MITSUBISHI ELECTRIC
Changes for the Better

for a greener tomorrow

MITSUBISHI ELECTRIC RESEARCH LABORATORIES
Cambridge, Massachusetts

Integration of unsupervised acoustic, lexicon, and language models toward language acquisition from speech

Shinji Watanabe

JHU mini workshop

July 2012

My research background

- Automatic speech recognition
  - Acoustic modeling (mainly): Bayesian acoustic modeling, adaptation, discriminative training
  - Language modeling (sometimes): topic tracking language model
  - Discriminative model for WFST based ASR decoder

- I recently started unsupervised (zero resource) spoken language processing
  - Good application of Bayesian approaches
  - It’s ongoing research

Hierarchical dynamics in speech (recognition)

Frame unit (10ms) $o_1, o_2, o_3, o_4, o_5,$

HMM state unit(~50ms)

Phoneme unit (~0.1s) /w/, /a/, /t/, /a/, /n/, /a/, /b/, /e/

Word unit (~0.3s) Hello, my name is Watanabe.

N-gram unit(~1s) my+name+is, name+is+Watanabe,

Utterance unit (~5s)
Or chunk unit (~1m)

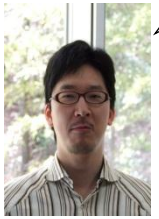
Hello

My name is Watanabe

Signal processing
Feature extraction

Acoustic model

Lexical model
Language model



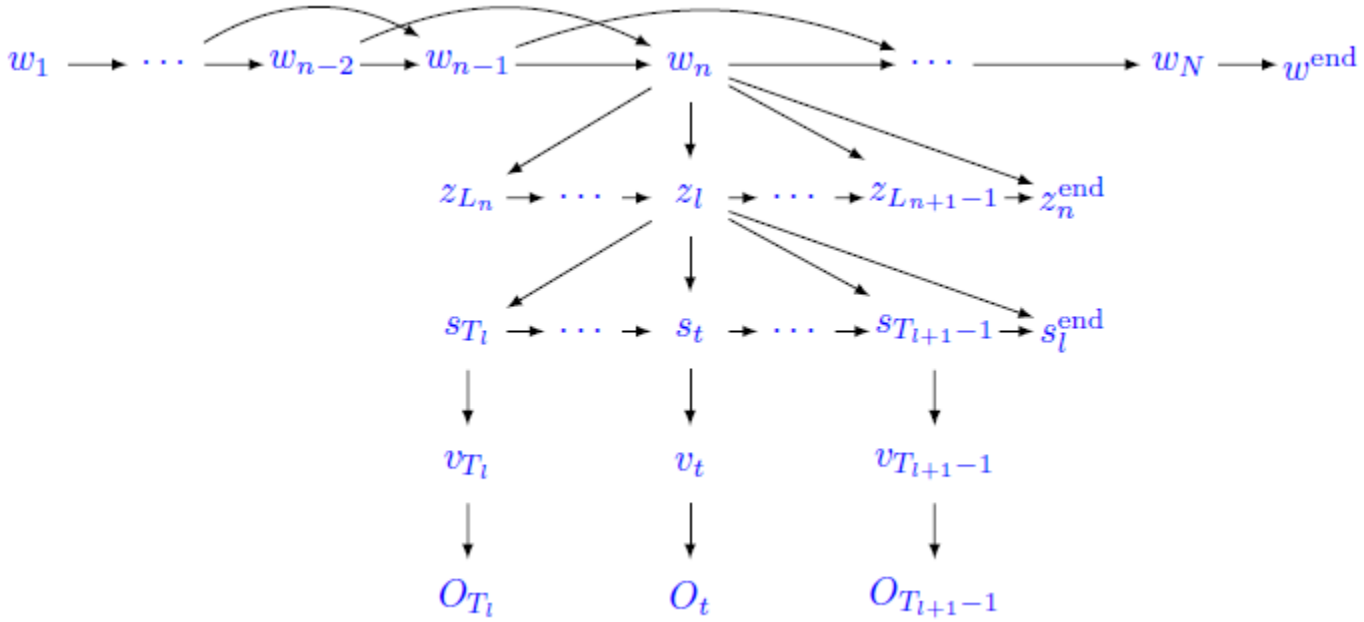
There are topic transition
speaking style changes, speaker changes

Strategy

- Represent a dynamics on each layer of speech with a straightforward generative model
  - Phoneme: HMM, word: n-gram

  Basically using automatic speech recognition techniques

5+ layer speech generative model



Speaker change dynamics
Topic dynamics
…..

