AGILE Speech to Text (STT)

Contributors:

BBN: Long Nguyen, Tim Ng, Kham Nguyen, Rabih Zbib, John Makhoul

CU: Andrew Liu, Frank Diehl, Marcus Tomalin, Mark Gales, Phil Woodland

LIMSI: Lori Lamel, Abdel Messaoudi, Jean-Luc Gauvain, Petr Fousek, Jun Luo

GALE PI Meeting

Tampa, Florida

May 5-7, 2009

Overview

- AGILE STT progress in P3 (Nguyen)
- Morphological decomposition for Arabic STT (Nguyen)
- Sub-word language modeling for Chinese STT (Lamel)
- MLP/PLP acoustic features (Gauvain)
- Language model adaptation (Woodland)
- AGILE STT future work (Woodland)

AGILE STT Progress for P3 and P3.5 Evaluations

Long Nguyen

BBN Technologies

AGILE P3 Arabic STT System

- ROVER combination of several outputs from BBN, CU and LIMSI
- Acoustic models trained on ~1400 hours of Arabic audio data
- Language models trained on 1.7B words of Arabic text
- 16% relative improvement in WER in P3 system compared to P2 system

System	dev07	dev08	P3 test
P2	10.3	----	
P3	8.6	10.0	8.1

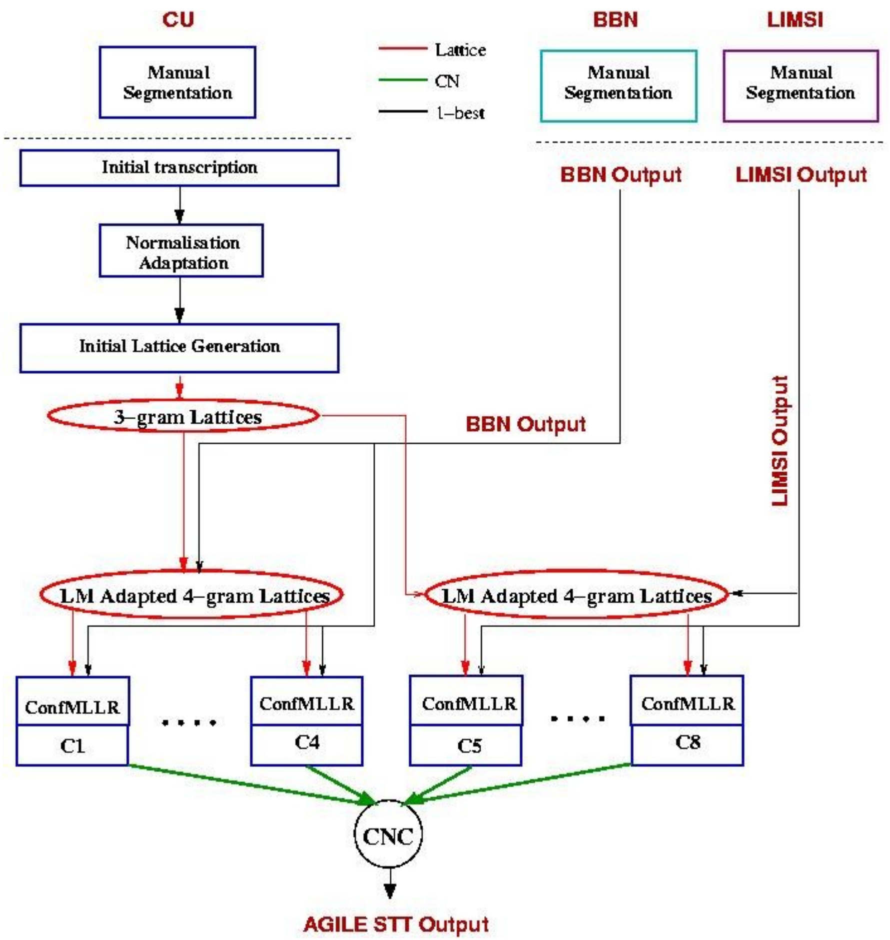
Key Contributions to Improvement

- **Extra training data**
- **Multi-Layer Perceptron (MLP) acoustic features***
- **Improved phonetic pronunciations**
  - **Augmented Buckwalter analyzer's list of MSA affixes with some dialect affixes to obtain pronunciations for dialect words**
  - **Developed procedure to automatically generate pronunciations for words that cannot be analyzed by Buckwalter analyzer**
- **Class-based and continuous-space language models**
- **Morphological decomposition***

