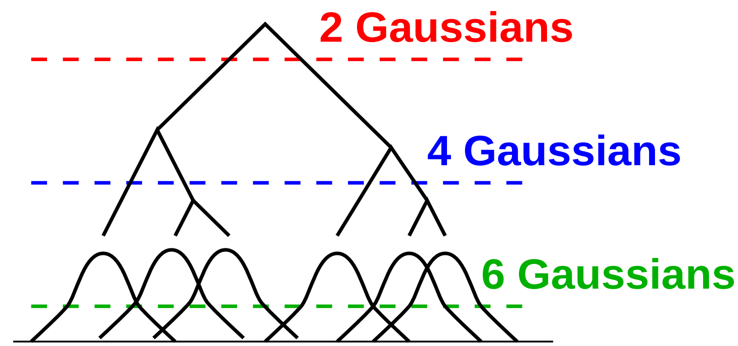
% Error	Phone Classification	Phone Recognition	Word Rec.
M=24 N=1	22.1	29.3	4.5
M=6 N=4	20.7	28.4	4.2
M=1 N=24	20.2	28.1	4.0
% Reduction	8.3	4.0	12.0

Model Aggregation

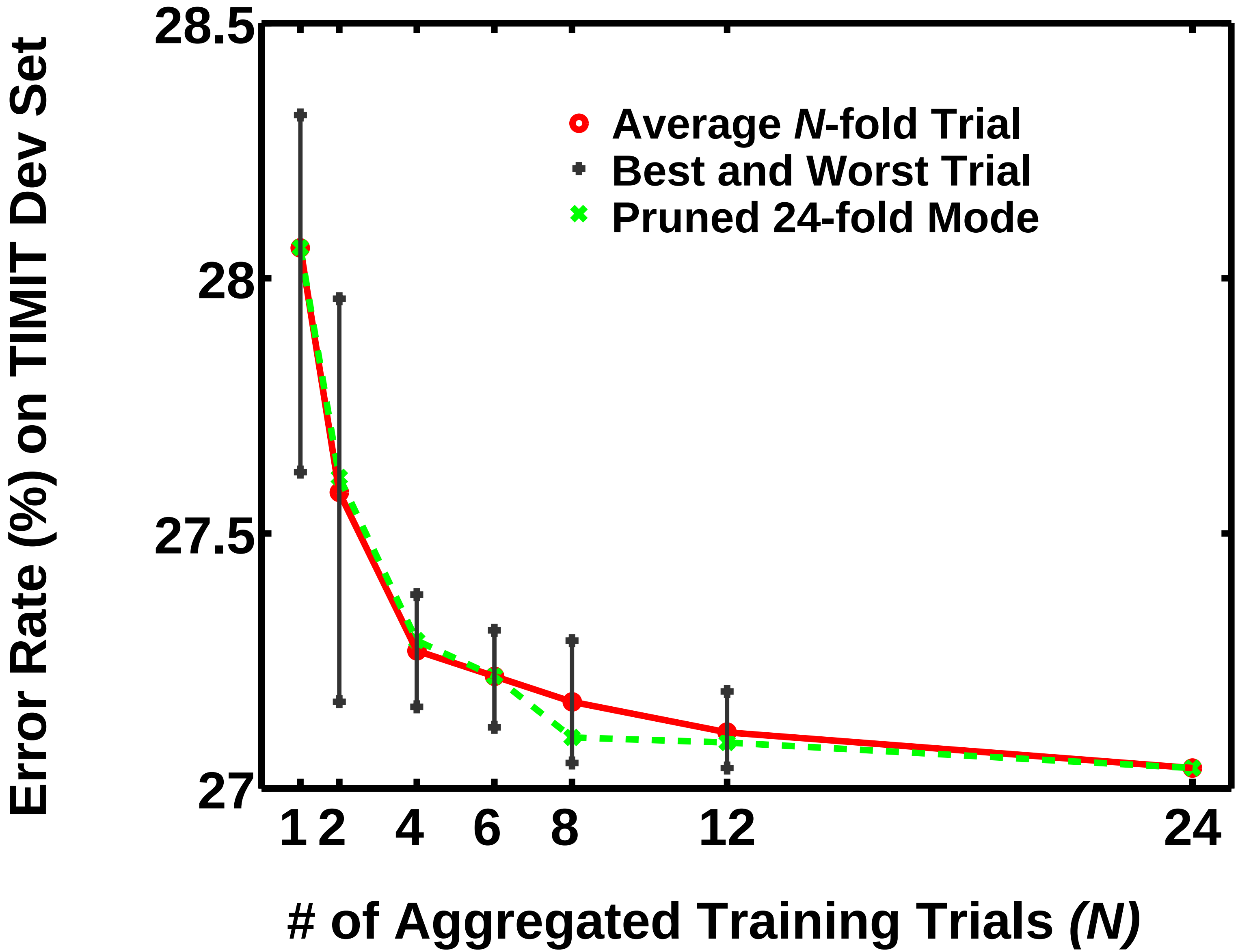
- **Aggregation combines *N* classifiers, with equal weighting, to form one aggregate classifier**