Word n-gram

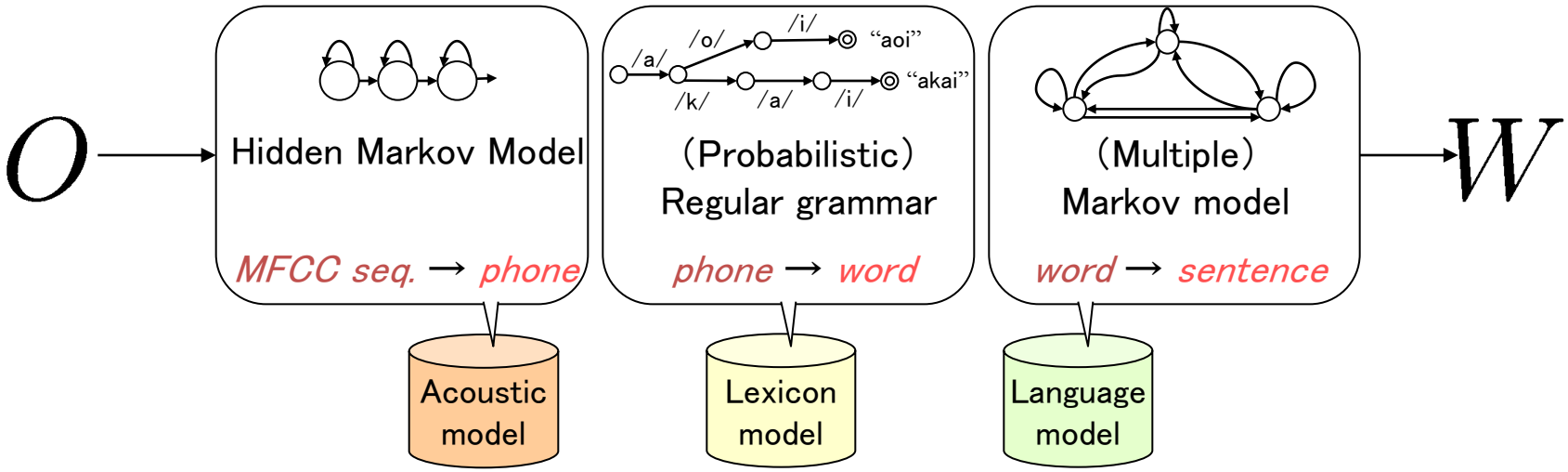
Phoneme n-gram

Left to right HMM

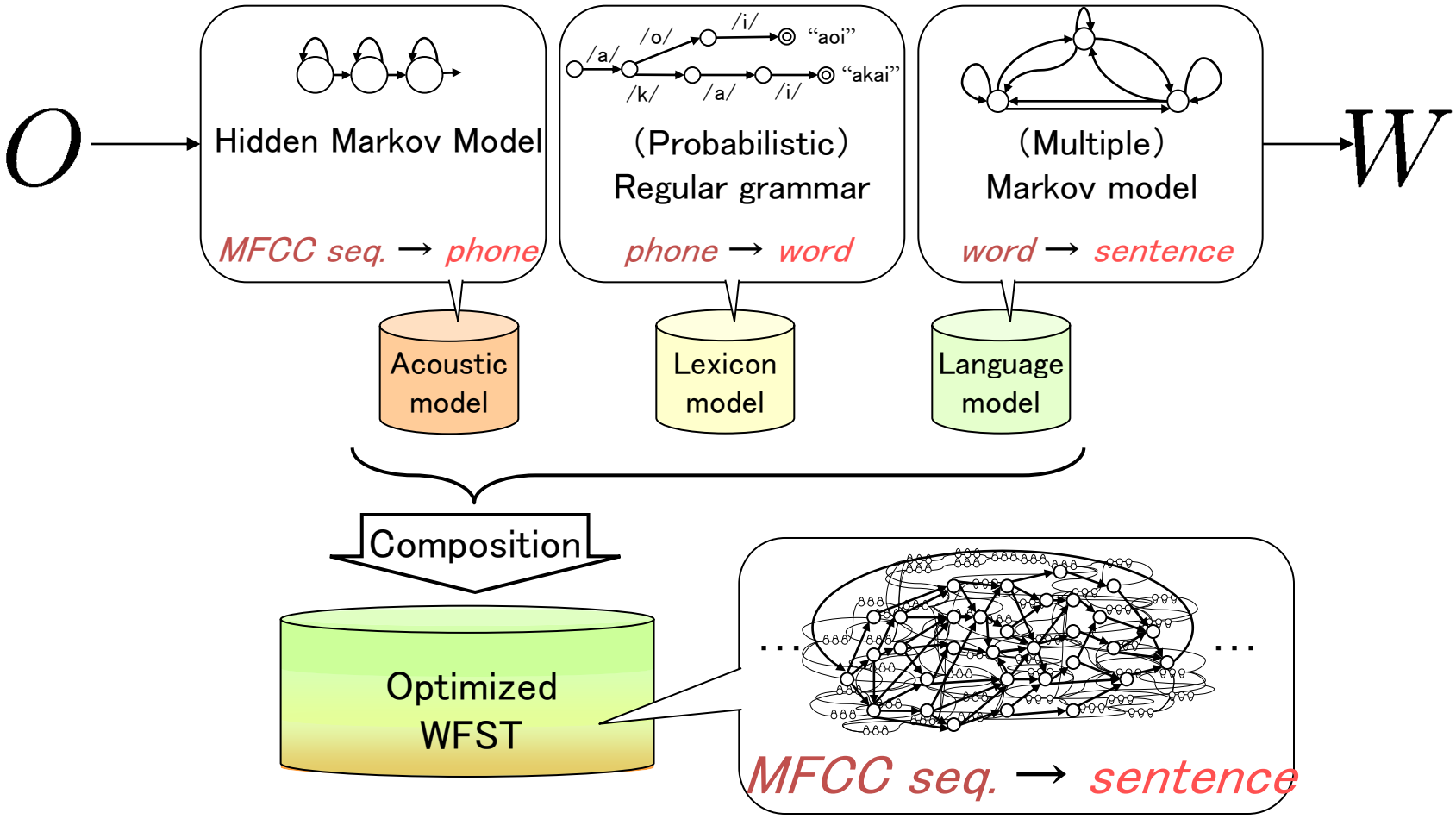
GMM

Speech feature (MFCC)

WFST based representation



WFST based representation

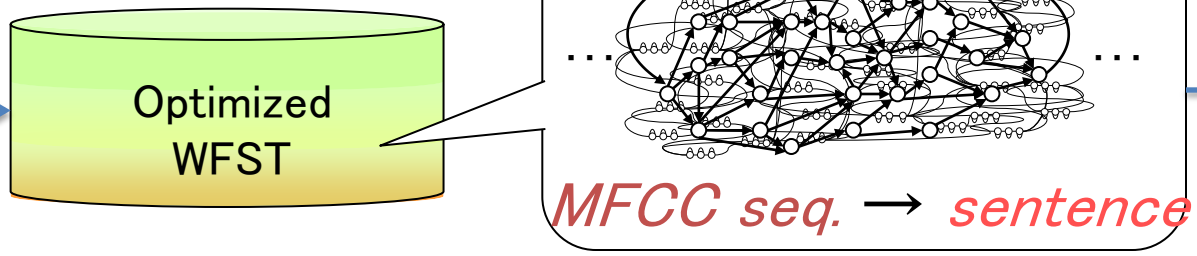


WFST based representation

O

W

Flat representation
⇒ makes inference algorithm simple



Inference algorithm (skip details)