*** Full presentations later**

AGILE P3.5 Mandarin STT System

- **Cross-adaptation framework**
  - CU adapts to BBN and to LIMSI output
  - Acoustic and LM adaptation
- **8-way final combination**
- **Acoustic models trained on 1700 hours**
- **Language models trained on ~4B characters**



Improvement for P3.5 Mandarin STT

GALE The AGILE Team

- 0.9% CER absolute improvement from P2.5 system to P3.5 system

	P2.5 Test	dev08	P3.5 Test
P2.5 System	8.0	8.4	11.2
P3.5 System	7.1	7.3	10.3

- Key contributions to improvement
  - Extra training data
  - MLP/PLP features*
  - Linguistically-driven word compounding
  - Continuous-space language model
  - Language model adaptation*
- CER of P3.5 test is 47% higher than that of P2.5 test

… and Most of the Errors are Due to:

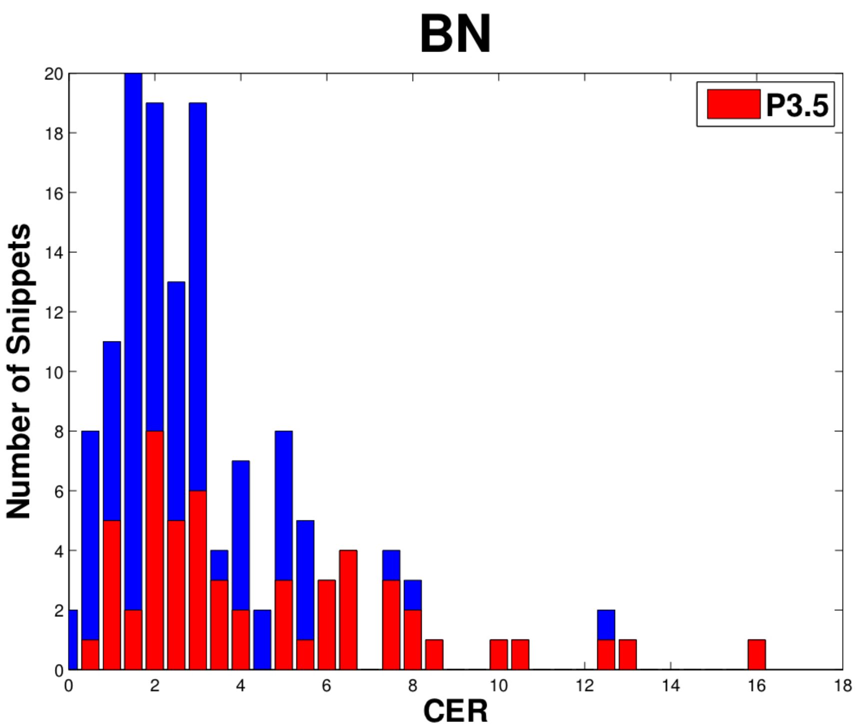
- **More overlapped speech in P3.5 compared to P2.5**

