Dialectal Chinese Speech Recognition

Richard Sproat, *University of Illinois at Urbana-Champaign*
Thomas Fang Zheng, *Tsinghua University*
Liang Gu, *IBM*
Jing Li, *Tsinghua University*
Yi Su, *Johns Hopkins University*
Yanli Zheng, *University of Illinois at Urbana-Champaign*
Haolang Zhou, *Johns Hopkins University*
Philip Bramsen, *MIT*
David Kirsch, *Lehigh University*
Izhak Shafran, *Johns Hopkins University*
Stavros Tsakalidis, *Johns Hopkins University*
Dan Jurafsky, *Stanford University*

Closing Day Presentation, August 16, 2004

Recognition of Accented Speech

- A crucial ASR task
- The world is ever more globalized
  - More people speak foreign languages (English, Spanish) for economic reasons, immigration, etc.
- Arabic and Chinese are key languages for ASR, and have many dialects/accents
- Accent is hard for current ASR paradigm
  - Test speech very different than training speech
  - Too expensive to collect training data on every accent of every language

Chinese Accented Speech

- Chinese is a key language for ASR
- But most of China speaks Chinese with an accent
- National language of China is a dialect of Mandarin called Putonghua = ‘Common Language’
- Chinese comprises 7 distinct language groups:
  - Mandarin, Yue (includes Cantonese), Min (Fujianese Taiwanese), Wu (includes Shanghainese), Xiang, Gan, Hakka
- Speakers of these languages speak Putonghua (Mandarin) with an accent.

Wu-Accented Putonghua

- Our project goal: improve recognition of accented Chinese
- We chose one particular accent: people from Shanghai whose native language is Shanghainese
- Shanghainese is one of the Wu languages
- Wu is the largest language in China besides Mandarin (81 million speakers)
- Wu is very different from Putonghua (Mandarin)
- So many of those 81 million Wu speakers have very strong accents

Wu-accented Putonghua (Mandarin)

- This is the Wu region
- It includes important cities like Shanghai
- And important dialects like Shanghainese
- Our project: recognizing Putonghua (Mandarin) spoken by people whose first language is Wu: **Wu-Dialectal Chinese.**

Wu vs. Putonghua vs. Wu-Accented Putonghua

Wu vs. PTH

"*There are over 1200 students.*"

PTH vs. Wu-Accented PTH

"*Hua Temple --- Longhua Temple, how did it come about, right? I, that is, I saw a story that is often told about this.*"

How to adapt to accented speech? Previous Work

- Training on accented speech
- Acoustic model (AM) adaptation
- Lexicon adaptation (pronunciation modeling)

Training on Accented Speech

- Ikeno *et al* (2003) *Spanish-accented test set*
  - Train on 100 hours native English: 68.5%
  - Train on 20 hours accented English: 39.2%

  } 20 hours

- Tomokiyo & Waibel (2001) *Japanese-accented*
  - Train on native English speakers: 63%
  - Pooled with 3 hours accented English: 53%

  } 3 hours

- Wang *et al* (2003) *German-accented test set*
  - Train on 34 hours native English: 49.3%
  - Train on 52 minutes accented English: 43.5%
  - Train on both: 42.3%

  } <1 hour

Acoustic Adaptation - MLLR

- MLLR: standard unsupervised speaker adaptation technique; learns a transform for gaussians
- Huang, Chang, Zhou (2000) *Shanghai-accented test*
  - No MLLR 23.18%
  - Used MLLR on individual test speakers: 21.48%
- Tomokiyo & Waibel (2001) *Japanese accented*
  - MLLR on individual test speaker: 63%
  - MLLR on 3 test speakers: 58%
  - MLLR on 15 test speakers: 53%
- Wang et al (2003) *German accented*
  - No MLLR 49.5%
  - 7 minutes MLLR on 64 speakers 46.8%
  - 50 minutes MLLR on 64 speakers 44.0%

Acoustic Adaptation-MAP

- Wang et al (2003) *German accented*
  - No MLLR 49.5%
  - 50 minutes MLLR on 64 speakers 44.0%
  - 50 minutes MAP on 64 speakers 38.0%

Acoustic Adaptation: Summary

- MLLR on multiple speakers useful
- Previous multispeaker MLLR only used single transform
- MAP better than MLLR with enough data
- No previous work on combining MAP and MLLR on accented data
- Suggests the following plan for our work:
  - Try more complex use of MLLR
  - Combine MLLR and MAP

Lexicon Adaptation: Standard Approach

- Create rules/CARTs to add pronunciation variants.
  - Hand-written rules or
  - Rules induced from phonetically transcribed data
- Use rules to expand lexicon
- Force-align lexicon with training set to learn pronunciation probabilities.
- Prune to small number of pronunciations/word.

Cohen 1989; Riley 1989, 1991; Tajchman, Fosler, Jurafsky 1995; Riley et al 1998; Humphries and Woodland 1998, inter alia

Lexicon Adaptation: Problems

- Limited success on dialect adaptation:
  - Mayfield Tomokiyo 2001 on Japanese-accented English: *no WER reduction*
  - Huang et al. 2000 on Southern Mandarin: *1% WER reduction over MLLR*
- Probable main problems:
  - Most gain already captured by triphones and MLLR
  - Speakers vary widely in their amount of accent so *dialect*-specific lexicons are insufficient

Project Goals

- Explore techniques for improving recognition of accented speech
  - Better acoustic model adaptation
  - Better lexicon (pronunciation model) adaptation
- Demonstrate that “accentedness” is a matter of degree, and should be modeled as such.
  - Automatic detection of accent severity
  - Dynamically adjust acoustic model based on accent detection

Overview

- Data Collection: Wu-accented PTH Data
- Analysis of Wu-accented Data
- Baselines/Oracles
- Pronunciation Modeling: IF Mapping experiments
- Automatic Age/Accentedness Detection
  - Using speaker clusters
- New Models of Acoustic Adaptation
- Dynamically adjusted acoustic model based on accent detection
- *Minimal Perplexity Word Segmentation*
- Implications and Future Work

Our new corpus

- Wu-Dialectal Chinese Corpus
  - –100 native Shanghai speakers
  - –~5 minutes spontaneous speech, 3 minutes read speech per speaker
  - –Total: 13 hours of accented broadband speech
- –Standard Chinese Corpus
  - –Matched for domain
  - –20 standard Chinese speakers
  - –6 minutes spontaneous speech per speaker

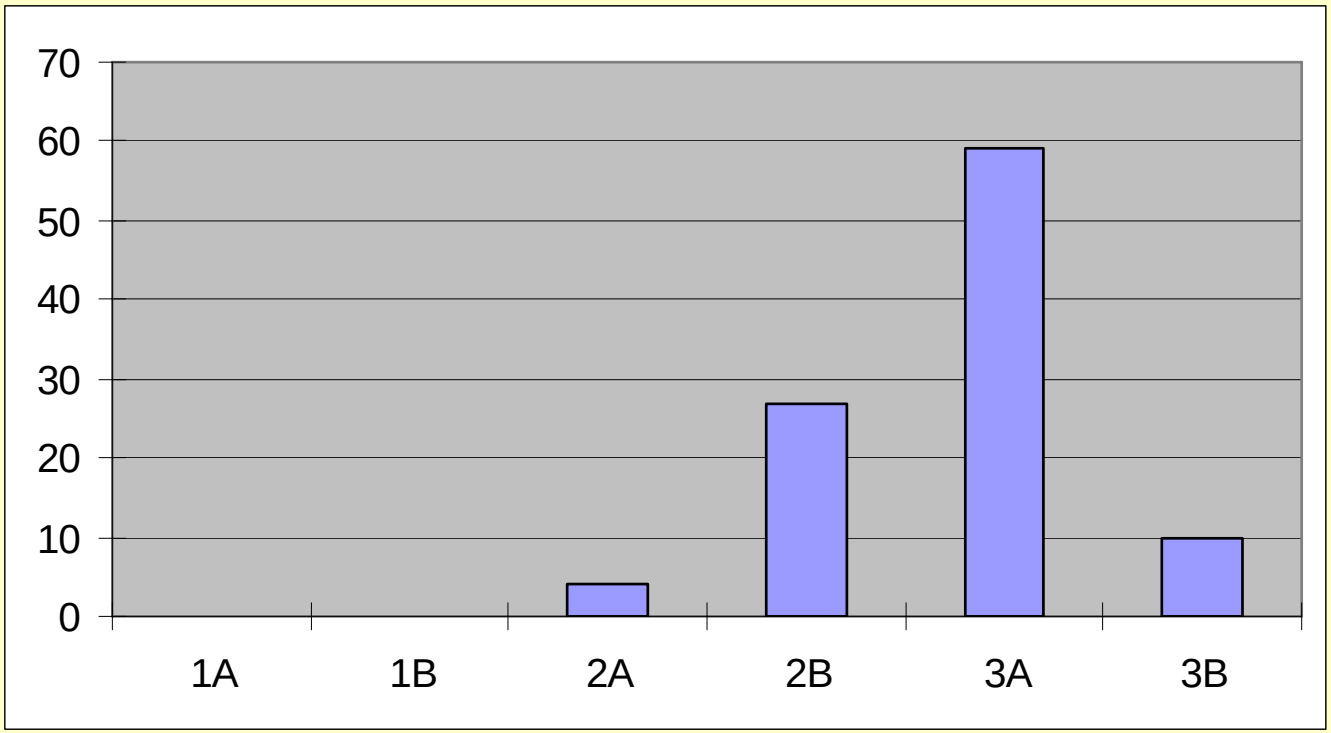
Speakers' Age Distribution

Num of speakers		Male	Female	Total
Age	26-40	27	25	52
	41-50	23	25	48

Speakers' Education Levels

Num of speakers		Male	Female	Total
Education	High	41	41	82
	Low	9	9	18

Accent Assessment

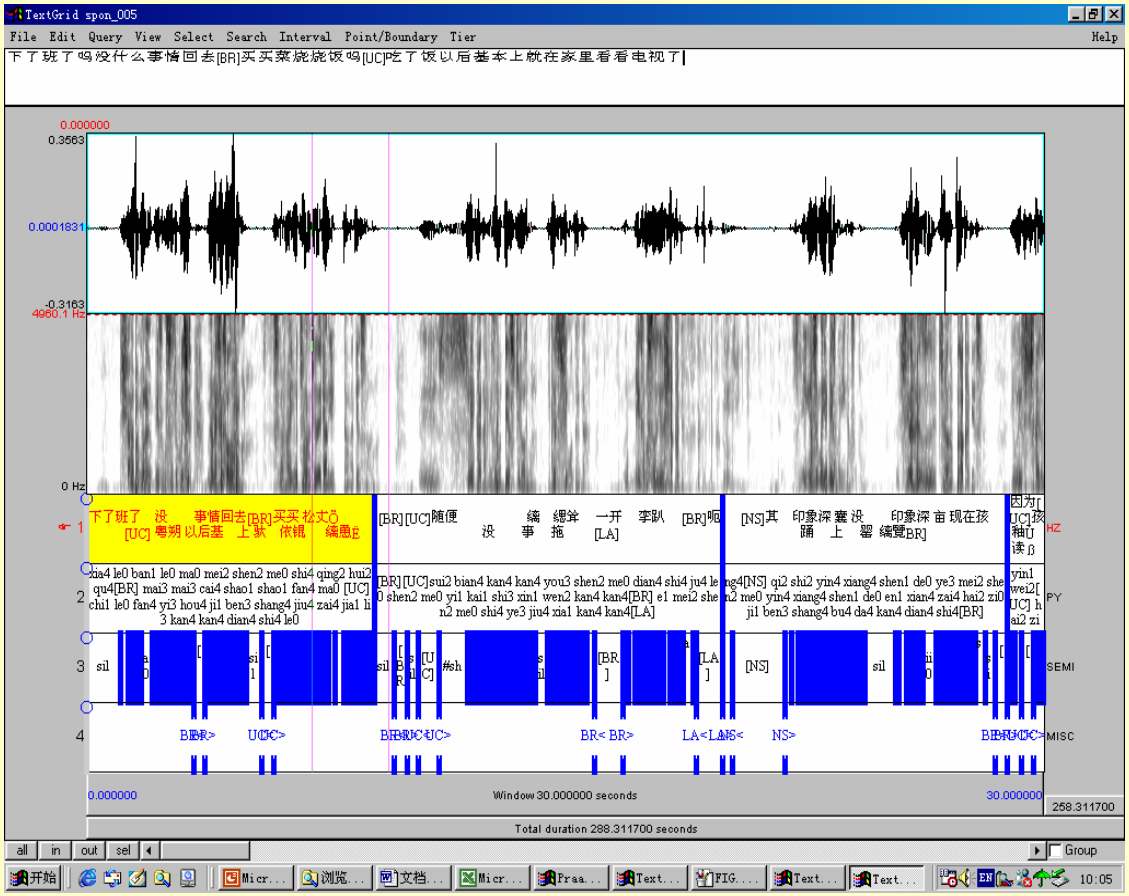


1A. State-level radio broadcaster; 1B. Province-level radio broadcaster; 2A. Quite good; 2B. Less accented; 3A. More accented; 3B. Hard to understand but know it is PTH

Data Annotation

- Orthographic Tier
- Canonical Pinyin Tier
- Surface Initial-Final (IF) Tier:
  - *In Chinese ASR, people typically model at the level of Initial-Finals, rather than phones*
- Miscellaneous Non-Speech Tier

Data Annotation Using Praat



Dialectal Chinese Speech Recognition

Annotation Example

hua2 si4 long2 hua2 si4 zen3 me0 dai4 lai2 de0 dui4 ba0

hua2 shiii4 loong2 hua2 sii4 zen3 me0 dai4 lai2 de0 dui4 b_va0

wo3 na4 shi4 wo3 kan4 le0 yi1 pian1 shi4 chang2 shuo1 de0 jiu4 shi4 ha0

uo3 na4 sii4 uo3 kan4 le0 i14 pian1 sii4 cang2 suo1 d_ve0 jiuu4 sii4 h_va0

Data Division

- Data divided into 80 training speakers and 20 test speakers.
- 20 test speakers were balanced for gender and accentedness

Specifics on Test Speakers

Speaker ID	Gender	Age	Education	PTH	Fluency	Rec. Loc.	Rec. Mi
008	Male	26	UG	2A	Yes	1	2
009	Male	35	UG	2B	Yes	1	1
011	Male	30	UG	2B	Yes	2.1	2
012	Male	34	UG	2B	Yes	2.1	1
016	Male	26	JC	2B	Yes	2.2	1
032	Male	34	UG	3B	Yes	4	1
035	Male	36	JC	3B	Yes	2.2	1
043	Male	44	UG	3A	Yes	4	1
046	Male	50	SHS	3A	Yes	2.1	2
047	Male	40	SHS	3A	No	3	1
053	Female	45	TSS	3B	Yes	3	1
054	Female	45	TSS	2B	Yes	3	2
059	Female	40	SHS	3B	Yes	1	1
061	Female	30	UG	2B	Yes	2.1	2
064	Female	34	UG	2B	Yes	2.1	1
066	Female	26	UG	2A	Yes	2.1	2
067	Female	33	UG	2A	Yes	2.1	2
076	Female	41	UG	3B	Yes	1	1
098	Female	41	SHS	3A	Yes	3	1
099	Female	41	JC	3B	Yes	3	2