- Acoustic modeling:
  - Segmental k-means (ML-EM) or utterance-unit blocked Gibbs (not a nonparametric Bayes) using WFST

$$p(\mathbf{O},\mathbf{Z},\mathbf{S},\mathbf{V},\mathbf{W}|\mathbf{\Theta}) = p(w^{end}|w_N)\prod_{n=1}^{N} p(w_n|w_{n-1},w_{n-2})p(z_n^{end}|z_{L_{n+1}-1}) \prod_{l=L_n}^{L_{n+1}-1} p(z_l|z_{l-1},w_n)p(s_l^{end}|s_{T_{l+1}-1}) \prod_{t=T_l}^{T_{l+1}-1} p(s_t|s_{t-1},z_l)p(v_t|s_t)p(\mathbf{o}_t|v_t) \quad (2)$$

where

$$p(w_n|w_{n-1},w_{n-2}) = \begin{cases} p(w_n) & n=1 \\ p(w_n|w_{n-1}) & n=2 \\ p(w_n|w_{n-1},w_{n-2}) & n>2 \end{cases} \quad (3)$$

$$p(z_l|z_{l-1},w_n) = \begin{cases} p(z_l|w_n) & l=L_n \\ p(z_l|z_{l-1}) & L_n<l<L_{n+1} \end{cases} \quad (4)$$

$$p(s_t|s_{t-1},z_l) = \begin{cases} p(s_t|z_l) & t=T_l \\ p(s_t|s_{t-1},z_l) & T_l<t<T_{l+1} \end{cases} \quad (5)$$

Inference algorithm (skip details)

- Acoustic modeling:
  - Segmental k-means (ML-EM) or utterance-unit blocked Gibbs (not a nonparametric Bayes)

Algorithm 2 Unsupervised language acquisition.

Initialize $\Psi \leftarrow \Psi^0$
Sample $\mathbf{Z} \sim p(\mathbf{Z}|\Psi)$
for $\tau = \{1, \cdots\}$ **do**
  Compute $\{\Omega^u\}_{u=1}^U$
  for $u =$ shuffle $(1, \cdots, U)$ **do**
    Update $\Omega^{\setminus u}$
    Update $\Psi \leftarrow \Psi^{\setminus u}$
    Prune $q(\mathbf{Z}^u)$ from $p(\mathbf{Z}^u|\Psi)$
    Sample $\mathbf{Z}^u \sim q(\mathbf{Z}^u)$
    Compute Ω^u from $\mathbf{Z}^u$
  end for
  Sample $\mathbf{W} \sim p(\mathbf{W}|\mathbf{O}, \mathbf{Z})$
  Sample $\mathbf{Z} \sim p(\mathbf{Z}|\mathbf{O}, \mathbf{W})$
end for