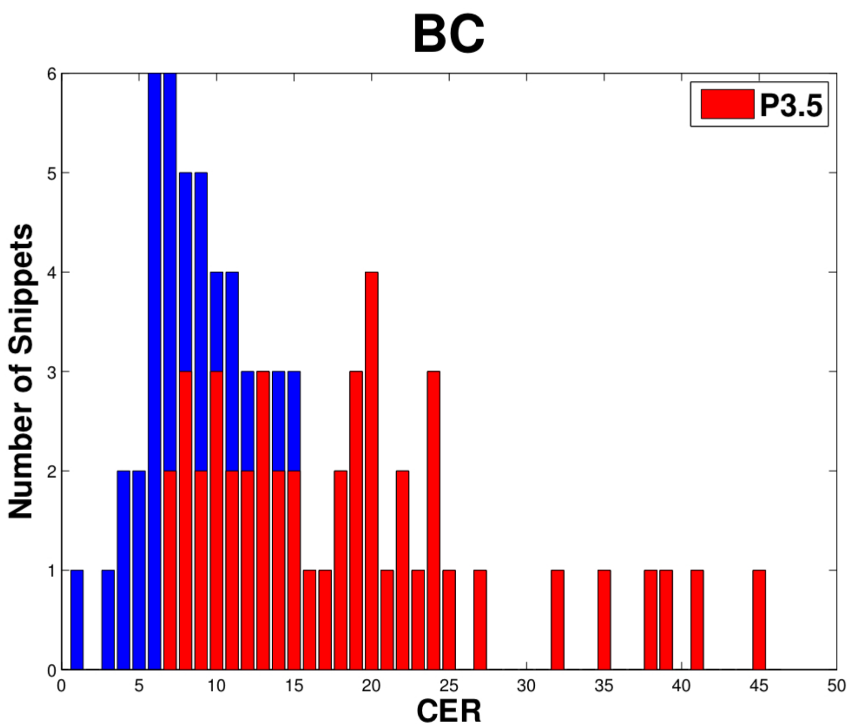
Eval Sets	Overlapped / Total Duration (sec)	Percentage
P2.5	198 / 8760	2.3%
P3.5	305 / 10168	3.0%

- **Accented speech (Taiwanese, Korean and others)**
- **Poor acoustic channel (phone-in)**
- **Background music or laughter**
- **Names (personal, program and foreign)**
- **English words (GDP, Cash, FDA, EQ …)**

Mandarin P3.5 Test vs. P3.5 Data Pool

- **Overall CER for P3.5 Pool is 7.7% (similar to that of P2.5 Test) while CER for P3.5 Test is 11.6%**





Summary

- Significant improvements for the team's combined results as well as individual site results
- More work to be done to improve STT further, especially for Mandarin (to be presented in Future Work slides)

Morphological Decomposition for Arabic STT

Long Nguyen

BBN Technologies

Outline

- **BBN work on morphological decomposition using Sakhr's morphological analyzer**
  - Comparison of out-of-vocabulary (OOV) rates and word error rates (WER) of four word-based and morpheme-based systems
  - System combination
- **CU work on morphological decomposition using MADA**
- **LIMSI work on morphological decomposition derived from Buckwalter morphological analyzer**

Word-Based Arabic STT Systems

- **Implemented two traditional word-based systems**
  - **Phonetic system (P)**
    - **Each word was modeled by one or more sequences of phonemes of its phonetic pronunciations**
    - **Vocabulary consisted of 390K words derived from the 490K most frequent words in acoustic and language training data (i.e. only words having phonetic pronunciations)**
  - **Graphemic system (G)**
    - **Each word is modeled by a sequence of letters of its spelling**
    - **Vocabulary included all of the 490K frequent words**
- **Arabic STT word-based systems require very large vocabulary to minimize out-of-vocabulary (OOV) rate**

Simple Morphological Decomposition (M1)

- Decomposed words into “morphemes” using a simple set of context-independent rules
  - Used a list of 12 prefixes and 34 “suffixes”
- Words belonging to the 128K most frequent decomposable words were not decomposed
- Recognition lexical units were morphemes that were composed back into words at the output stage

B. Xiang, et al., “Morphological Decomposition for Arabic Broadcast News Transcription,” ICASSP 2006

Sakhr Morphological Decomposition (M2)