Pronunciation differences in Wu-accented Putonghua

- **Standard [sh] is pronounced [s]**
- **Standard [ch] is pronounced [c]**
- **Standard [zh] is pronounced [z]**
- **Standard [ing] and [in] are interchangeable**
- **Standard [eng] is pronounced [en]**

Standard	Shanghai PTH	
shan	san	mountain
chan	can	cicada
zhuozi	zuozi	table

Factors influencing Wu accent

(with Rebecca Starr, Stanford)

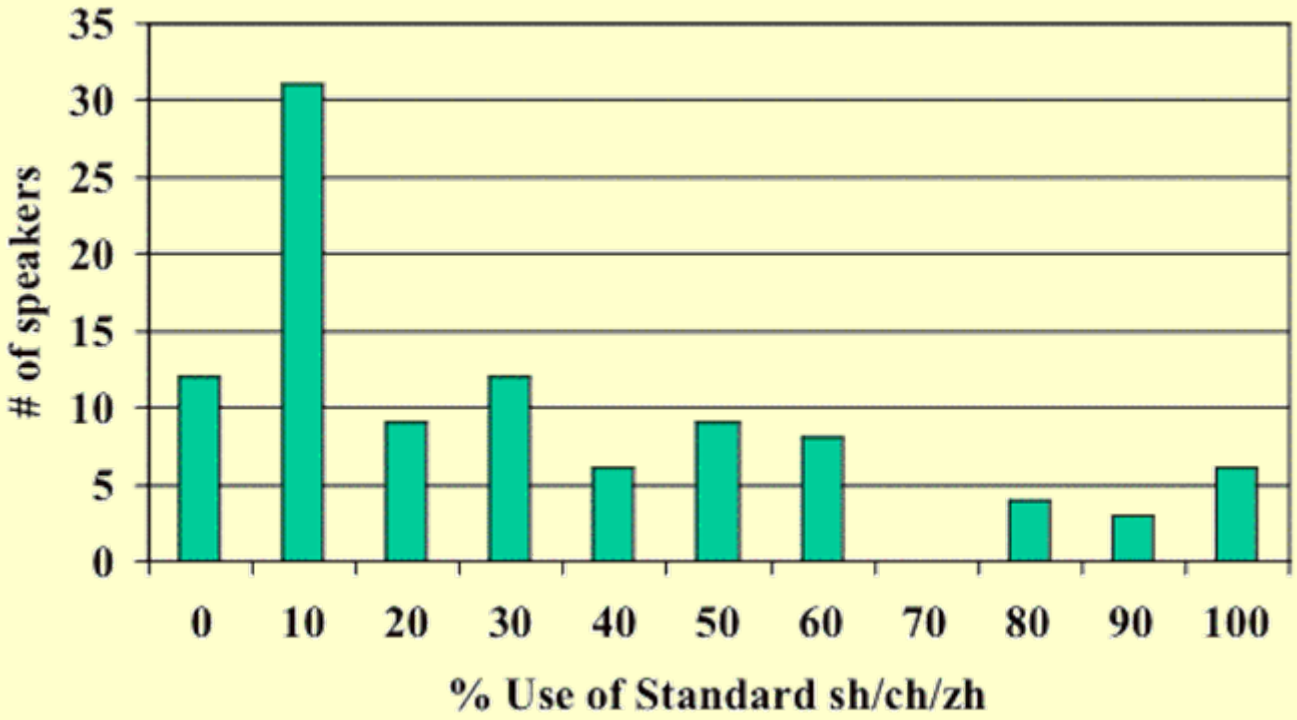
- We examined every **sh, zh, ch** in our corpus

  19,662 tokens of **sh/zh/ch,** coded for
  - Did they turn into s/z/c?
  - Age
  - Gender
  - Education
  - Phone (**sh, zh, ch**)
  - Phonetic context
- Logistic Regression

Results from phonological analysis

1. **Massive variation between speakers**
   - 0%-100% use of standard pronunciation

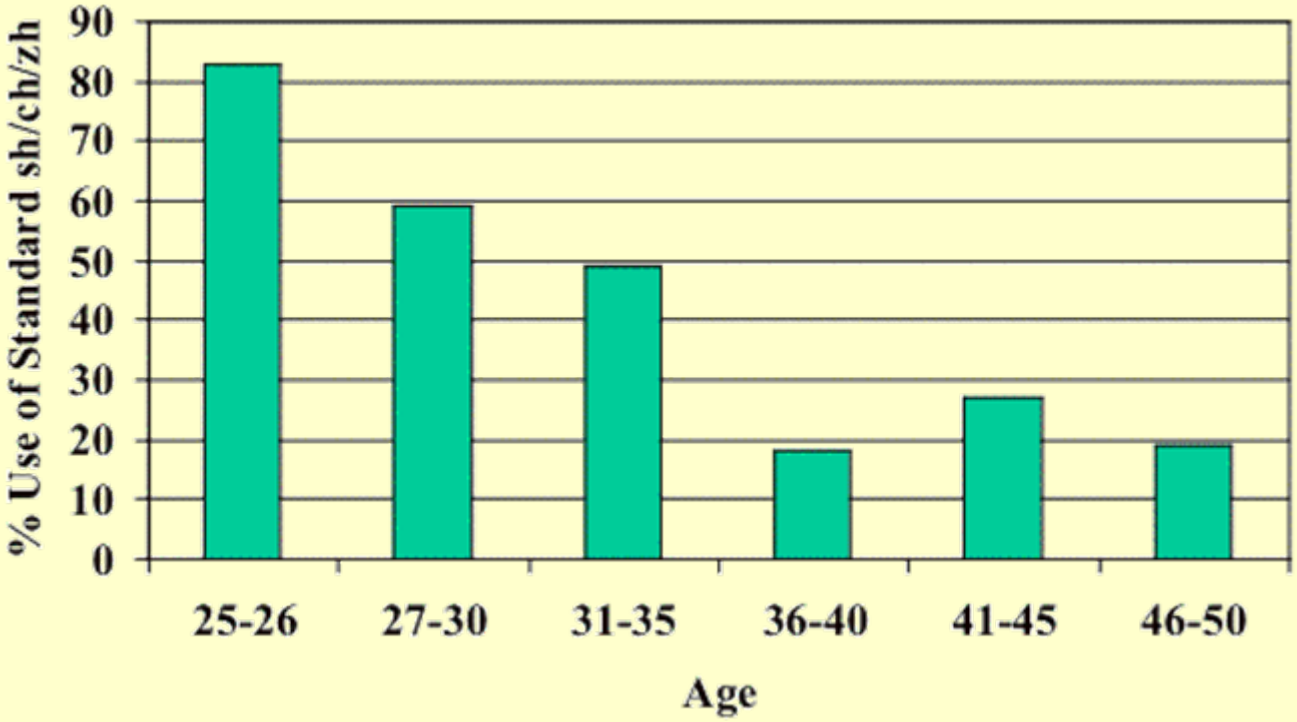
Massive variation among speakers



Results from phonological analysis

1. **Massive variation between speakers**
   - 0%-100% use of standard pronunciation
2. **Age** and education are predictors of more standard speech
   - Younger speakers are more standard

Younger speakers more standard



Results from phonological analysis

1. **Massive variation between speakers**
   - 0%-100% use of standard pronunciation
2. **Age** and education are predictors of more standard speech
   - Younger speakers are more standard
3. Percentage of sh versus s correlates with other indicators of accent:
   - **The more [s], the more accented**
   - The more [sh], the more standard

Conclusions from Analysis

- Massive variation in severity between speakers:
  - **Accent modeling needs to be continuous not binary: need to model accent severity**
- Age and education predict standard speech:
  - **Can use age-type features to predict accent severity**
- The more [s], the more accented:
  - **Can use count of [s] & [sh] to predict accent severity**
- Clear phonological characteristics of accent in sh/ch/zh/ng
  - **Lexical adaptation/pronunciation modeling seems good bet**

Baseline Experiments

- Language model: Consistent across all conditions
- Acoustic Models:
  - Mandarin Broadcast News (MBN)
  - Wu Devtrain

Language Model

- Training data:
  - Mandarin HUB5 (200 telephone conversations of up to 30 minutes each)
  - 100 hours of conversational Putonghua collected by HKUST
  - 6.3 hours Wu-accented devtrain data
- Dictionary: 50,500+ word dictionary from Tsinghua University

Language Model

- Text segmented using *maximal matching* with the dictionary
- Word bigram language model with Katz backoff built using AT&T GRM and FSM tools

Baseline Acoustic Model Datasets

- Mandarin Broadcast News (MBN):
  - 30 hours
  - Wideband recordings
  - Mostly professional speakers
- Wu accented training data (WUDEVTRAIN):
  - 6.3 hours
  - Wideband recordings
  - In domain

Specifics of Acoustic Models and Decoding

- Standard 39 dimensional MFCC
- 14 GMM per state
- Acoustic models constructed using HTK 3.2
  - Convert to AT&T BLASR format
- Decoding used AT&T *drecog*

Baseline Results

- Results for beam of 14 and grammar weight of 14.

AM Training	CER
MBN	61%
WUDEVTRAIN	44.2%

Lexicon Oracle

- How much gain can one generally expect from *pronunciation modeling*?
  - If one knew exactly which pronunciation(s) a test speaker would use for a word this is already better than what could be hoped for
  - Optimize these pronunciations for the given acoustic model with forced alignment

Lexicon Oracle

- Alter the dictionary to allow alternate pronunciations for some sounds:
  sh→s, zh→z, ch→z, in→ing
- Force align the dictionary on each test speaker
- Choose single most common pronunciation for each word

Lexicon Oracle

Speaker	Oracle CER	Baseline CER
008:	63.9	63.9
009:	62.8	63.8
011:	65.3	70.1
012:	59.0	58.9
016:	67.7	67.7
032:	45.3	48.1
035:	57.9	59.3
043:	57.2	58.6
046:	70.1	71.0
047:	71.7	72.2
053:	81.2	84.3
054:	59.7	59.8
059:	66.4	71.8
061:	50.7	51.6
064:	39.7	40.0
066:	48.6	49.7
067:	50.9	50.9
076:	49.4	50.5
098:	73.1	75.1
099:	70.2	73.6
Total:	59.6	61.0

- Only a 1.4% gain overall even with speaker-specific lexicons
- Suggests that gains from lexicon modification will not come easily
- But perhaps there are more sophisticated methods

Pronunciation Modeling: IF Mapping Experiments

Presenter: Thomas Zheng

Project Goal as Proposed

- To develop a general framework to model phonetic variability, pronunciation variability and lexical variability in dialectal Chinese ASR tasks.
- To find suitable methods to modify PTH recognizer so as to obtain a dialectal Chinese recognizer for the specific dialect of interest, which employ :-
  - Dialect-related knowledge, and
  - Training data (in relatively small quantities, or even no)
- Expectation: the recognizer should also work for PTH, in other words, it should be good for a mixture of PTH and dialectal Chinese.

Observation on WDC Data

- IF-mapping / Syllable-mapping:
  - Influenced by Wu dialect, a Wu dialectal Chinese (WDC) speaker often pronounce any of a certain set of IFs into another IF, and there are rules to follow, such as *zh* -> *z*, *ch* -> *c*, *sh* -> *s*, and so on.

- Observations on three sets - *train* (80 speakers), *devtest* (20), and *test* (20):
  - Mapping pairs almost the same among all three sets;
  - Mapping pairs almost identical to experts' knowledge;
  - Mapping probabilities also almost equal;

- Remarks:
  - Experts' knowledge could be useful;
  - Mapping rules can be learned from less data.

Workshop Experiment

- A total different roadmap
  - Using HTK 3.2.1 (latest version downloadable on web)
  - Using only 20 speakers' data + dialect-based knowledge

- Step 1: Apply PTH-IF mapping rules;
- Step 2: Apply WDC-IF mapping rules;
- Step 3: Apply syllable-dependent mapping rules;
- Step 4: Perform multi-pronunciation expansion (MPE) based on unigram probability;
- Step 5: Perform rank-based AM rescoring.

- Why trying this method?
  - "IF-mapping" in dialectal Chinese is the fact (human uses it);
  - "In-domain data training" will sure get a good result but collecting data is a huge task, especially for 40 sub-dialects of Chinese;
  - "Mere Adaptation" will be easier and better but might make it hard to distinguish those mapping pairs, each pair tends to become a single IF;
  - This is not practical in such applications where you have no more information about the speakers and a mixture of WDC and PTH is used as Call Centers;
  - It is expected that knowledge based method would result in an overall good performance for both WDC and PTH.