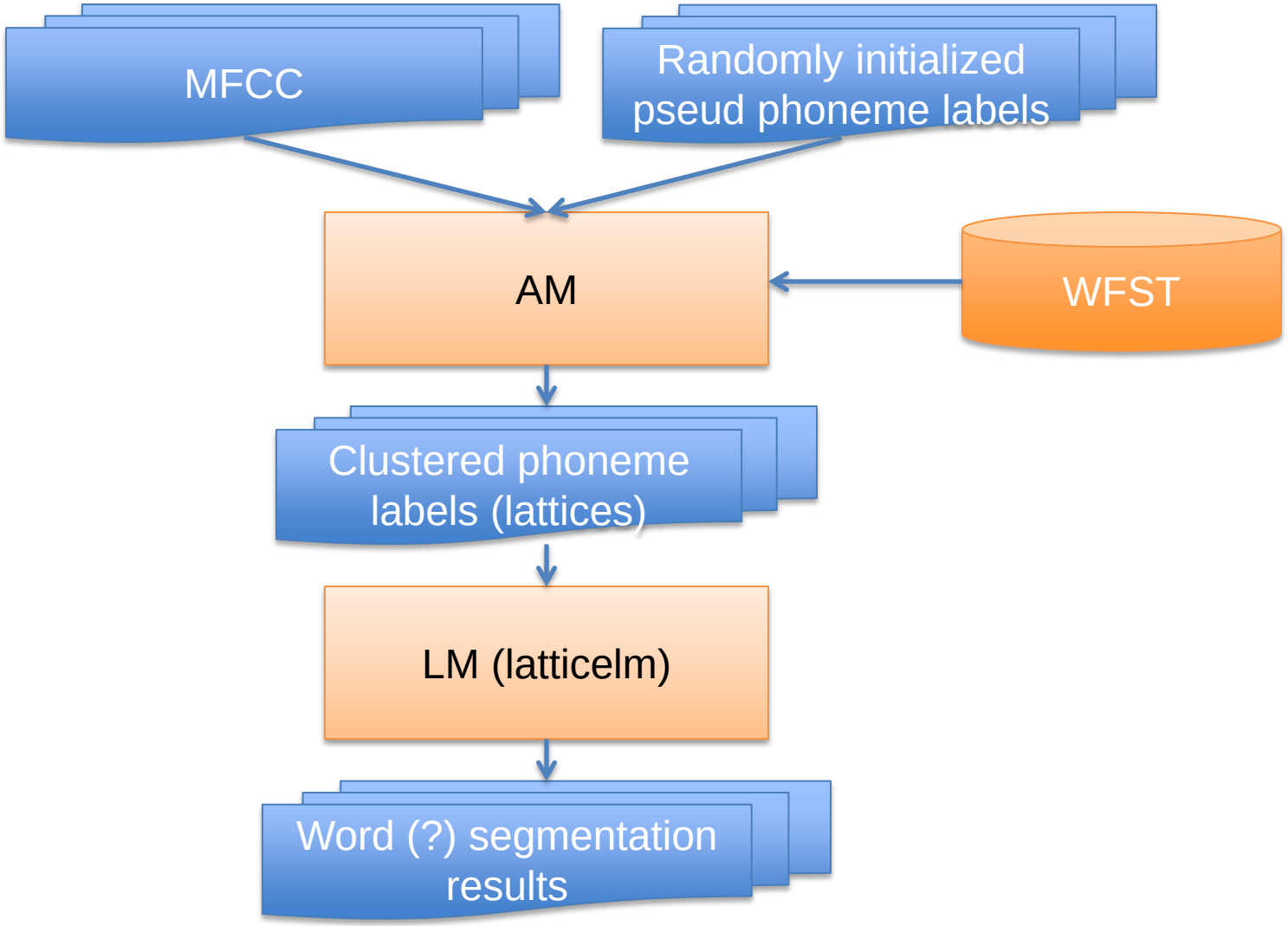
- Lexicon and language modeling:
  - Gibbs sampling based on Hierarchical Pitoman-Yor Process [Mochihashi (2009), Neubig (2010)]

LatticeIm (Graham Neubig)

- Open source tool for word segmentation based on HPY
  - Input: phoneme sequences for “phoneme latteices”
  - Output: word sequences
- Implemented based on openfst

  ⇒ easily integrated with other components

Flow chart



Examples of pseud phoneme labels

- Random initialization

```
1p 31p 31p 35p 20p 21p 32p 2p 11p 17p 13p 31p 19p 43p 28p 21p 20p 34p 21p 33p 4p 22p 14p 9p 26p 24p
28p 37p 43p 25p 16p 23p 40p 43p 40p 10p 48p 31p 5p 36p 48p 20p 33p 16p 12p 35p 1p
```

  - Uniformly samples numbers from [1p, 2p, …, 48p]. The length of phoneme sequence is proportional to the length of utterance
  - Beginning and end phonemes are fixed as “1p” <-silence
- After unsupervised AM