- **Used Sakhr's context-dependent, sentence-level morphological analyzer to decompose each word into *[prefix] + stem + [suffix]***
- **Did not decompose the 128K most frequent decomposable words**

Comparison of OOV Rates

- Overall, morpheme-based systems (M1 and M2) have lower OOV rates than word-based systems (P and G)

System	vocab	dev07	eval07	dev08
Phonetic (P)	390K	4.36	2.88	1.44
Graphemic (G)	490K	3.78	2.07	0.84
Morpheme1 (M1)	289K	2.82	1.89	0.94
Morpheme2 (M2)	284K	0.81	0.66	0.56

- M2 system has a much lower OOV rate than M1 system

Performance Comparisons (WER %)

System	dev07	eval07	dev08
Phonetic (P)	10.6	11.6	12.1
Graphemic (G)	11.6	12.2	12.5
Morpheme1 (M1)	10.3	11.1	11.6
Morpheme2 (M2)	10.2	10.8	11.8

- Morpheme-based systems performed better than word-based systems
- Morpheme-based system (M2) based on Sakhr's morphological analysis had the lowest word error rate (WER) for most test sets

System Combination Using ROVER

ROVER	dev07	eval07	dev08
P+G	10.5	10.9	11.6
P+M1	10.1	10.9	11.4
P+M2	10.2	10.7	11.5
P+G+M1	9.9	10.6	11.0
P+G+M2	9.8	10.4	11.0
P+M1+M2	9.8	10.5	11.1
P+G+M1+M2	9.7	10.3	10.8

- Combination of all four systems (P+G+M1+M2) provided the best WER for all test sets

CU: Morphological Decomposition

- Decomposed words using MADA tools (v1.8)
  - Used option D2: separating prefixes and modifying stems (e.g. wll$Eb ==> w+ l+ Al$Eb)
  - Ngram-SMT-based MADA-to-word back mapping used
  - Reduced OOVs by 0.5-2.0% absolute
  - Approximately 1.19 morphemes per word
- Built a graphemic morpheme-based system (G_D2)
  - WER gains of up to 1.0% abs. over graphemic word baseline
  - Further gains from combining with phonetic word-based system

System	dev07	eval07	dev08
G_Word (P3a)	13.1	14.4	15.2
G_D2 (P3b)	12.5	13.6	14.2
V_Word (P3c)	11.6	13.2	14.2
P3a + P3c	11.5	12.7	13.4
P3b + P3c	11.0	12.1	12.0

LIMSI: 3 Variant Buckwalter Methods

- Affixes specified in decomposition rules (32 prefixes and 11 suffixes)
- Added 7 dialectal prefixes
- Variant 1: split all identifiable words with unique decompositions to have 270k lexicon of stems, affixes, and uncomposed words
- Variant 2: + did not decompose the 65k frequent words ==> 300k lexical entries
- Variant 3: + did not decompose ‘Al’ preceding solar consonants ==> 320k lexical entries
- Variant 3 slightly outperformed word-based systems
- Additional gain from ROVER with word-based systems

Conclusion

- **Morpheme-based systems perform better than word-based systems for Arabic STT**
- **Morphological decomposition of Arabic words taking their context into account produces better morphemes for morpheme-based Arabic STT**

Character vs Word Language Modeling for Mandarin

Lori Lamel

LIMSI

Motivation

- Is it better to use word-based or character-based models for Mandarin
- No standard definition of words, no specific word separators
- Characters represent syllables and have meaning
- Lack of agreement between humans on word segmentation
- Segmentation influences LM quality

Language Models for Chinese

- Recognition vocabulary typically includes words and characters (no OOV problem)
- Is there an optimal number or words?
- Is it viable to model character units?
- Is there a gain from combining word and character LMs?
- Range of options for combining LM scores (CU)
  - Hypothesis combination using ROVER
  - Linearly interpolate LM scores
  - Use lattice composition - log-linear score combination