- Step 1: Applying PTH-IF mapping rules
  - Rules are based on experts' knowledge (with AM unchanged)
    - (zh, z) (z, zh)
    - (ch, c) (c, ch)
    - (sh, s) (s, sh)
    - (eng, en) (en, eng)
    - (ing, in) (in, ing)
    - (r, l)
  - Gain not so significant: 0.5% CER reduction
  - Pronunciation entry probability does not help improve performance

- Step 2: Applying WDC-IF mapping rules
  - There indeed are some Wu dialect Chinese specific IFs, such as *iao* -> *io^*;
  - Rules learned from *devTest*
  - Newly introduced WDC specific IFs trained from *devTest* using adaptation method
  - 8.66% absolute CER reduction
  - MLLR adaptation outperforms MLLR+MAP
    - About 10% difference
    - Possibly due to less data

- Step 3: Apply syllable-dependent mapping rules
  - Assumption: most IF-mappings are context-independent, but some are syllable-dependent (such as *iii*|(*sh iii*) -> *ii*|(*s ii*)), we believe there are others
  - Rules learned from *devTest*
  - We do not succeed in improving the accuracy, on the contrary, the character accuracy reduced by about 6%
  - We do not have a clear explanation yet
  - So we keep using context-free mapping rules

- Step 4: Multi-pronunciation expansion (MPE) based on unigram probability
  - Motivation: more pronunciations help model pron. variations, but lead to more confusion, there should be tradeoff;
  - Accumulated unigram probability (***AccProb***) used as the criterion
    - Only words with higher unigram probabilities will have multiple pronunciations each;
    - Words with lower unigram probabilities will have a single standard pronunciation each;

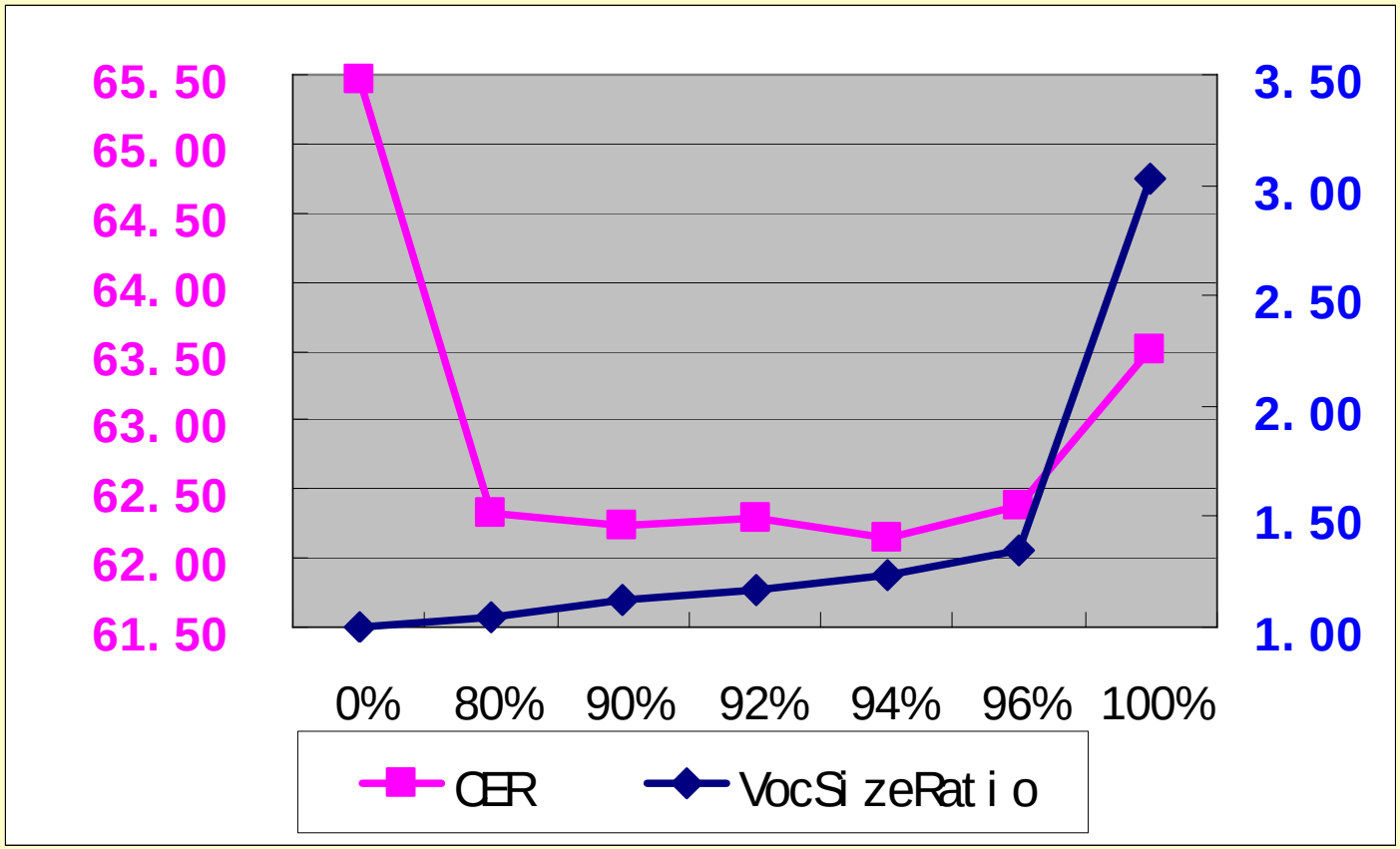
Word	Prob. (descending)	Acc. Prob.		
		0.000		
</ s>	0.10782136	0.108		
	0.03608752	0.144		
	0.02161165	0.194		
	0.01907339	0.213		
…				
	0.00005742	0.899	←	Actual minimum
…	**0.00005742**			
	0.00005742	**0.900**	←	Desired point
…	**0.00005742**			
	0.00005742	0.901	←	Actual maximum
…				
	0.00000124	1.000^{-}		
	0.00000124	1.000^{-}		

Multi-Pronunciation Expansion

Single-Pronunciation (Standard)

The Multi-Pronunciation Expansion Criterion

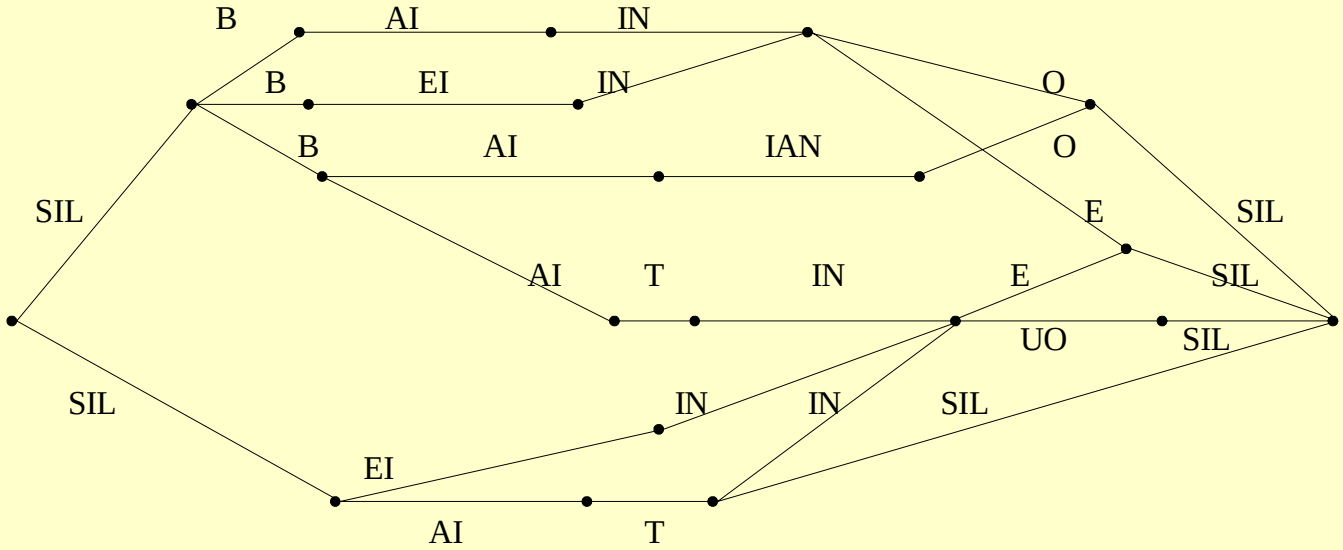
AccProb: 0% means no multiple pronunciation expansion, while 100% full expansion;



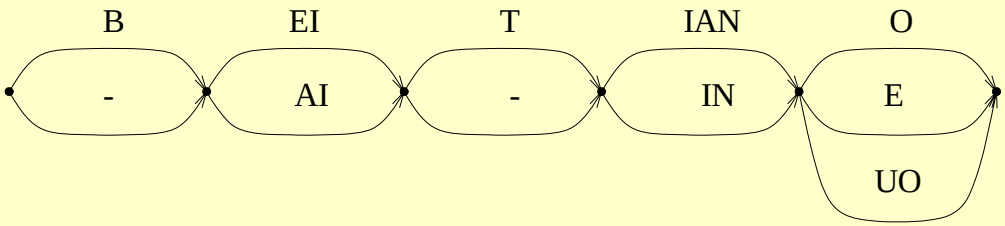
Best result achieved at a suitable *AccProb* value, say 94%, with VocSizeRatio=1.24

- Step 5: Rank-based AM Rescoring
  - Assumption: ranks in lattice when using the recognizer derived from the PTH one to recognize WDC speech has a relatively stable distribution

Generate lattice ("SIL" marks pauses) for each sentence in *devTest*



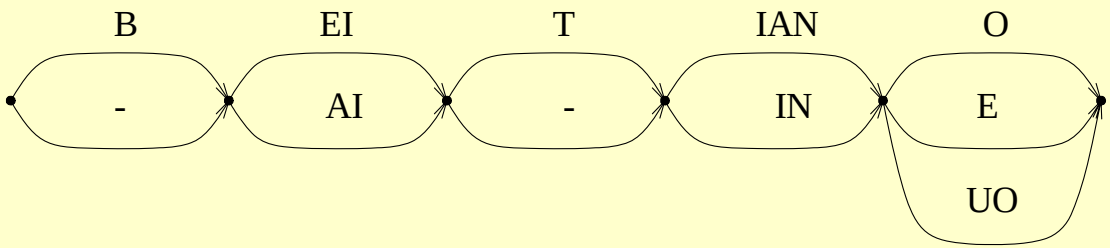
Turn the lattice into multiple alignment ("-" marks deletions) -
information of arcs in the lattice will be remembered for later back-tracking.



Lidia Mangu
et al [1999]

Learn P (*a* | *a*, *rank*): probability of *a* if seen in the *rank*-th position

Recognition:



Transcription:

B AI T IAN E

Learning:

- Count (B) ++; Count (B | B, 1)++
- Count (AI)++; Count (AI | AI, 2)++
- Count (T)++; Count (T | T, 1)++
- Count (IAN)++; Count (IAN | IAN, 1)++
- Count (E)++; Count (E | E, 2)++

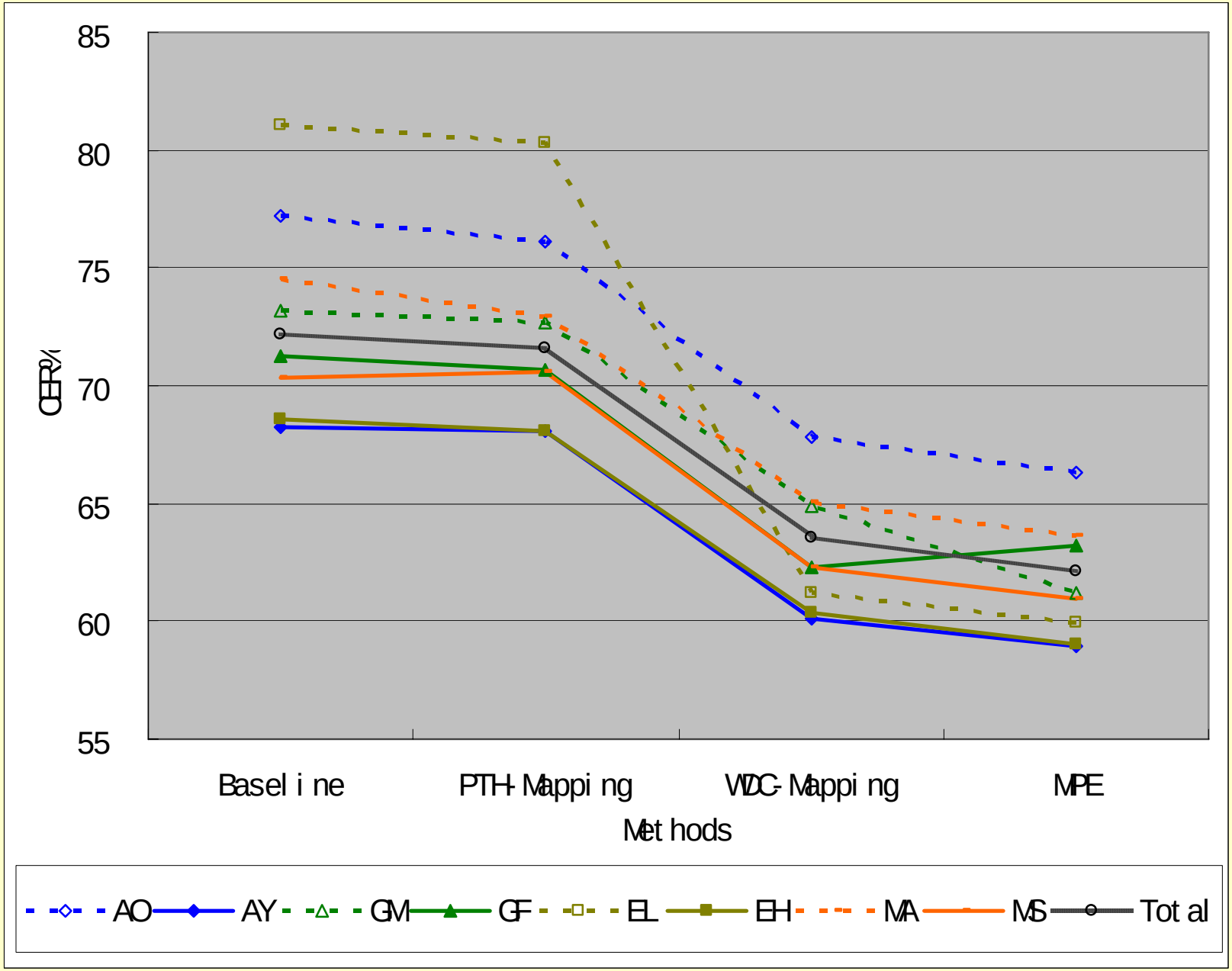
Post-processing:

Prob. (*a* | *a*, *rank*) =

Cnt (*a* | *a*, *rank*) / *Cnt* (*a*)

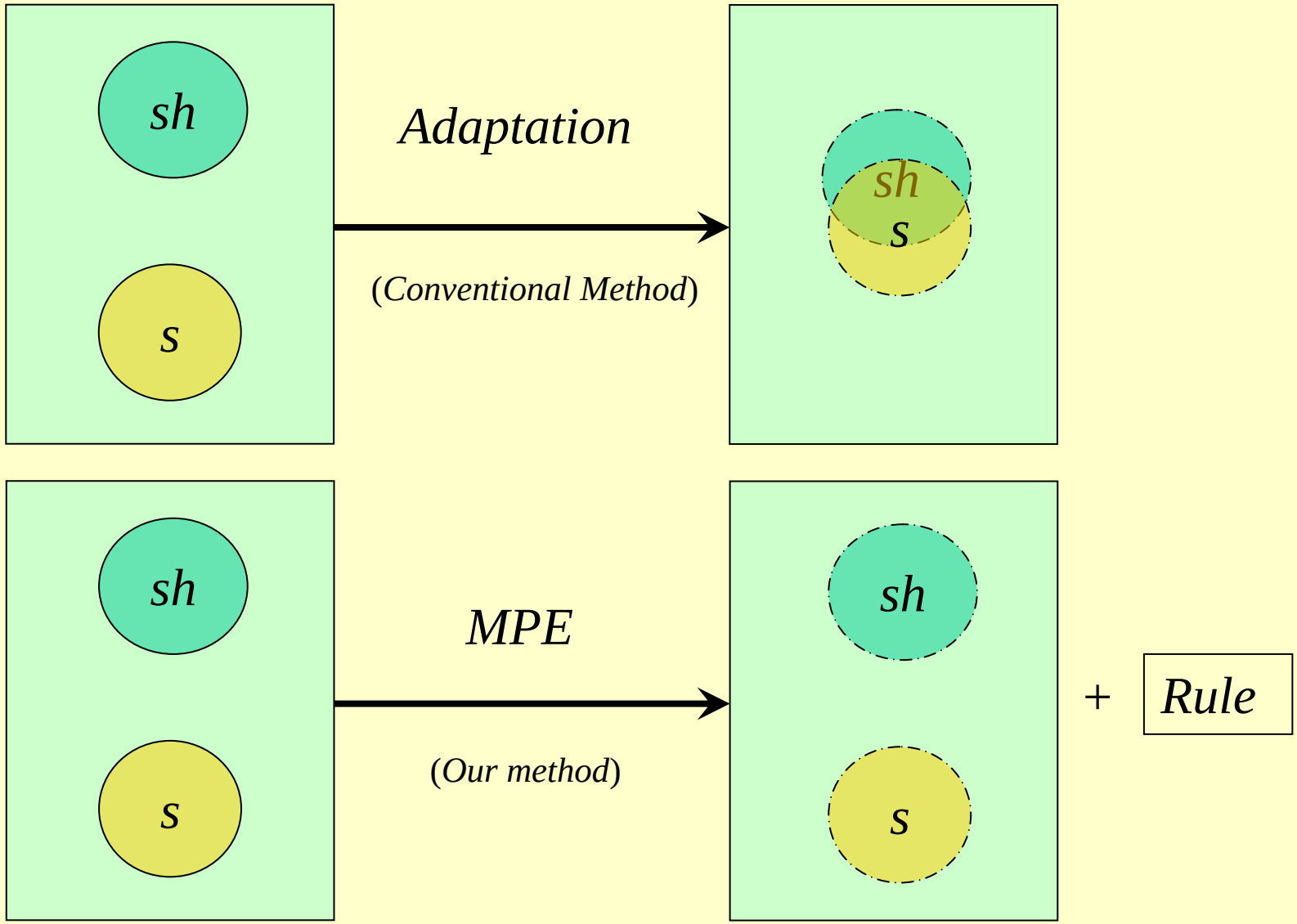
- Rescoring during recognition:
  - Original lattice
  - Multi-alignment lattice
  - Original lattice rescoring: using the ranks in this multiple alignment and the back-tracking information, modify the probability of the WDC-IF in each arc in the lattice.
- *This part is unfinished because there is not any direct way for this kind of rescoring*

Performance improvement comparison: overall, and in terms of speaker clusters



Q: Recognize PTH using WDC recognizer?