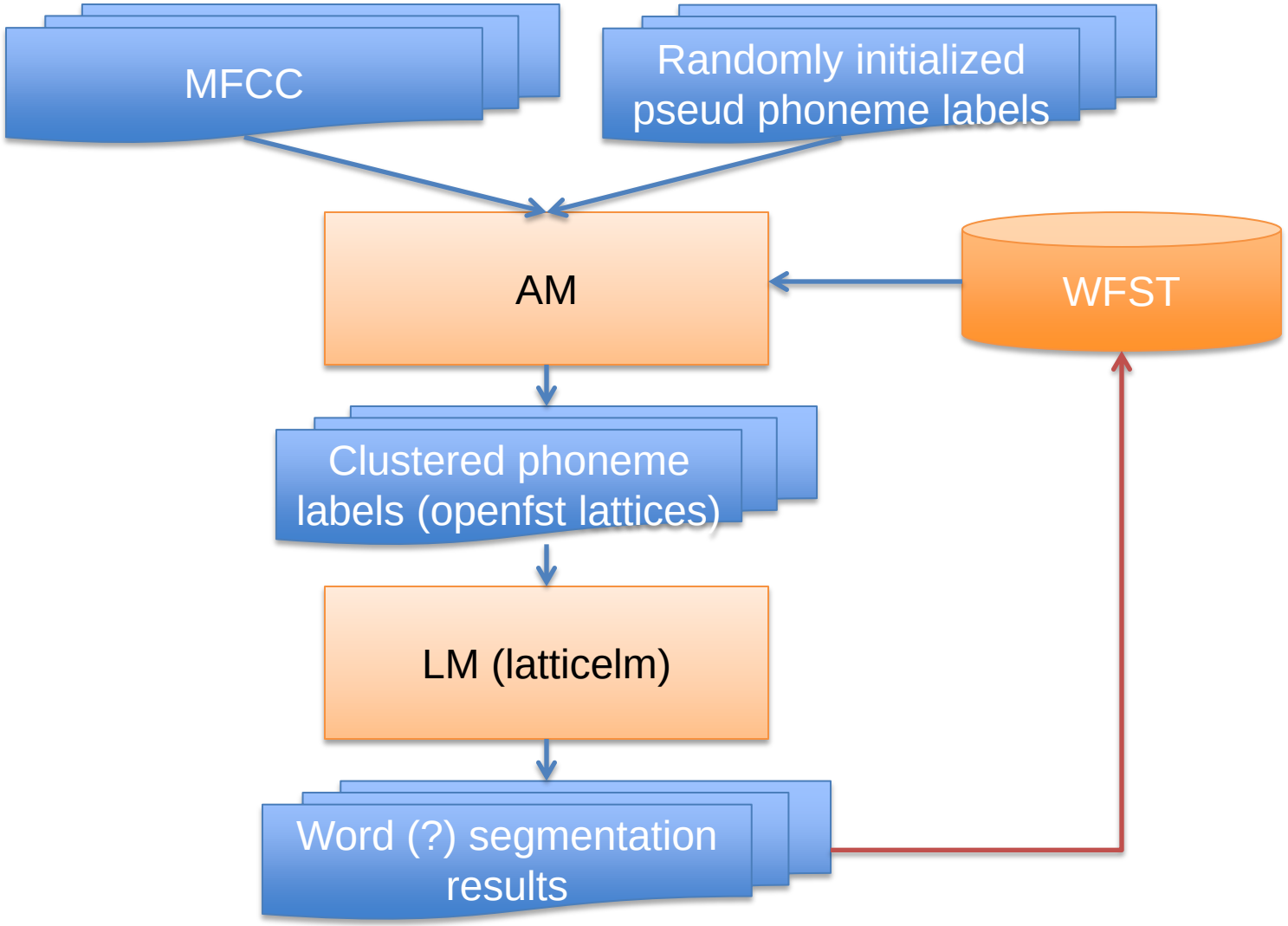
```
1p 45p 18p 33p 45p 31p 21p 9p 21p 18p 17p 29p 19p 42p 21p 14p 29p 18p 24p 29p 30p 40p 29p 19p 29p 32p
40p 31p 9p 21p 4p 41p 27p 4p 18p 12p 28p 4p 5p 4p 15p 42p 11p 1p
```

  - Segmental k-means
- After unsupervised LM

```
1p 45p18p 33p45p31p 21p9p 18p 17p29p 19p42p21p 14p 28p18p 24p 30p 40p 29p19p 29p32p 40p 31p 21p4p 41p
4p 18p 12p28p 4p 5p4p 15p 42p11p
```

  - Concatenating phonemes to obtain word sequences

Flow chart (feed back from LM)



Openfst format of lattices

0	1	1p	1p	486.167
1	2	40p	40p	55.7665
2	3	47p	47p	173.766
3	4	33p	33p	357.023
3	87	43p	43p	161.133
4	5	38p	38p	586.824
5	6	32p	32p	311.078
6	7	31p	31p	659.849
6	81	39p	39p	662.414
7	8	3p	3p	518.258
8	9	46p	46p	266.61
9	10	14p	14p	571.64
9	70	4p	4p	256.732
10	11	10p	10p	323.303
10	72	19p	19p	625.699
10	77	13p	13p	163.565
11	12	19p	19p	301.427
12	13	48p	48p	208.436
13	14	13p	13p	263.555
14	15	18p	18p	510.772
15	16	42p	42p	210.877
15	83	46p	46p	215.286