Experimental Results

LM	1-best CER	Lattice CER
Word	5.1	1.7
Word -> Char	5.3	1.7
Char	6.9	2.9

- bnmdev07
- CER and lattice quality better for word LMs
- Deterministic constraints on words
- Pronunciation issues

Multi-Level Language Model Performance

- Performance evaluated on P2-stage CU-only system
  - Lattices generated using word LMs
  - New lattices generated by rescoring with character LMs
  - Linear combination of LM-scores no performance gain

LM	bnd06	bcd05	dev07	dev08	P2ns
Word (4-gram)	7.2	16.4	9.8	9.6	9.6
Character (6-g)	7.6	17.9	11	10.4	10.5
ROVER	7.1	16.5	10.2	10.4	9.8
Compose (log-linear)	7.1	16.3	9.7	9.6	9.4

- ROVER combination gave mixed performance
  - Confidence scores not accurate enough
- Lattice intersection (log-linear combination)
  - Consistent (small) gains over word-based system

MLP Features for STT

Jean-Luc Gauvain

LIMSI

Goals/Issues

- Improve acoustic models by using MLP-features
- Way to incorporate long term features such as wLP-TRAP which are high dimensional feature vectors (e.g. 475)
- Combination with PLP features (appending features, cross-adaptation, Rover)
- Model and feature adaptation
- Experiments on both the Arabic and Mandarin STT tasks (and other languages)
- Used in Jul'07 Arabic STT (LIMSI) system and Jul'08 Arabic and Dec'08 Mandarin systems (CUED, LIMSI)

Bottle-Neck MLP

- **4 layer network [Grezl et al, ICASSP'07]**
- **Input layer: 475 features (e.g. wLP-TRAP, 19 bands, 25 LPC, 500ms)**
- **2nd layer: 3500 nodes**
- **3rd layer: bottleneck features (LIMSI 39, CUED 26)**
- **Output layer:**
  - **LIMSI uses HMM state targets (210-250)**
  - **CUED uses phone targets (40-122)**

MLP Training

- Training using ICSI QuickNet toolkit
- Separate MLLT/HLDA transforms for PLP and MLP features
- Discriminative HMM training: MMI/MPE
- Single-pass retraining approach, use PLP lattices for MMI/MPE estimation of the PLP+MLP HMMs
- Experiment with various amount of training data to train the MLP:
  - WER is significantly better using entire training set

MLP-PLP Feature Combination (LIMSI)

- Experimented various combination schemes: feature vector concatenation, MLP combination, cross adaptation, …
- Evaluate 2 sets of raw features for MLP in combination with PLP (wLP-TRAP and 9xPLP)
- Evaluated cross-adaptation and rover combination
- Findings:
  - feature vector concatenation outperforms MLP combination
  - PLP+MLP combination outperfoms PLP features
  - MLP based on wLP-TRAP combines better than MLP based on 9xPLP
  - cross-adaptation and rover provide additional gains on top of feature combination

MLP Model Adaptation

- Experimented with CMLLR, MLLR, and SAT
- Findings:
  - standard CMLLR, MLLR and SAT techniques work for MLP features but the gain is less than with PLP features
  - after adaptation PLP+MLP combination still outperforms PLP features
    LIMSI: 1.0% absolute on Arabic
    CUED: 0.5% absolute on Arabic

CUED Specific Results for Arabic

- Combine a graphemic and phonemic system
- Use 40 phonemic targets for both systems
- MLP gives twice as much gain for the graphemic case than for the phonemic one (0.6 vs 0.3 for a 3-pass system)
- Implicit modeling of short vowels via the MLP features
- 0.5% absolute gain using 4-way combination over 2-way

Summary & Future Work

- MLP features based on wLP-TRAP are very effective in combination with PLP features
- Very significant gains have been obtained by using feature combination, cross-adaptation, and system output combination on both Arabic and Mandarin
- LIMSI also successfully used these features for Dutch and French
- Experimenting with alternative raw features to replace the costly wLP-TRAP features
- Linear adaptation of raw features in front of MLP
- Better feature combination schemes