- We obtain WDC recognizer from PTH recognizer;
- We get a CER reduction of over 10% when recognizing WDC on an average;
- How about using it to recognize PTH?



- We can expect that using WDC recognizer to recognize PTH, the performance will degrade;
- But we would expect it will not decrease too much;
- Results: using WDC recognizer, you get
  - Over 10% CER reduction to recognize WDC;
  - 0.62% CER increase to recognize PTH.

Summary & Future Plan

- The use of knowledge is useful and effective
- In this project, there are several problems to solve: *channel*, *speaking-style*, *dialect background*, and *domain* problems.
  - It is easier to solve all these problems by simply using the adaptation method;
  - Our method focuses only on the *dialect* problem;
  - The results using our method could be better if we integrate those methods related to *channel*, and *speaking-style*.

- The proposed method needs much more efforts in programming and data preparation for each step because there are not existing tools to use -- this leads to low efficiency
- We choose to use HTK because we can continue using it in post-workshop experiments

- Continue on the current project, including:
  - Investigating the syllable-dependent mapping;
  - Rank-based Rescoring
- Language Model Adaptation
  - Different word form with same meaning
    - Such as: vs. - *like*; vs. - *cook*
    - Linguists say the vocabulary similarity rate between Putonghua and *Wu* dialect is about 60~70%.
  - Different word order
    - (*you first go*) vs. (*you go first*)

Unsupervised Continuous Lexicon Adaptation

David W. Kirsch

Lehigh University

CLSP – 16 August 2004

Dialectal Chinese Speech Recognition

Pronunciation Lexicon

- W_1: $p_{1,1}$ $p_{1,2}$ $p_{1,3}$
- W_2: $p_{2,1}$ $p_{2,2}$
- W_3: $p_{3,1}$ $p_{3,2}$ $p_{3,3}$
- W_4: $p_{4,1}$

- Keeps track of likely word pronunciations
- Produced from linguistic knowledge or observation
- Static lexicon bad for accented speakers
- Need to adapt based on accent

Past Approaches

- Train on speaker
- Hand-construct pronunciation rules
- Cluster speakers

- Often impractical or impossible
- Requires expert domain knowledge
- Accented-ness is continuous, not discrete

Proposed Approach: Criteria

- Continuously adaptable models
- Continuously classified speakers
- Adapt to new speakers without supervision

Proposed Approach: Methods

- Continuous speaker classification:

  Detect phoneme ratios

- Continuous speaker models

  Reweight pronunciation lexicon

Phoneme Substitution

- /sh/->/s/
- /zh/->/z/
- /ch/->/c/
- /eng/->/en/
- /ing/->/in/

- Important aspect of Wu-accented speech
- Strongly tied to other accentual features
- Easily observable
- Continuous

Phoneme Ratios

- Substitution /sh/->/s/
- Ratio count(sh)/count(s)

0----------1----------→

^StrongAccent Standard^ ^Hypercorrect

Detecting Ratios

- 80% confidence in ratio with 20 phones
- 20 phones:

  Sh/s: 5 sentences

  Ng/n: 5 sentences

  Zh/z: 7 sentences

  Ch/z: 10 sentences
- Reasonable for many ASR tasks

Reweighting

Given a word with 3 pronunciation counts:
W: count(shi) count(si) count(i)

Transform counts to:
W: count(shi) / X count(si) * X count(i)

- Need to restore original probability mass of altered frequencies

$$X^2 = \frac{\frac{spkr.count(s)}{spkr.count(sh)}}{\frac{all.count(s)}{all.count(sh)}}$$

- Reweight according to speaker's phone ratio
- Normalize with respect to training set

Strengths

1) Continuity
2) Uses speaker-level information while preserving word-level information
3) Can be applied for each phoneme substitution independently
4) Can be pruned to 1-best after reweighting

Goals

- Compare to baseline:
  - ASR with single static lexicon
- Compare to discrete PM approach:
  - ASR with multiple static lexica
- Refine Algorithm
- Find similar continuous approaches to other accentual features

Advisors

Project Advisor	• Thomas Fang Zheng (Tsinghua Univ.)
Academic Advisor	• Brian Davison (Lehigh Univ.)

Accentedness Detection

Presenter: Richard Sproat

“Accentedness” Classification

- Two approaches:
  - Classify speakers by age, then use those classifications to select appropriate models
  - Do direct classification into accentedness

- The former is more interesting, but the latter seems to work better.

Age Detection

- Shafran, Riley & Mohri (2003) demonstrated age detection using GMM classifiers including MFCC's and fundamental frequency. Overall classification accuracy was 70.2% (baseline 33%)
- The AT&T work included 3 age ranges: youth (< 25), adult (25-50), senior (>50)
- Our speakers are all between 25 and 50. We divided them into two groups (<40, >=40)

Age Detection

- Train three-state HMM's with up to 80 mixtures per state on:
  - Standard 39 MFCC + energy feature file
  - The above, plus three additional features for (normalized) f0: *f0*, *Δf0*, *ΔΔf0*
  - Normalization: *f0norm = log(f0) – log(f0min)* (Ljolje, 2002)
- Use above in decoding phase to classify speaker's utterances into "older" or "younger"
- Majority assignment is assignment for speaker

Age Detection (Base = 11/20)

Train \ Test	*Spontaneous*		*Read*	
	MFCC	**MFCC+f0**	**MFCC**	**MFCC+f0**
Spontaneous	**13**	**14**	**14**	**10**
Read	**13**	**12**	**13**	**14**

Accent Detection

- Huang, Chen and Chang (2003) used MFCC-based GMM's to classify 4 varieties of accented Putonghua
- Correct identification ranged from 77.5% for Beijing speakers to 98.5% for Taiwan speakers

Accent Detection (Base = 10/20)

Train \ Test	*Spontaneous*		*Read*	
	MFCC	**MFCC+f0**	**MFCC**	**MFCC+f0**
Spontaneous	**12**	**15**	**11**	**10**
Read	**14**	**15**	**15**	**15**

Automatic Speaker Clustering Using Fronting Ratios

- Features that correlate with accentedness:

$$\frac{count(s)}{count(s)+count(sh)}$$

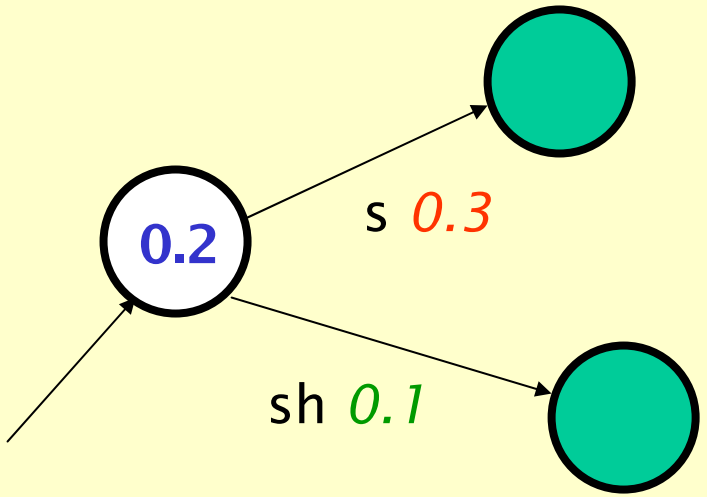
$$\frac{count(z)}{count(z)+count(zh)}$$

$$\frac{count(c)}{count(c)+count(ch)}$$

Phone Population Estimates from Decoder Lattices

- Compute continuous "counts" for phones

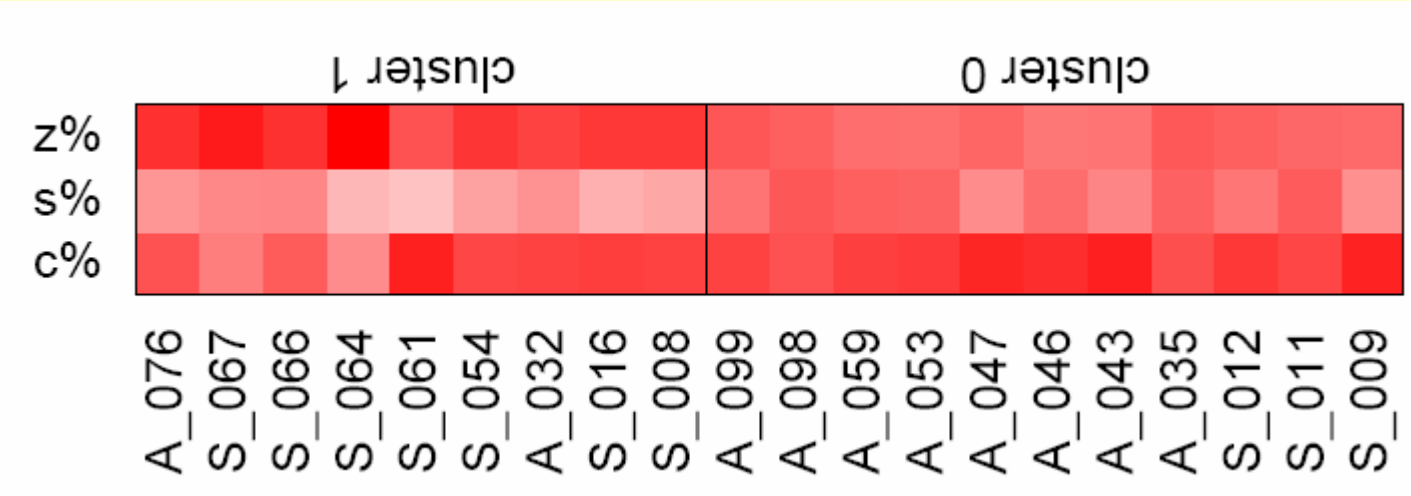
$$
\begin{aligned}
C(l|L) &= \sum_{\pi \in L} p(\pi) C(l|\pi) \\
&= \sum_{\pi \in L} \left(p(\pi) \sum_{a \in \pi} \delta(a,l) \right) \\
&= \sum_{a \in L} \left(\delta(a,l) \sum_{\pi \in L : a \in \pi} p(\pi) \right) \\
&= \sum_{a \in \mathcal{I}(l) : L[a] = L} p(a) \\
&= \sum_{a \in \mathcal{I}(l) : L[a] = L} f(k[a]) p(a|k[a])
\end{aligned}
$$



c(s) = **0.2** * *0.3* = 0.06
c(sh) = **0.2** * *0.1* = 0.02

“Clustering”

- Cluster into two groups using repeated bisections algorithm with cosine distance measure:



Research Proposal

Adaptation Techniques for Accented Speech

Yanli Zheng

University of Illinois

Dialectal Chinese Speech Recognition

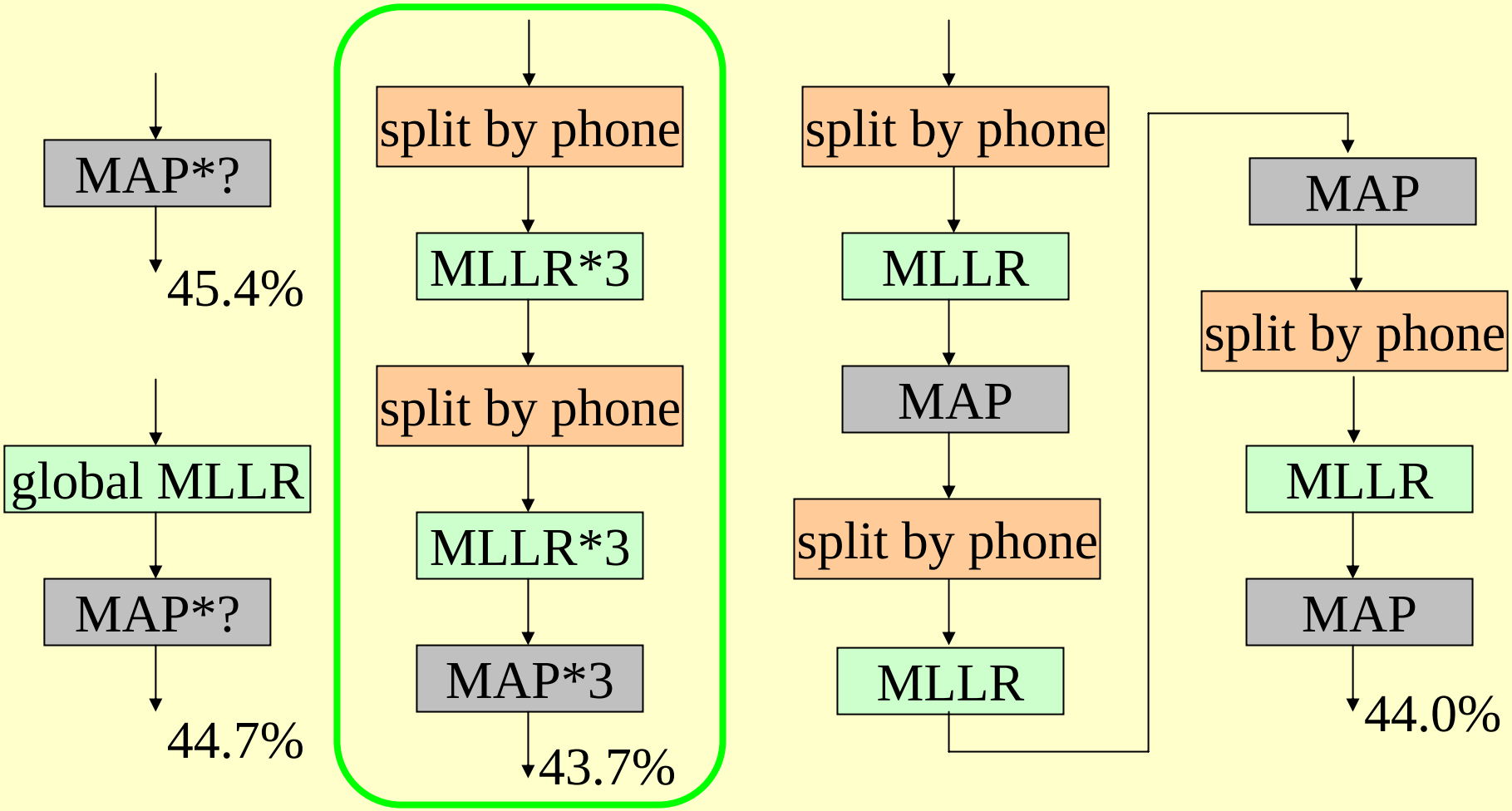
Adaptation Techniques for Accented Speech Recognition