$$\varphi_A(\vec{x}) = \frac{1}{N}\sum_{n=1}^{N}\varphi_n(\vec{x})$$

- **The expected error of an aggregate classifier is less than the expected error of any randomly chosen constituent**
- ***N*-fold aggregate classifier has *N* times more computation**
- **Gaussian kernels of aggregate model can be hierarchically clustered and selectively pruned**
  - Experiment: Prune 24-fold model back to size of smaller *N*-fold models

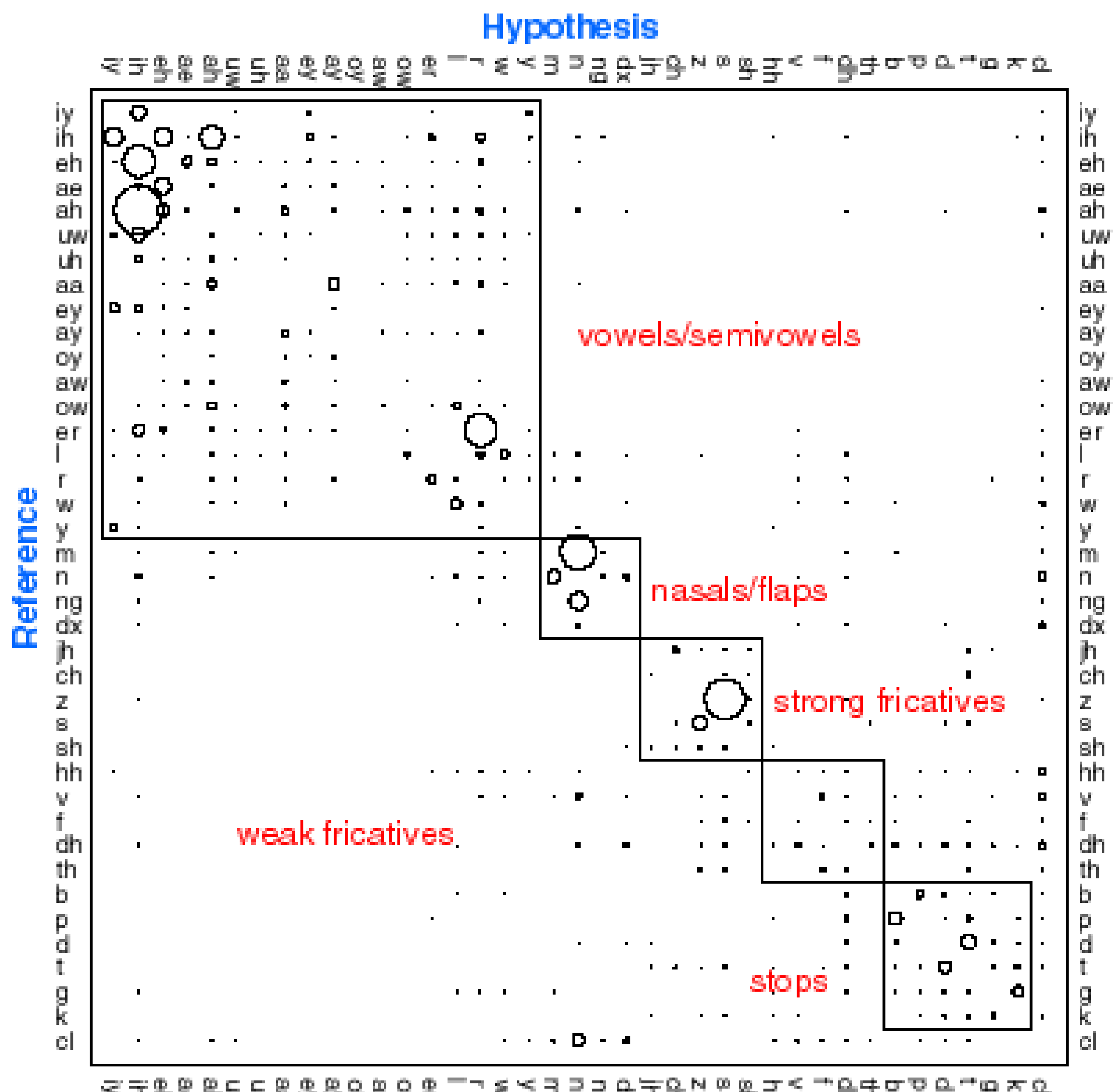


Aggregation Experiments



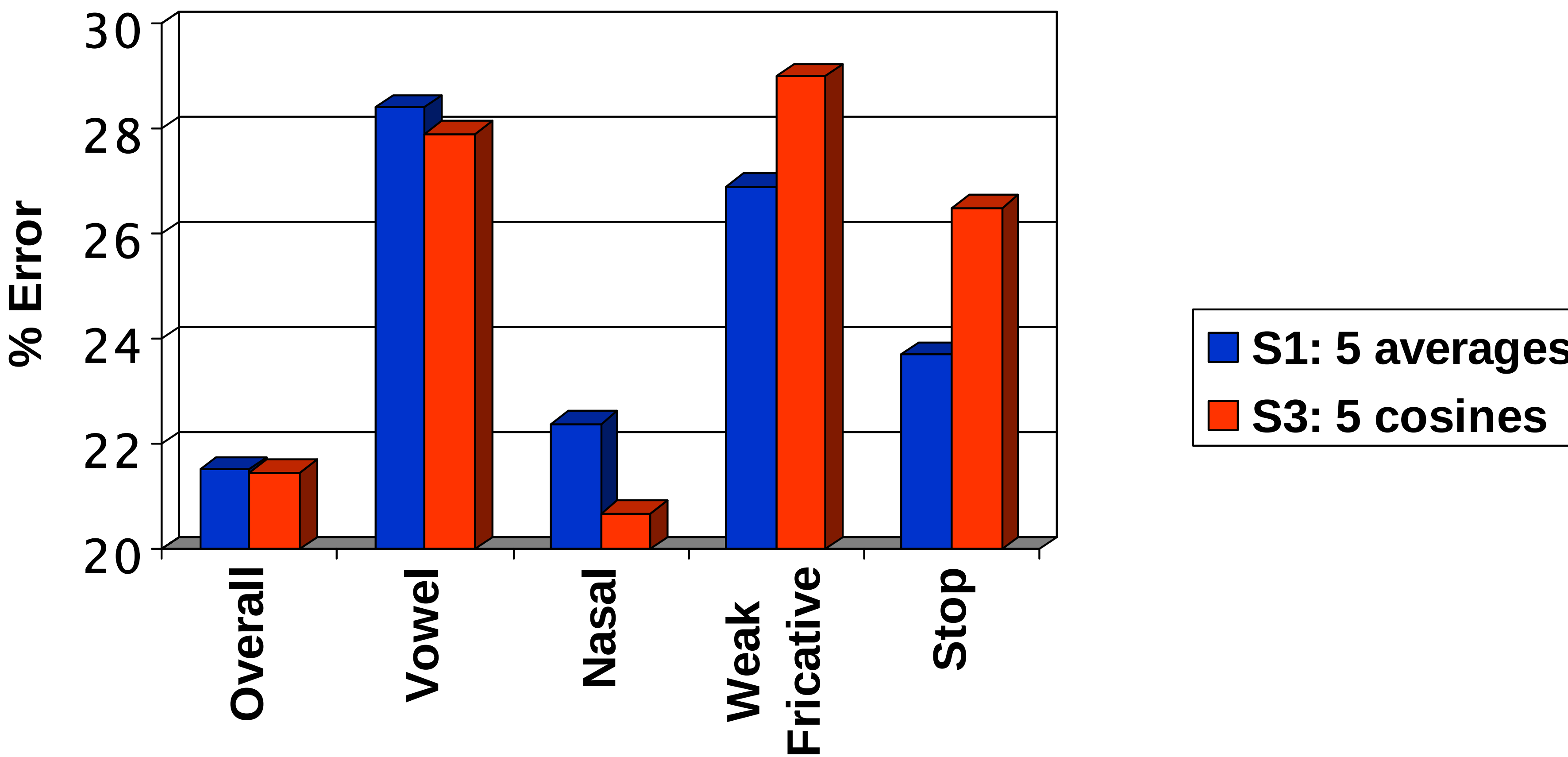
Phonetic Classification Confusions

- **Most confusions occur within manner class**



Committee-based Classification

- Change of temporal basis affects within-class error
  - Smoothly varying cosine basis better for vowels and nasals
  - Piecewise-constant basis better for fricatives and stops



- Combining information sources can reduce error

Committee-based Classifiers (Halberstadt, 1998)