- Weight: acoustic score (minus log likelihood)+ language scores (minus log probability with a scaling factor)

Experiments

"Preliminary experiments"

Experimental condition

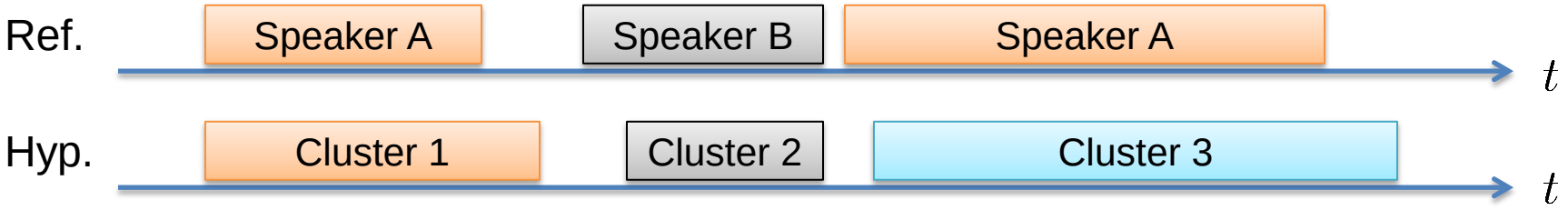
- TIMIT female training set (112 utterances, 1088 utterances)
- Acoustic modeling (Kaldi: tightly integrated with openfst)
  - Acoustic condition

Sampling Rate	16 kHz
Quantization	16 bit
Feature Vector (39 dims.)	12 - order MFCC wih log energy with Δ and $\Delta\,\Delta$
Window	Hamming
Frame Size/Shift	25/10 ms

  - Acoustic model
    - **48**, 100, 500, 1000 phonemes
    - 3 state left-to-right HMMs
    - 8 Gaussian mixture components
  - Phoneme bigram:
- Lexicon and language modeling (latticelm)
  - Phoneme 3gram, Word 3gram

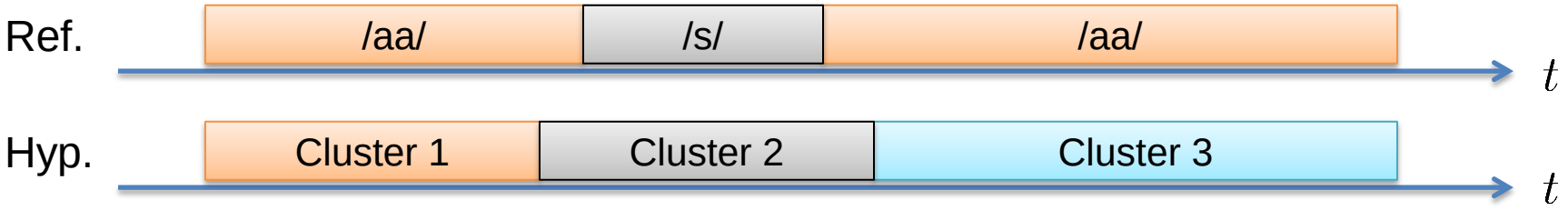
Performance criterion

- Diarization error rate
  - Who speaks when?