Research Proposal, Yanli Zheng

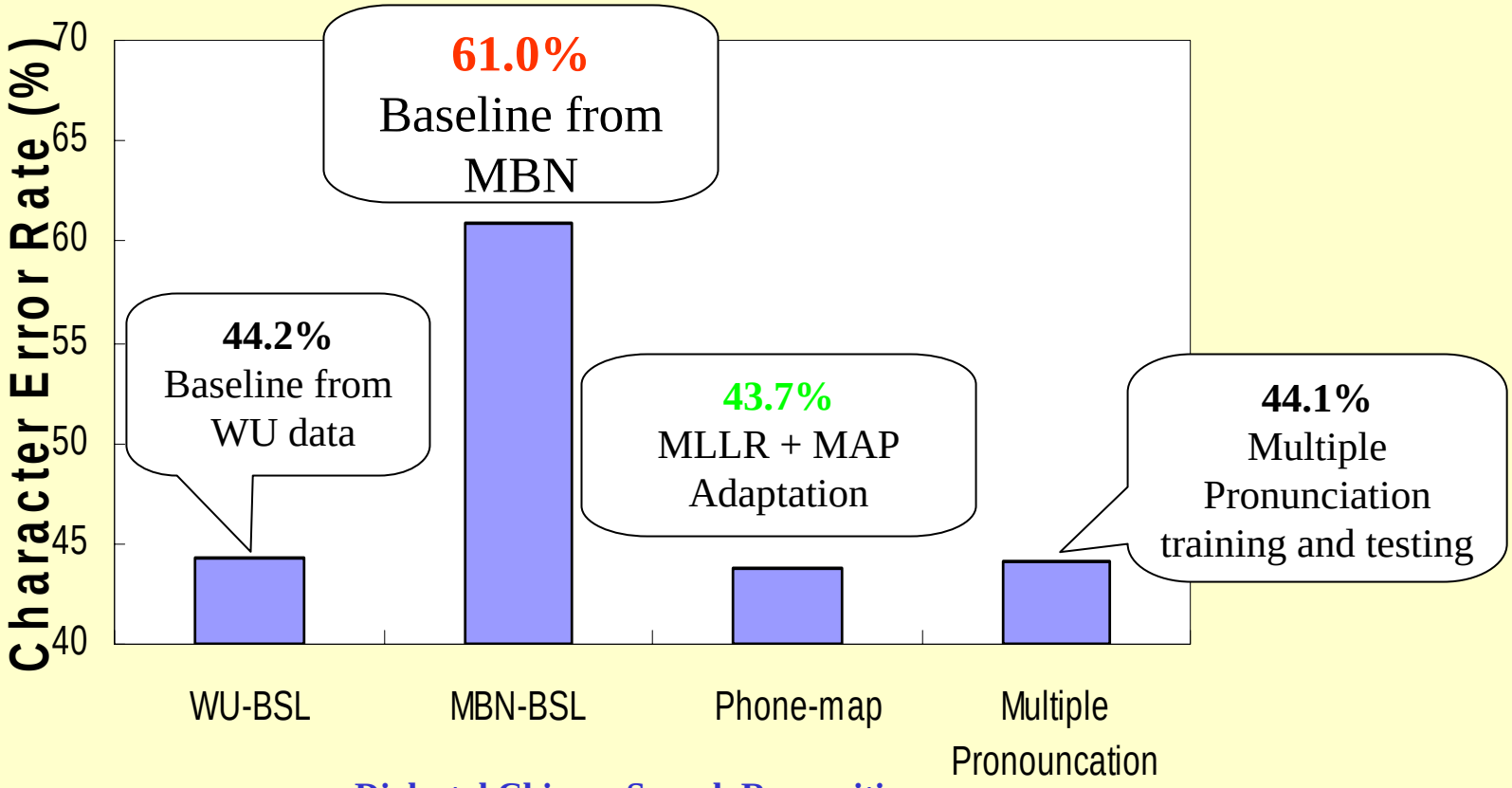
- Adapting to Wu accented speech
  - Supervised MLLR+MAP on 80 speakers in accented training set
- Analysis of Results
  - Adaptation effects different phones differently
- Proposal for Accent based Acoustic Modeling
- Proposal for Detection of Accent Degree

Different Adaptation Methods

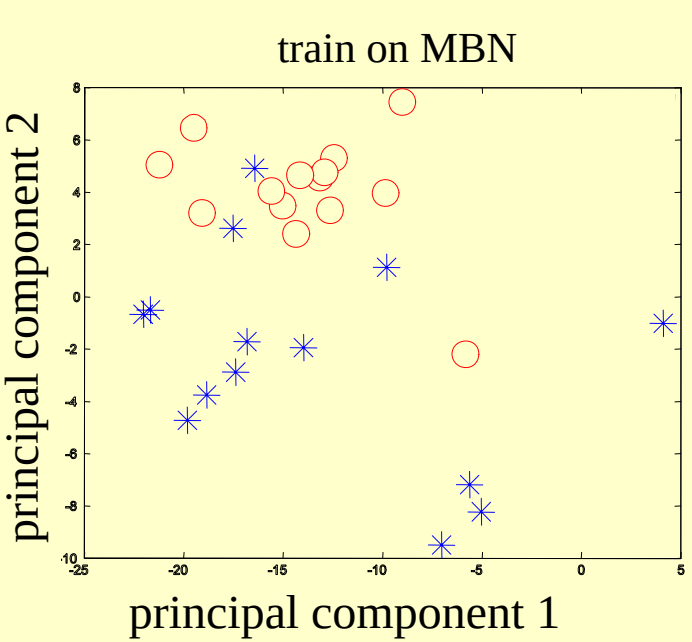
(MBN Baseline CER 61%)

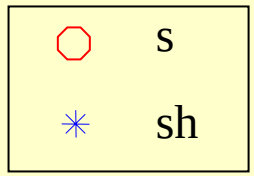


Accent ASR Experiments on Various Acoustic Models



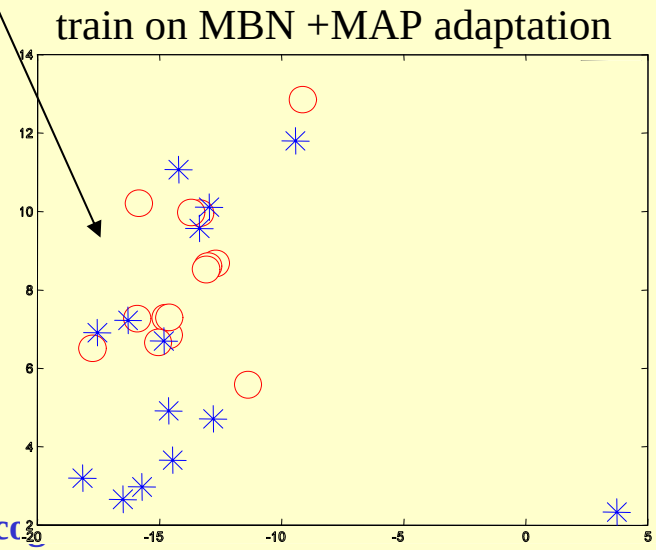
Comparison of Gaussian Probability Distributions before and after Adaptation



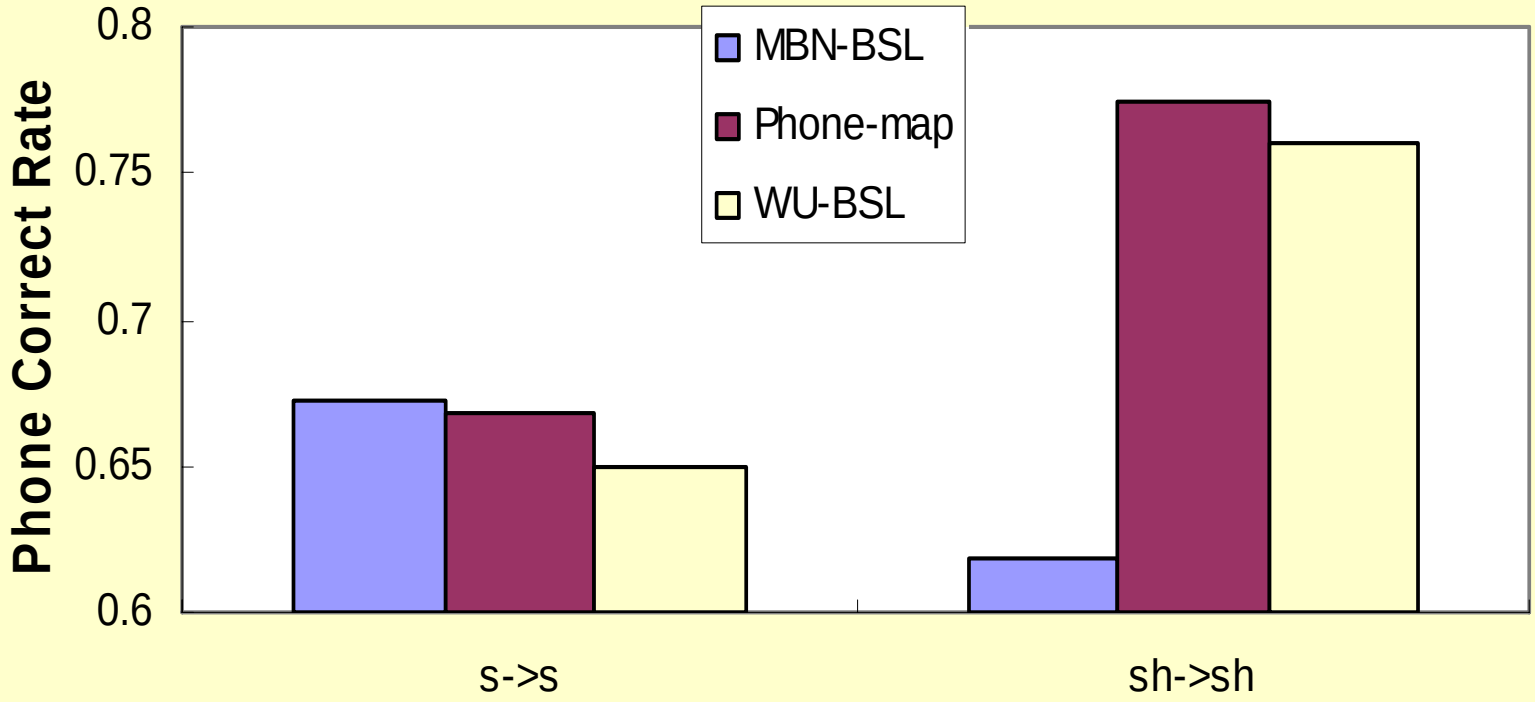


Acoustic adaptation models phone confusion

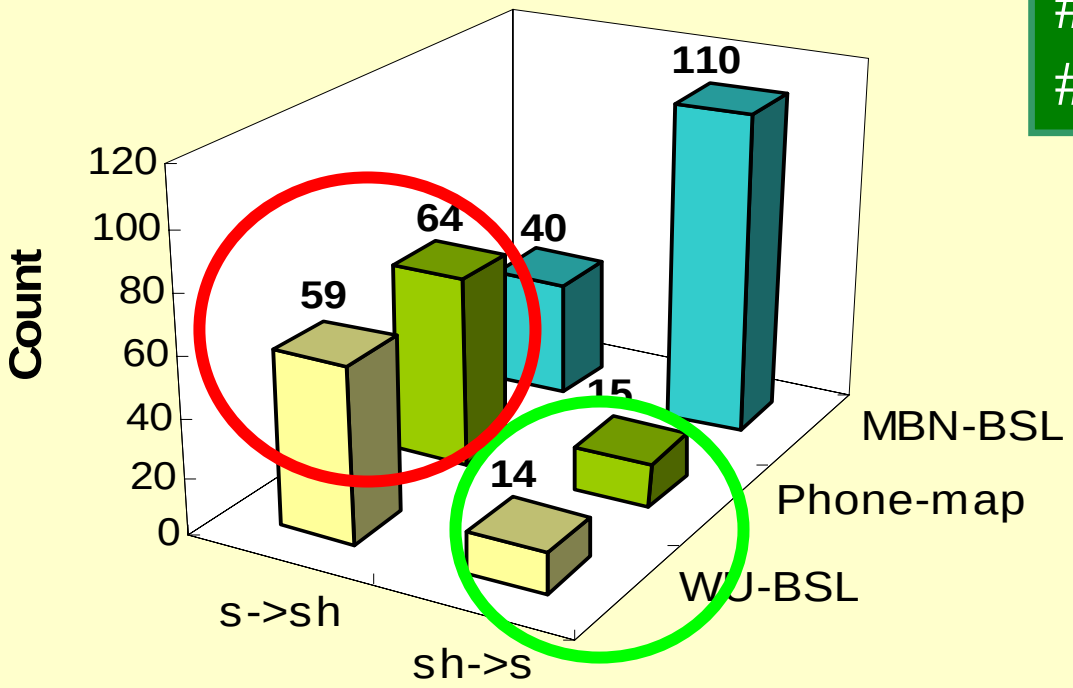
adapt



Adaptation improves "sh" recognition but not "s" recognition



"s" & "sh" Substitution Rate for different Acoustic Models



# s	208
# sh	2454

A priori statistics favor "sh"