Language Model Adaptation and Cross-Adaptation

Phil Woodland

University of Cambridge

Context Dependent LM Adaptation

- **Interpolated language models combines multiple text sources**
  - **allows weighting of LMs trained on different sources (e.g. text sources vs audio transcripts)**
  - **Can adapt weights on test data for particular test data types: normally do unsupervised adaptation to reduce perplexity**
- **"Usefulness" of sources vary between contexts:**
  - **influenced by: resolution, generalization, topics, styles, etc**
  - **global interpolation unable to capture context specific variability**
  - **context dependent interpolation weights used for LM adaptation**
- **Context dependent interpolation weights allows more flexibility**

$$P(w|h) = \sum_m \Phi_m(h)\, P_m(w|h)$$

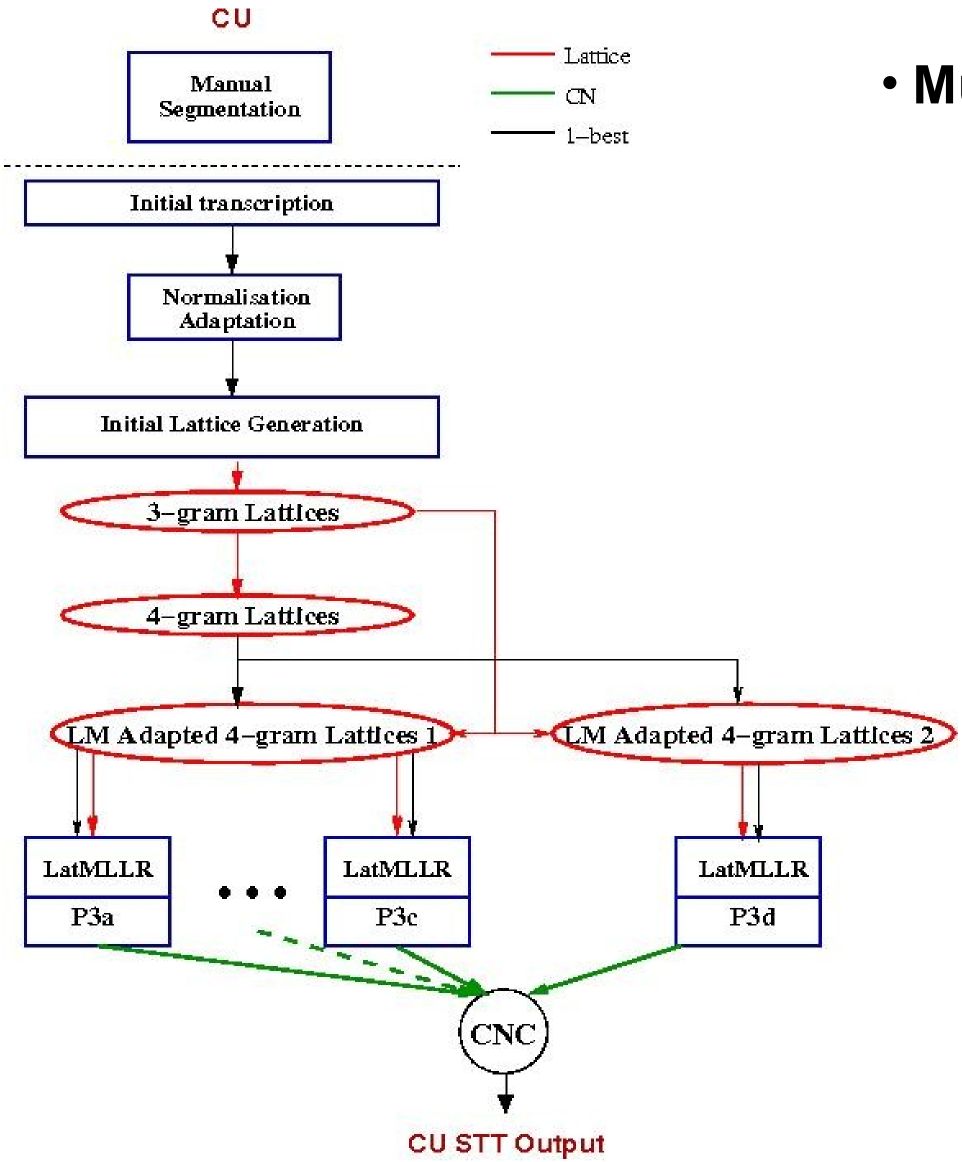
LM Adaptation Results

- **MAP adaptation used on test data**
  - Use hierarchical priors of different context lengths
  - Unsupervised adaptation for genre/style etc
  - Evaluated using single rescoring branch of Chinese CU system
  - CER improvements 0.4% abs

LM Adapt	eval06	eval07
No	16.4	9.5
Yes	16.0	9.1

- **Current/Future work**
  - CD weight priors estimated from training data
  - Discriminative weight estimation
  - More difficult to get improvements on Arabic

CU P3.5 Chinese STT System



- **Multi-pass combination framework**
  - **P3a: GD Gaussianised PLP system**
  - **P3b: GD PLP+MLP system**
  - **P3c: GD PLP (Gaussianised) +MLP**
  - **P3d: SAT Gaussianised PLP system**
  - **Rescore LM-adapted lattices**
  - **CNC combination gain over best branch typically 0.3% abs CER**