- **Uses multiple acoustic feature vectors and classifiers to incorporate different sources of information**
- **Explored 3 combination methods (e.g., voting, linear, indep.)**
- **Obtains state-of-the-art phonetic classification and recognition results (TIMIT)**
- **Combining 3 boundary models in Jupiter weather domain**
  - **Word error rate 10-16% relative reduction over baseline**
  - **Substitution error rate 14-20% relative reduction over baseline**

Acoustic Measurements	% Error	% Sub
B1 (30 ms, 12 MFCC, telescoping avg)	11.3	6.4
B2 (30 ms, 12 MFCC+ZC+E+LFE, 4 cos±50ms)	12.0	6.7
B3 (10ms, 12 MFCC, 5 cos±75ms)	12.1	6.9
B1 + B2 + B3	10.1	5.5

Related Work

- **ROVER system developed at NIST [Fiscus, 1997]**
  - **1997 LVCSR Hub-5E Benchmark test**
  - **"Recognizer output voting error reduction"**
  - **Combines confidence-tagged word recognition output from multiple recognizers**
  - **Produced 12.5% relative reduction in WER**
- **Notion of combining multiple information sources**
  - **Syllable-based and word-based [Wu, Morgan et al, 1998]**
  - **Different phonetic inventories [AT&T]**
  - **80, 100, or 125 frames per second [BBN]**
  - **Triphone and quinphone [HTK]**
  - **Subband-based speech recognition [Bourland, Dupont, 1997]**

References

- **E. Bocchieri. Vector quantization for the efficient computation of continuous density likelihoods. *Proc. ICASSP*, 1993.**
- **T. Hazen and A. Halberstadt. Using aggregation to improve the performance of mixture Gaussian acoustic models. *Proc. ICASSP*, 1998.**
- **J. Glass, T. Hazen, and L. Hetherington. Real-time telephone-based speech recognition in the Jupiter domain. *Proc. ICASSP*, 1999.**
- **A. Halberstadt. Heterogeneous acoustic measurements and multiple classifiers for speech recognition. Ph.D. Thesis, MIT, 1998.**
- **T. Watanabe et al. Speech recognition using tree-structured probability density function. *Proc. ICSLP*, 1994.**