Adaptation Techniques for Accented Speech Recognition

Research Proposal, Yanli Zheng

- Adapting to Wu accented speech
  - Supervised MLLR+MAP on 80 speakers in accented training set
- Analysis of Results
  - Adaptation effects different phones differently
- Proposal for Accent based Acoustic Modeling
- Proposal for Detection of Accent Degree

Accent Based MAP Estimation

Factor MAP into two components:
predict accent given data; predict model given accent

a: Accented Variable

$$\theta_{MAP} = \{\theta^{(1)}_{MAP}, \cdots, \theta^{(N)}_{MAP}\}$$
$$= \underset{\theta}{argmax}\ g(\theta^{(1)}, \cdots, \theta^{(N)} | x^{(1)}, \cdots, x^{(N)})$$
$$= \underset{\theta}{argmax}\ f(x^{(N)}_{(1)} | \theta^{(N)}_{(1)}) g(\theta^{(N)}_{(1)})$$
$$= \prod_{k=1,\cdots,N} \underset{\theta^{(k)}}{argmax}\ f(x^{(k)} | \theta^{(k)}) g(\theta^{(k)})$$
$$= \prod_{k=1,\cdots,N} \underset{\theta^{(k)}}{argmax} \sum_{a} g(\theta^{(k)} | a, x^{(k)}) p(a | x^{(k)})$$
$$= \prod_{k=1,\cdots,N} \underset{\theta^{(k)}}{argmax} \sum_{a} g(\theta^{(k)} | a) p(a | x^{(k)})$$

Accent Based Model Selection

Speaker specific MAP:

$$\theta^{k}_{MAP} = \underset{\theta^{(k)}}{argmax}\ g(\theta^{(k)} | x^{(k)})$$
$$= \underset{\theta^{(k)}}{argmax} \sum_{a} g(\theta^{(k)} | a) p(a | x^{(k)})$$

Accent Detection

Special Case of $g(\theta^{(k)} \mid x^{(k)})$

Approach 1: **Binary Accent Detection**

$$\theta^{(k)} = \begin{cases} \theta_{accent} & a = 1, \quad \text{Accent speaker} \\ \theta_{standard} & a = 0, \quad \text{standard speaker} \end{cases}$$

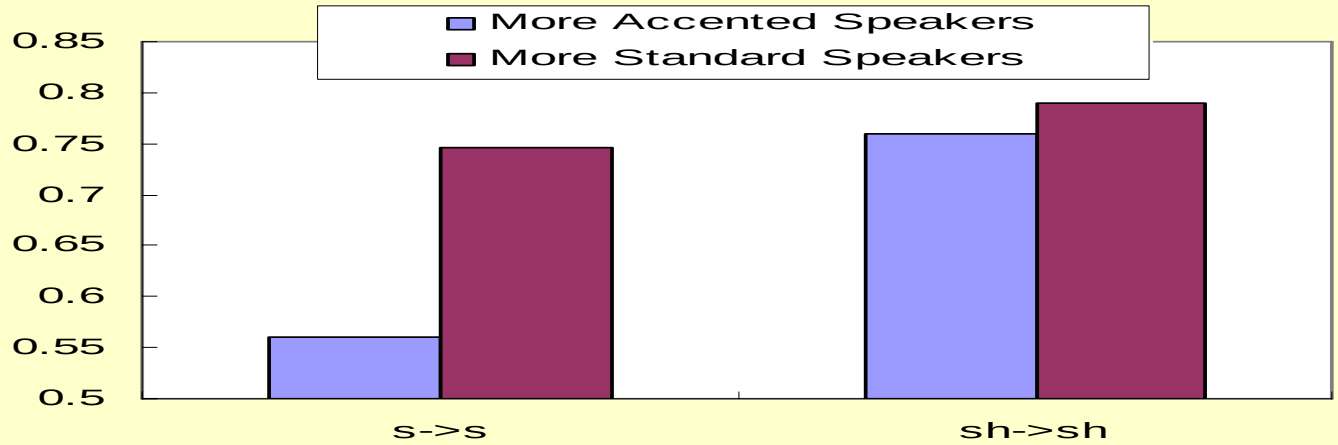
$$g(\theta^{(k)}|x^{(k)}) = \delta(a^{(k)} - 1)p(\theta_{accent}|x^{(k)}) + \delta(a^{(k)})p(\theta_{standard}|x^{(k)})$$

Approach 2: **Detection of Degree of Accent**. Model accent degree as linear combination of accented and standard models:

$$g(\theta^{(k)}|x^{(k)}) = a^{(k)}p(\theta_{accent}|x^{(k)}) + (1 - a^{(k)})p(\theta_{standard}|x^{(k)})$$

Approach 3: Different phone (classes) need to be treated differently

Why?



How?

Adaptation based on Accent Detection

$$
\begin{aligned}
g(\theta_{MAP}|x) &= \underset{\theta}{argmax}\ g(\theta|x) \\
&= \underset{\theta_1^M}{argmax}\ g(\theta_1, \ldots, \theta_M|x) \\
&= \prod_{m=1}^{M} \underset{\theta_m}{argmax}\ g(\theta_m|x)
\end{aligned}
$$

Accent Detection

$$\theta^k_{MAP} = \underset{\theta^{(k)}}{argmax}\ g(\theta^{(k)}|x^{(k)})$$

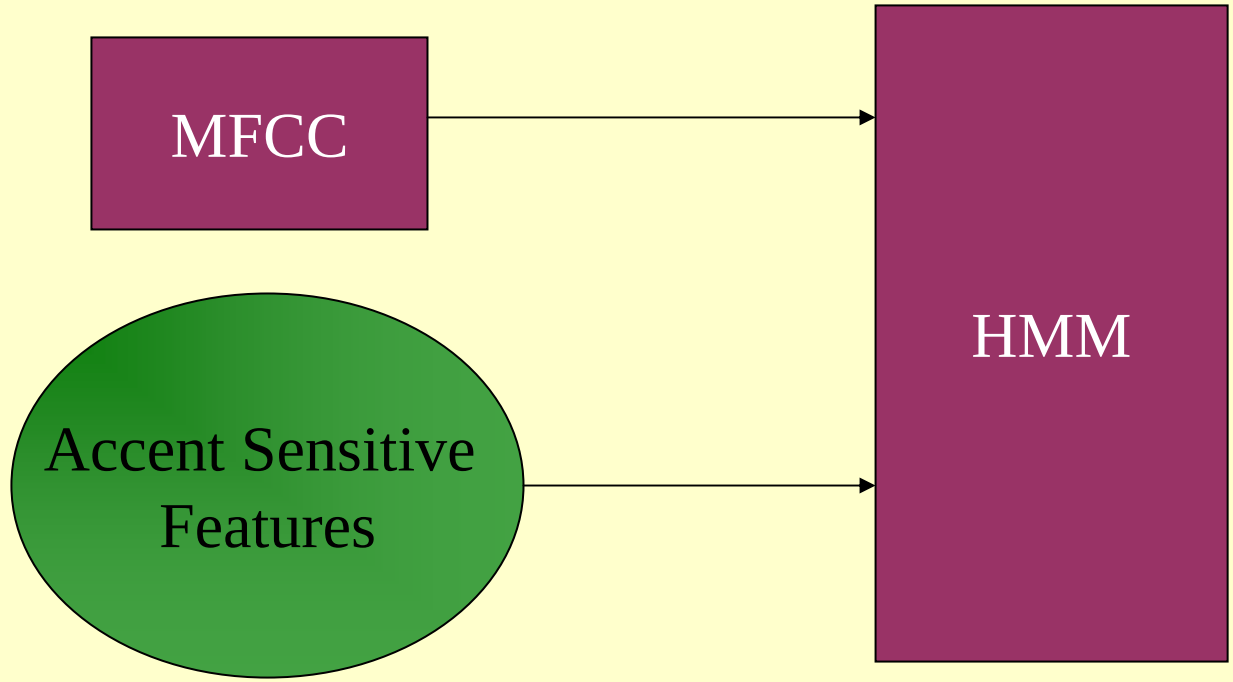
$$= \underset{\theta^{(k)}}{argmax} \sum_a g(\theta^{(k)}|a)p(a|x^{(k)})$$

Accent Based Model Selection

Accent Detection

Detection of Accent Degree

Approach 1: Combine with Accent Sensitive Features

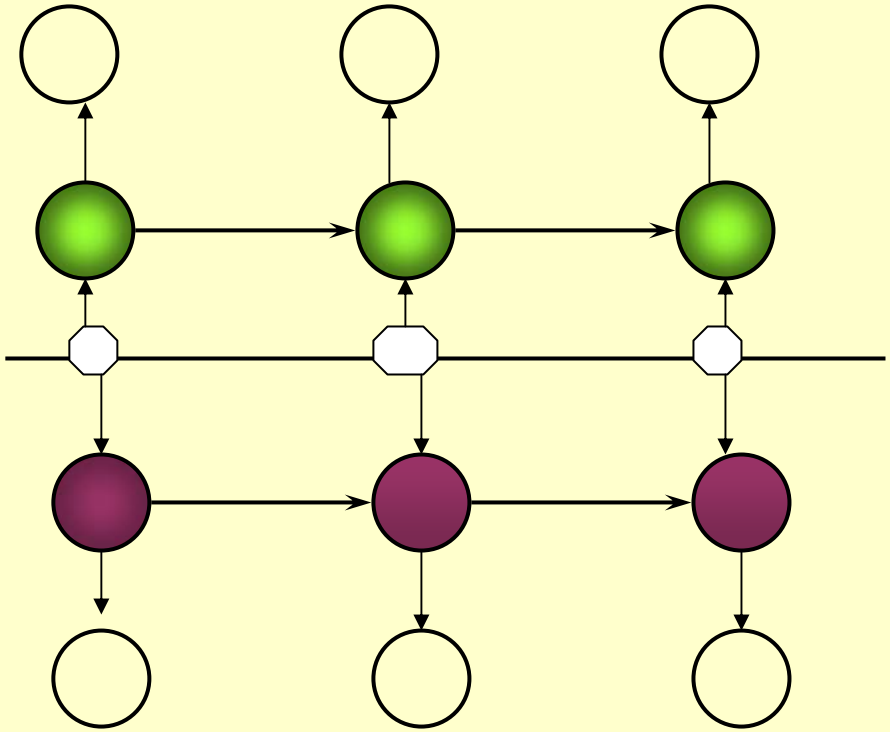


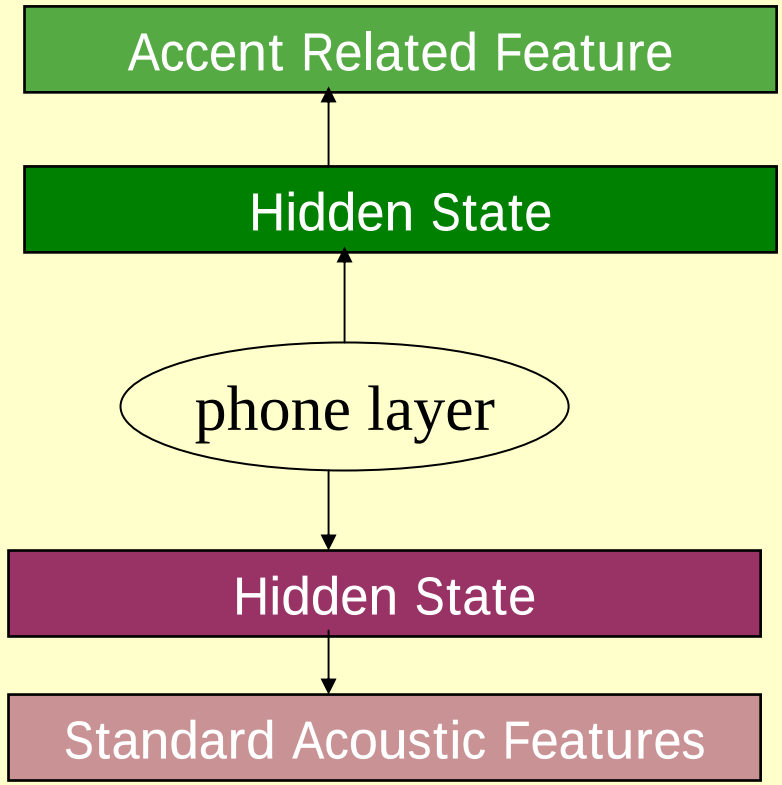
s/sh z/zh c/ch

Formant

Pitch

Approach 2: Multi-Stream Acoustic Model Method





Conclusion

Purpose

- Accent-based Modeling for accented speech

Strategies

$$\theta^k_{MAP} = \underset{\theta^{(k)}}{argmax}\ g(\theta^{(k)}|x^{(k)})$$

$$= \underset{\theta^{(k)}}{argmax} \sum_a g(\theta^{(k)}|a)\ p(a|x^{(k)})$$

Accent Based Model Selection

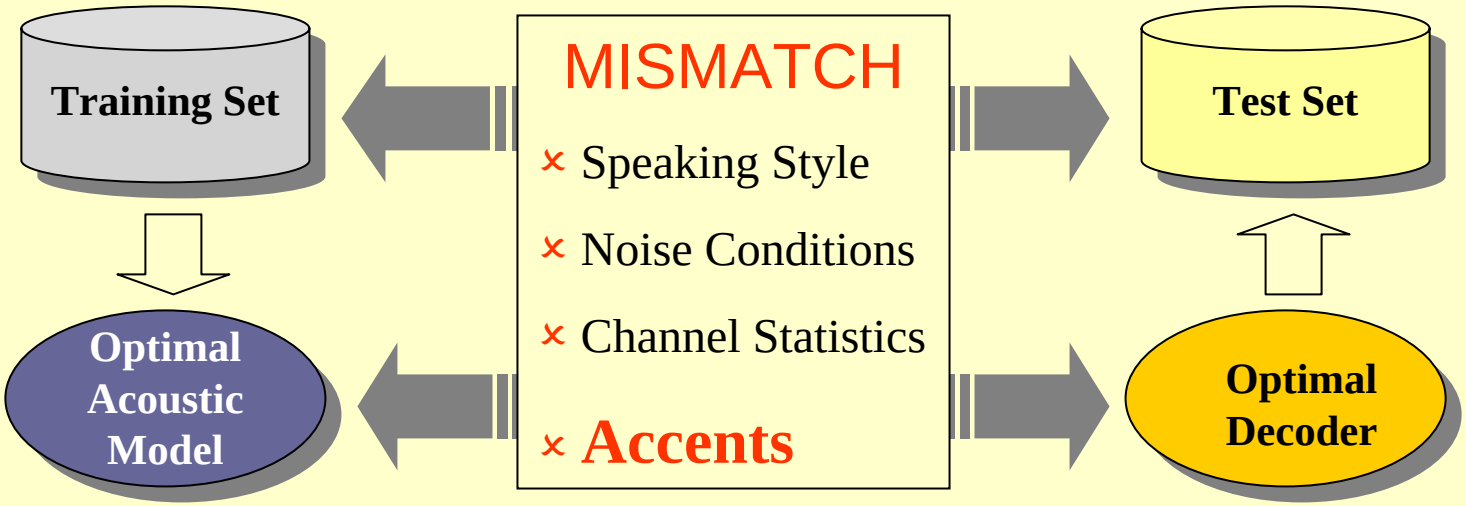
Accent Detection

Dynamically adjusted acoustic model based on accent detection

Presenter: Liang Gu

Dialectal Chinese Speech Recognition

Problem Statement



Example of Impact of Accent Mismatch

Acoustic Model trained on Mandarin Broadcast News (MBN)