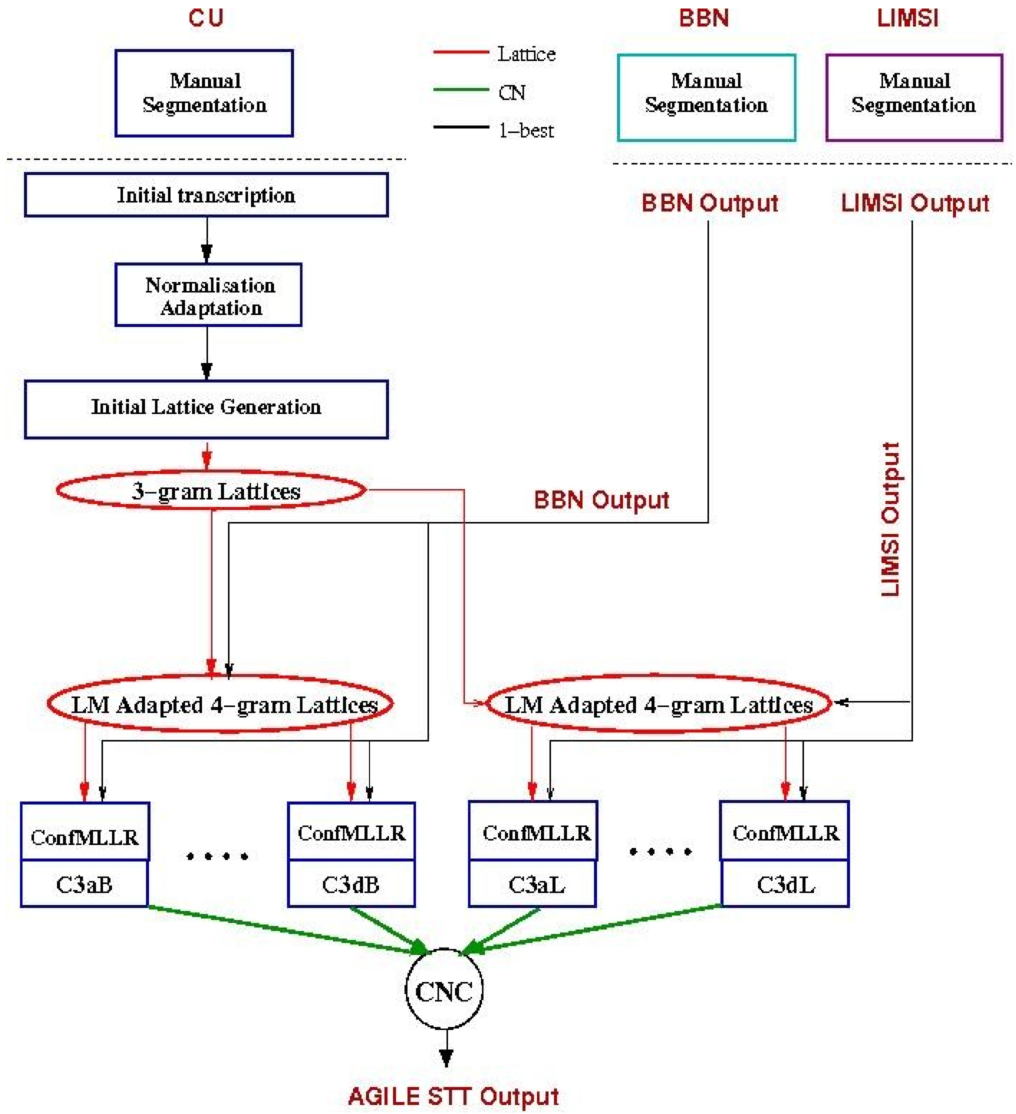
Language Model Cross-adaptation

- **Eval system combines outputs from multiple sites**
  - **Normally cross-adaptation transforms acoustic models only**
- **Also adapt language model used in rescoring**
  - **Context dependent adaptation**
  - **Confidence-based adaptation from 1-best of LIMSI and BBN outputs**

AGILE System	bnd06	bcd05	dev07	dev08	P2ns
ROVER	5.9	13.4	7.8	7.4	7.6
Xadapt (AM only)	5.8	13.6	7.8	7.4	7.6
Xadapt (AM+LM)	5.7	13.3	7.6	7.3	7.3

- **Consistent CER gains of 0.1%-0.3% over simple ROVER and acoustic model only cross-adaptation**

AGILE P3.5 Chinese STT System



Cross-adaptation framework

- BBN and LIMSI supervision
- CU system adapted
- Acoustic/LM adaptation
- Supervisions treated separately
- 4 cross-adapted branches for each of LIMSI and BBN supervision

8-way final combination

AGILE Chinese STT since P2.5 Eval

System	P2.5	P3.5
CU Dec 2007	8.9	12.0
CU Nov 2008	8.1	11.1
BBN Nov 2008	8.1	11.6
LIMSI Nov 2008	9.0	12.8
AGILE Dec 2007	8.0	11.1
AGILE Nov 2008	7.1	10.2

- **Significant improvements since P2.5 evaluation**
  - **CU system improved by 8%-9% relative**
  - **Combined AGILE system improved by 8%-11% relative**
  - **P3.5 data 3+% harder than P2.5 data**
  - **Tuned ROVER slightly lower CER: cross-adapt retained for MT**

Future Work in STT

Phil Woodland

University of Cambridge

Future Work: Core STT

- **Acoustic Model Training/Adaptation**
  - Improved discriminative training/large margin techniques
  - Discriminative adaptation (mapping transforms)
  - MLP features: improved inputs, better training/adaptation
  - Other posterior features
  - Accent/style dependent models
  - Explicit modelling of background/reverberant noise
- **Language Models**
  - Refinements of LM adaptations
  - Continuous space LMs (adaptation, fast training/decoding)
- **Improved Multi-Site System combination**
- **Sentence segmentation/punctuation estimation**

Future Work: Language Dependent

- **Arabic**
  - **Refined use of morphological decompositions**
  - **Use of generic vowel models**
  - **Automatic diacritisation of LM data**
  - **Dialect only models/systems**

- Chinese
  - Multi-level language models (character/word)
  - Compare/combine initial/final modeling with phone-based
  - Linguistically-driven word compounding
  - Improve accuracy on named entities