MBN: **19.1** %*

Wu-Accented Read Speech: **57.4** %*

Wu-Accented Spontaneous Speech: **61.0** %*

*CER: Character Error Rate

Acoustic Modeling on Accented Speech

Model Training on Limited Accented Data

Conventional Approaches

- Model Trained on Accented Data

  (Maximum Likelihood Estimation, etc.)
- Model Adaptation on Accented Data
  - MAP
  - MLLR

New Approaches

- **Automatic Accent-based Speaker Clustering**
- **Model Trained on Speaker-Clustered Accented Data**
- **Model Adaptation on Speaker-Clustered Accented Data**

Decoding

Conventional Approaches

- Decoding using Accented-Speech Acoustic Models
- Decoding using Phone-MAP-Adapted Acoustic Models
- Max-Posterior-Probability FSM Decoding

New Approaches

- Decoding using Accented-Speech and Phone-MAP-Adapted Acoustic Models
- **Best-path Selection** between Accented-Speech and Phone-MAP-Adapted Acoustic Models based on **Accent Detection**

Decoding based on Accent Detection

GMM-based Automatic Accent Detection

More Standard-PTH Speaker

More Accented Speaker

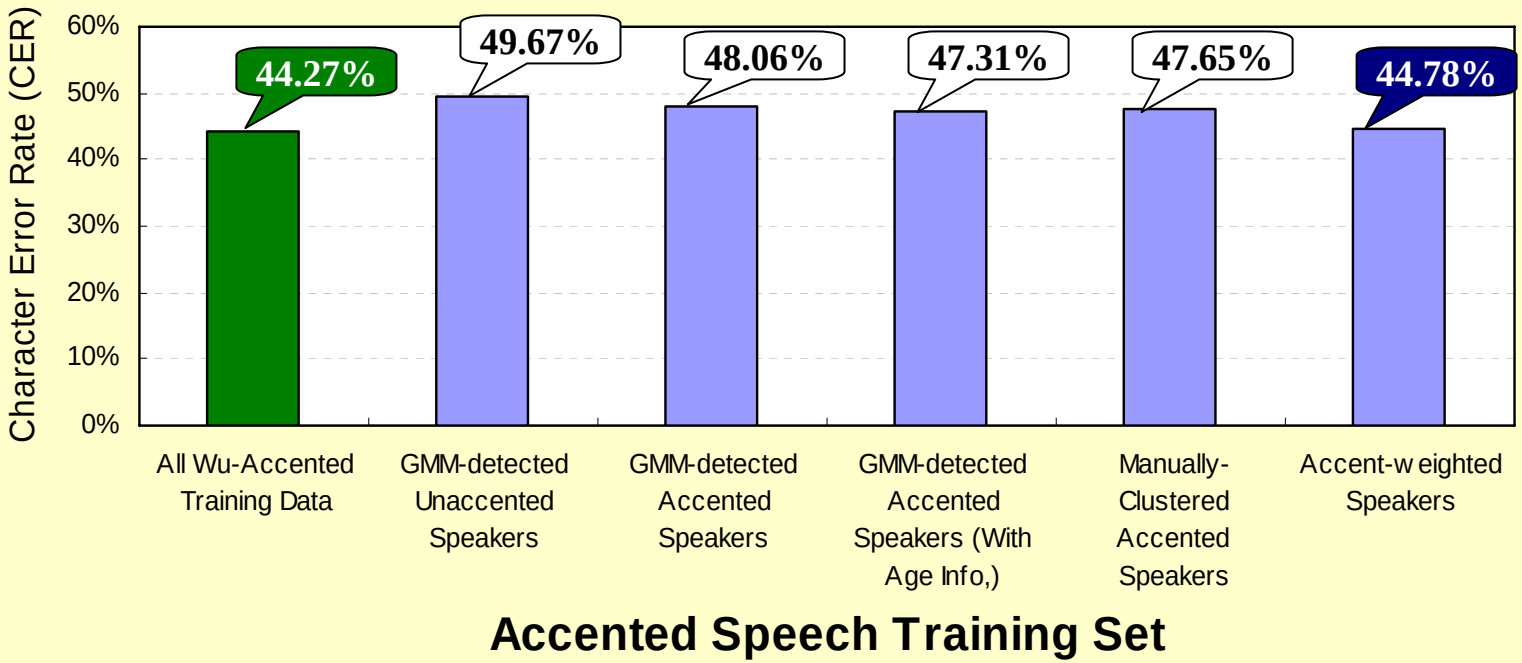
Decoding using Phone-MAP-Adapted Acoustic Models

Decoding using Accented-Speech Acoustic Models

Decoding Output

Acoustic Modeling on Accented Speech

Experiment on Accent-Clustered Training

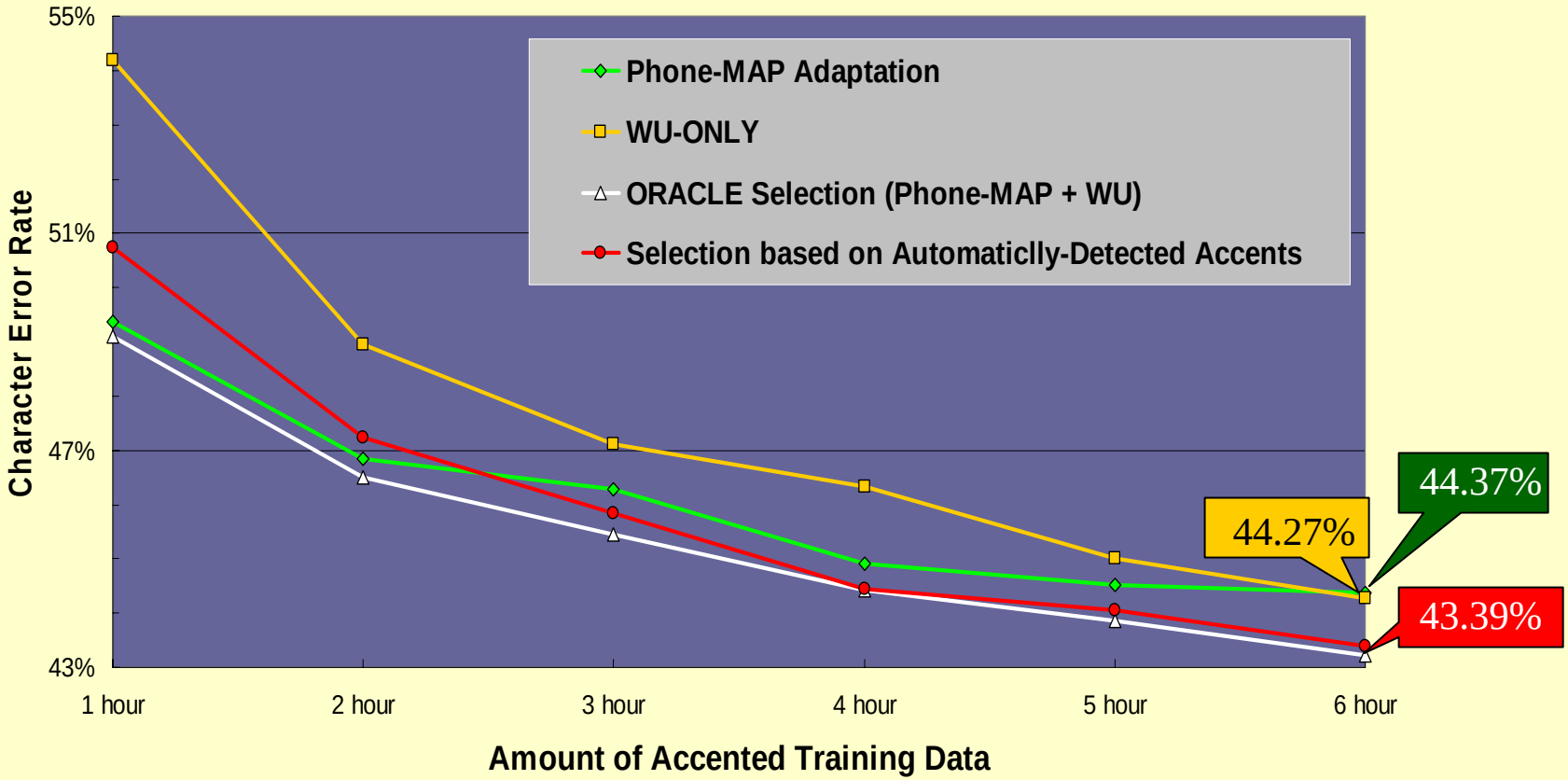


Insufficient Accented Speech Training Data

Dialectal Chinese Speech Recognition

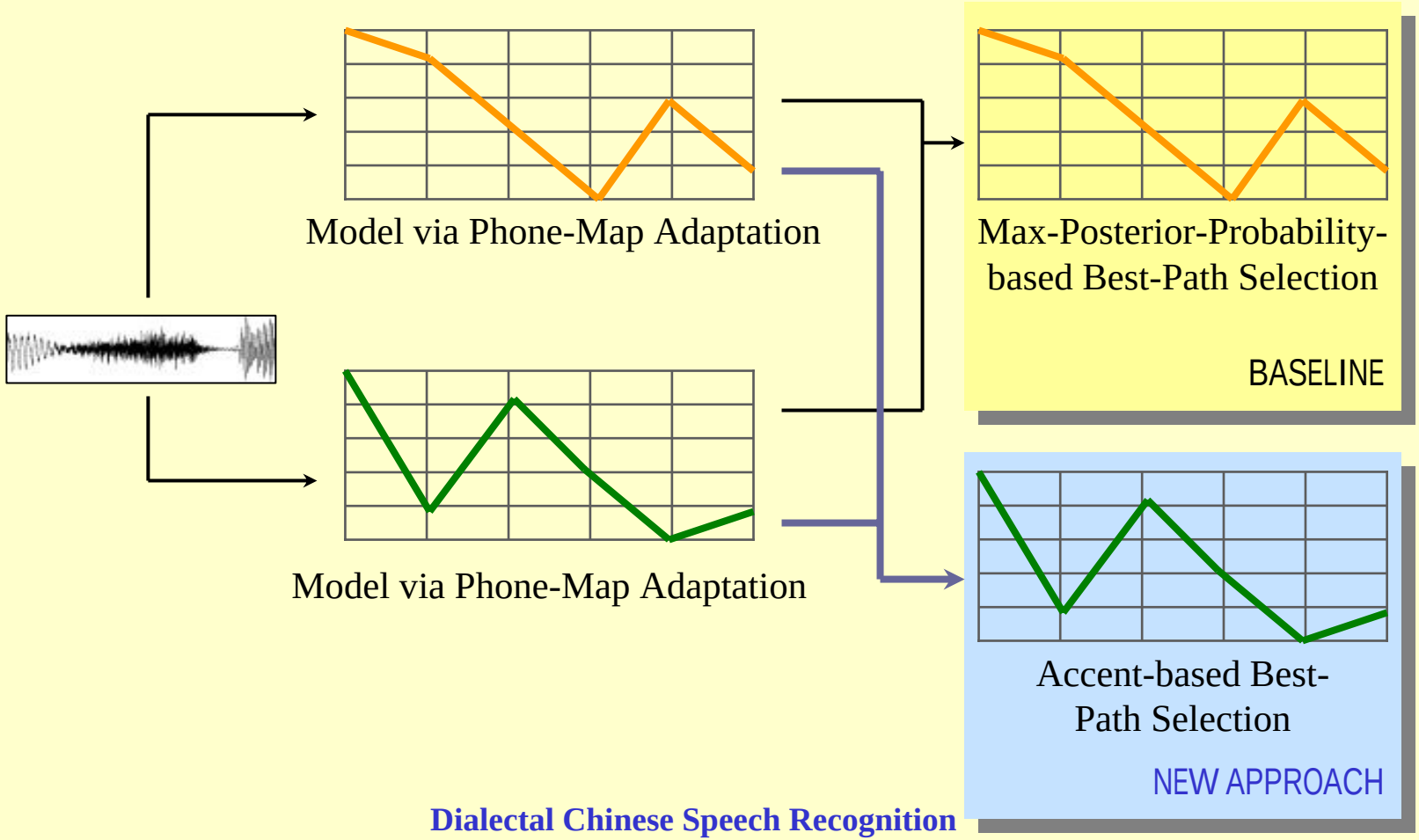
Acoustic Modeling on Accented Speech

Experiment on Accent-Clustered Decoding (I)



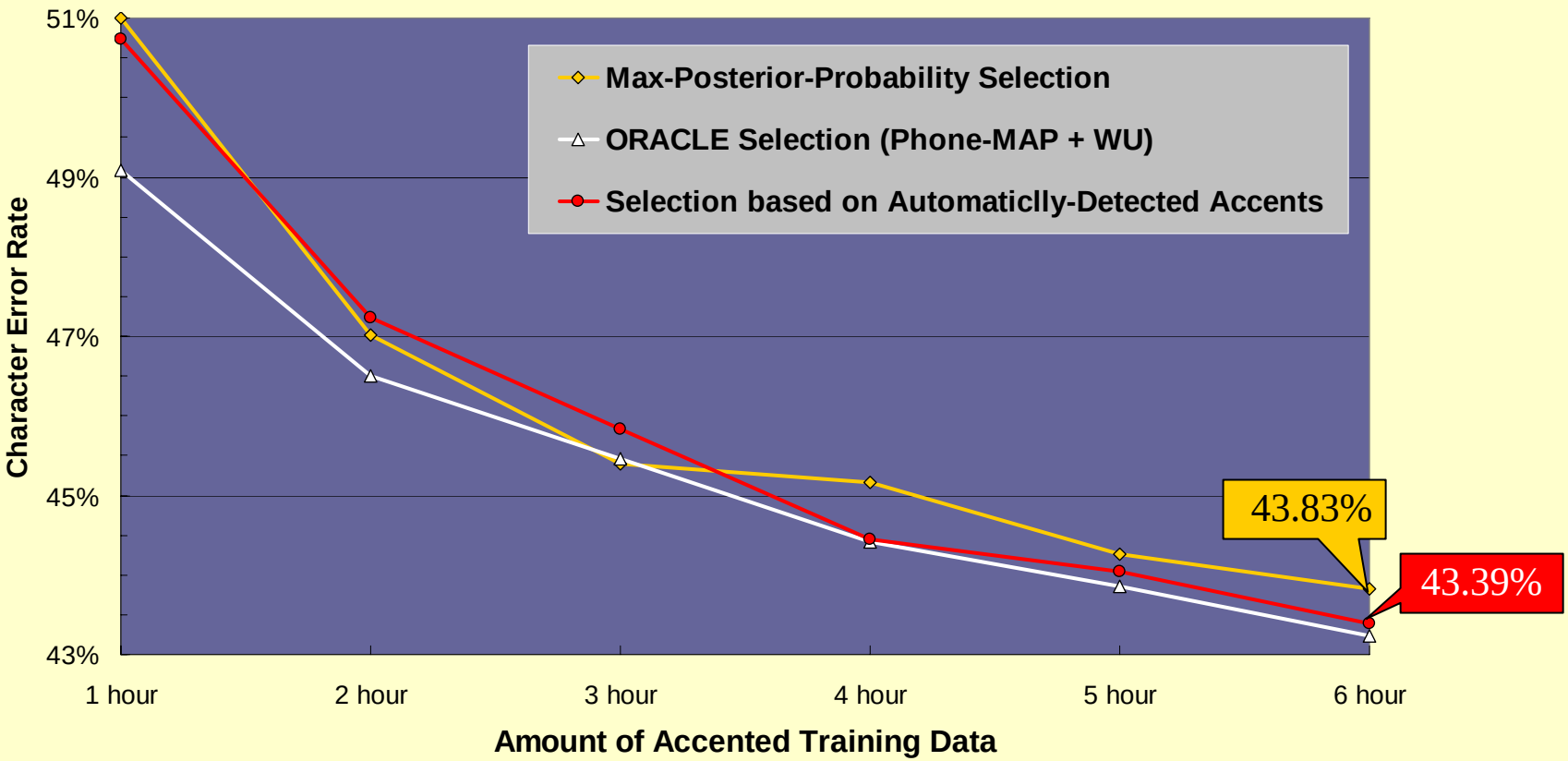
Acoustic Modeling on Accented Speech

Experiment on Accent-Clustered Decoding (II)



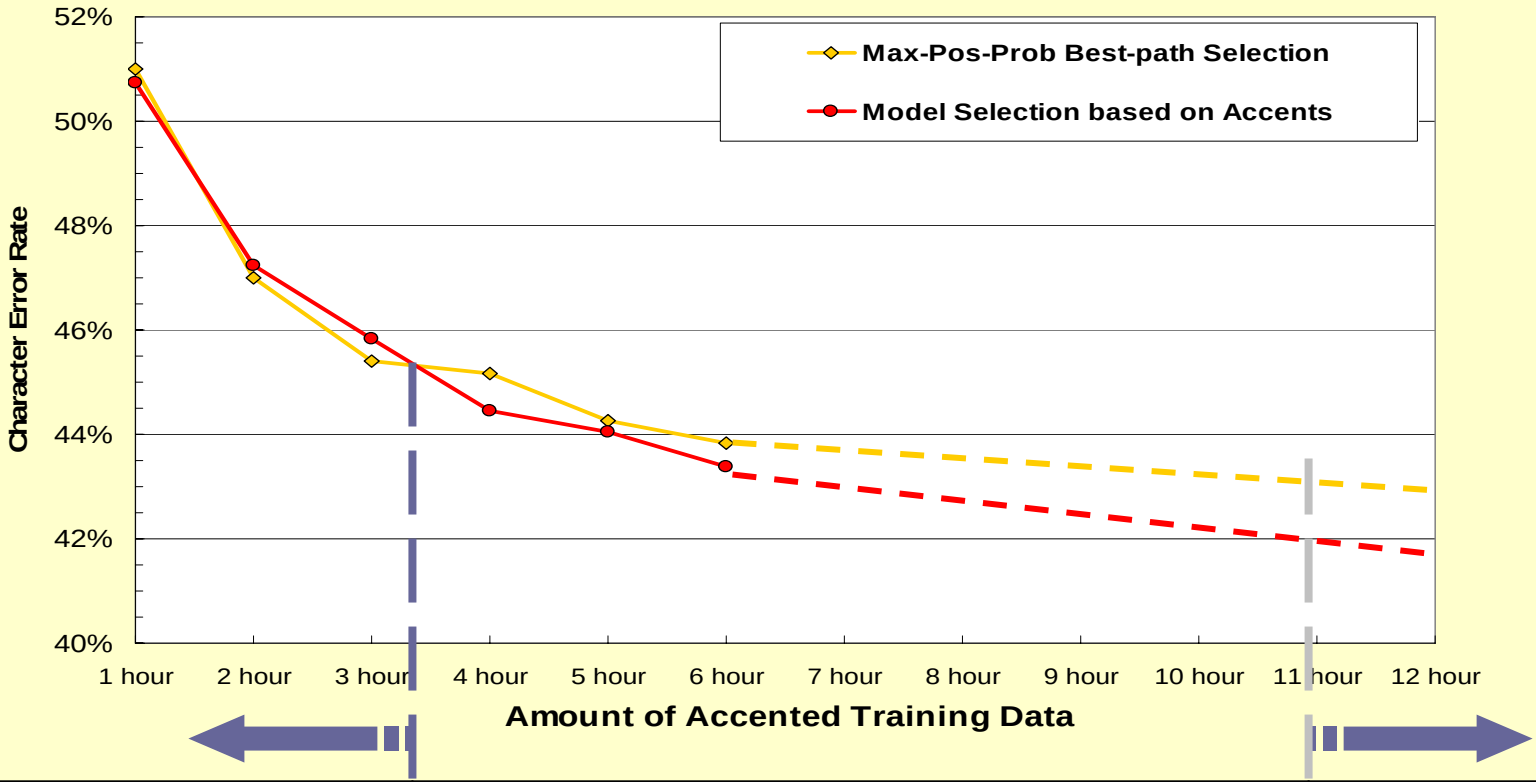
Acoustic Modeling on Accented Speech

Experiment on Accent-Clustered Decoding (II)



Discussion on Accented Speech ASR Experimental Results

Accented Data vs. Optimal Modeling Method



Acoustic Modeling on Accented Speech

Summary

- Wu-Accented Speech Recognition
  - Baseline acoustic models trained on 120-hour Standard PTH
  - 6.3-hour wu-accented acoustic training data
  - 20 wu-accented test speakers with various accent degree
- Conventional Approaches
  - MAP & MLLR adaptation
  - Model training on limited accented training data
  - Hybrid-Decoding by maximizing posterior probability
- New Approaches
  - Automatic "**More Accent**" / "**More Standard**" accent detection
  - Hybrid-Decoding by selecting **Accent-Matched acoustic models**
  - **Reduce CER (Character Error Rate) by 0.4% ~ 0.9% absolute**
  - **More improvement with larger accented training set**

Research Proposal

Minimal Perplexity Word Segmentation

Tweaking the segmentation of a small corpus to minimize language model perplexity

Philip Bramsen, MIT

Dialectal Chinese Speech Recognition

Motivation

- Chinese ASR typically based on (fixed) dictionary of words.
  - Character pronunciations often depend on word affiliation
- Therefore Chinese ASR LM's typically built out of words
- Problem: Chinese text lacks word boundaries

Word Segmentation

- Various approaches to Chinese word segmentation proposed
- For ASR, one segments the training corpus then builds a LM on segmented corpus
- For a fixed dictionary, *maximum matching* works about as well as anything

Maximum Matching

- Suppose English were written with no spaces:

  theirgardenisovergrown

  Dictionary: *the*, *their*, *gar*, *garden*, *den* …
- Left-to-right maximum matching would give:

  their garden is overgrown

Effect of Dictionary on LM Perplexity

- Compare the perplexity of a bigram model based on:
  - Dictionary of single characters only
  - Our baseline 50k word dictionary
- Character-based LM: 88.1
- Dictionary-based LM: 69.2
- Question: is there a minimal perplexity dictionary?

Sidenote: Why Bigram Model?

- Conversational speech training corpora small, so wide n-gram language models infeasible
- In our task:
  perp(trigram) > perp(bigram)

A Heuristic Iterative Approach to Dictionary Optimization

- Add to the dictionary all word bigrams $w_1 w_2$ where:

$$C(w_1 w_2) > Threshold_{C(w_1 w_2)}$$

$$\frac{C(w_1 w_2)}{C(w_1)} > Threshold_{\%w_1}$$

$$\frac{C(w_1 w_2)}{C(w_2)} > Threshold_{\%w_2}$$

A Heuristic Iterative Approach to Dictionary Optimization

- Related to mutual information but we are not using mutual information as a threshold

- Similar in spirit to segmentation methods that attempt to minimize entropy or description length (e.g de Marcken, 1996)

New Word Examples

- Some words found with thresholds 5_50_50

"say clearly"

"not easy"

"not know"

"masked palm civet" (*Paguna larvata*)

"cell phone"

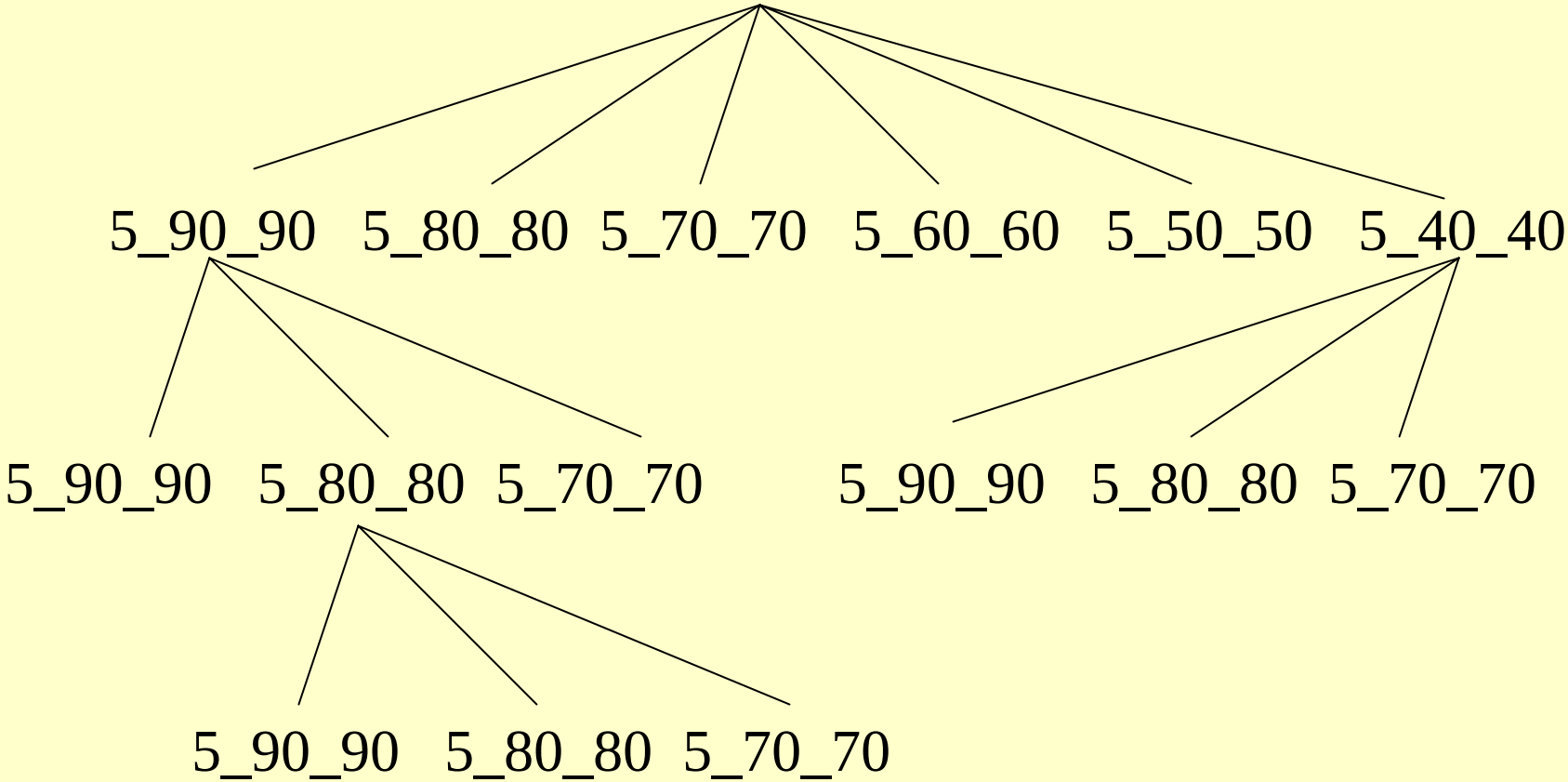
"electronic products"

"in school"

Algorithm

- Create new dictionary as described
- Use new dictionary to resegment the training/test corpora
- Rebuild LM on training corpus
- Compute perplexity on test corpus
- Recompute new words on newly segmented corpus
- Repeat process

Parameter Space



Perplexity Results

Baseline Dictionary	69.19
First Iteration Best 5_3_3	68.30
Second Iteration Best 5_4_4+5_4_4	68.09
Character-based variable-length (9)-gram model	77.29

Perplexity Results: Character Lexicon

Baseline Character Dictionary	88.05
First Iteration Best 5_2_2	75.03
Second Iteration Best 5_3_3+5_3_3	73.67
Character-based variable-length (9)-gram model	77.29

Example of Iterative Segmentation (Starting with Characters)

- No grouping:

  “What kind of pants do you like, jeans or slacks?”

- First iteration:

- Second iteration:

Reduction of CER

- Rescoring lattice with lowest perplexity LM reduces CER from best system from 43.7% to 43.6%
- Not entirely fair since lowest perplexity LM is selected on basis of same test corpus used for scoring decoding
- However we believe further reductions in perplexity are possible

Future Work

- Select path in tree using held-out training data (cross-validation)
- More systematic search of parameter space
- Explore sensitivity of other segmentation approaches than maximum matching
- Extend to other languages: how can one improve the dictionary for English ASR?

Advisors

- Project advisor: Richard Sproat
- Local advisor (MIT): James Glass

Conclusions

- New approach for accented speech ASR:
  1. Detect accent
  2. Select acoustic model based on accent
  - **1%** improvement in CER over best current system on accented speech
- Developed accent-specific transforms using supervised MLLR plus MAP on accented training corpus

Conclusions

- Novel accentedness detection method based on phone count ratios
  - Ratios computed from decoder lattices
  - Used for unsupervised speaker clustering and adaptation

Conclusions

- Consistent with previous results, our results suggest pronunciation modeling/lexicon adaptation may be worth about 1.5%

- New “IF Mapping” approach based on minimal training data, shows promise

Future Work

- Extend binary accent-based modeling to modeling based on continuous "accentedness" values, both for:
  - acoustic modeling (Yanli Zheng)
  - pronunciation modeling (David Kirsch)
  - model degree of accent using iterative application of transform matrix: $\mu' = W^n(\mu)$
- Promising approach to minimal perplexity word segmentation (Philip Bramsen)

Future Work

- Extend IF mapping to syllable mapping
- Language-model adaptation for accented speech
- Multi-stream HMM's with hidden accent variables and accent-related acoustic observations