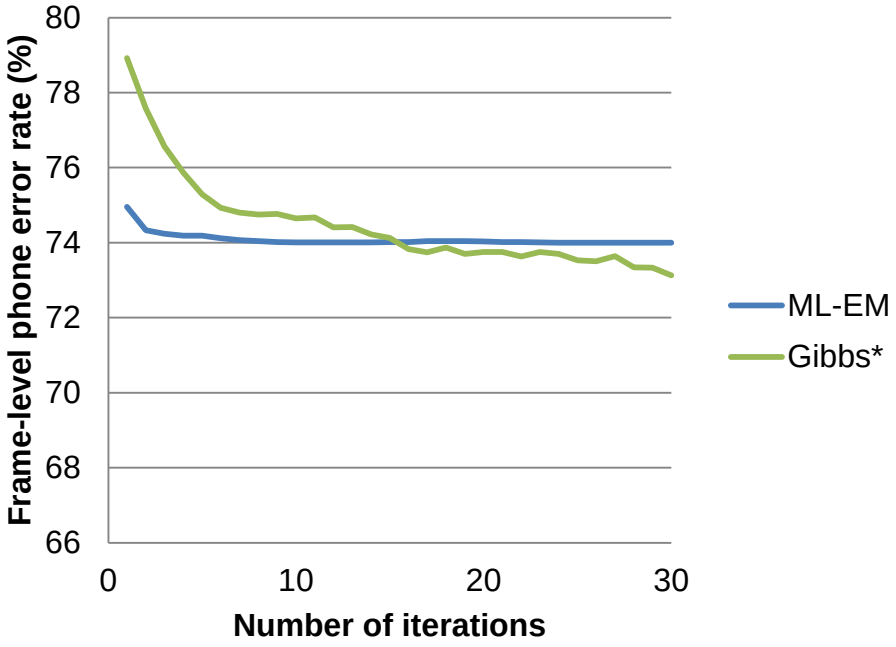
$$DER = \frac{\sum_{s=1}^{S} \mathrm{dur}(s) \cdot (\max(N_{ref}(s), N_{hyp}(s)) - N_{correct}(s))}{\sum_{s=1}^{S} \mathrm{dur}(s) \cdot N_{ref}}$$

  - Changing from speaker clusters to phoneme clusters
    - Phoneme alignment (obtained by Viterbi algorithm)

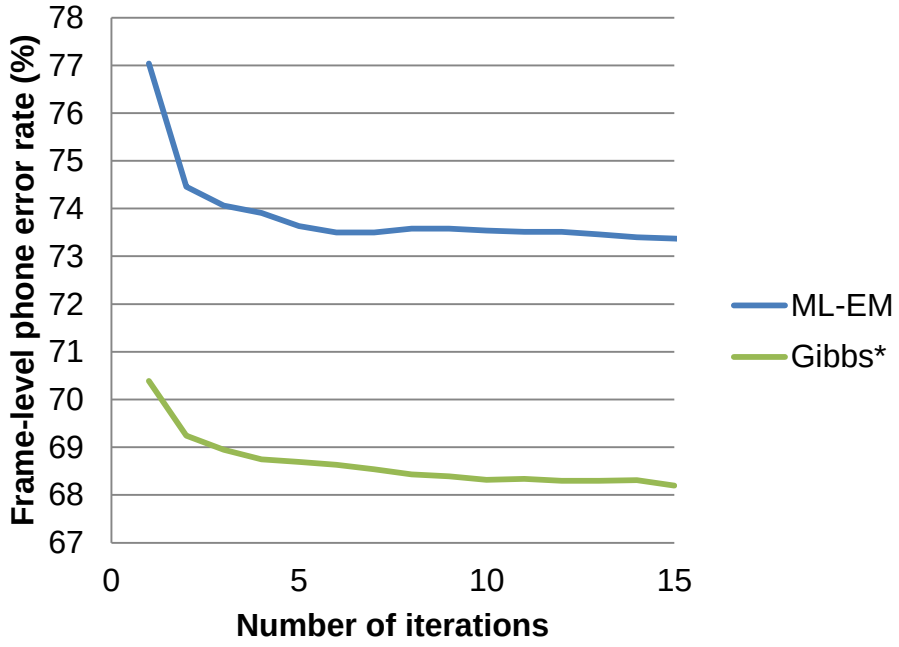


Experimental results (unsupervised AM)

Small data (112 utt.)



Large data (1,088 utt.)



- It seems to work to some extent, but not good performance
- Gibbs would be better (mitigating the local optimum problem (?)), but it takes so much time

Example (* different experimental setup)

Table 2: An example of word acquisition

tie	/t ay/	out	/aw t/	ill	/ih l/	per	/p er/
in	/ih n/	it	/ih t/	ice	/ay s/	sea/see	/s iy/
law	/l aa/	i'm	/ay m/	sew	/s ow/	so	/s ow/
ought	/aa t/	is	/ih z/	they	/dh ey/	jaw	/jh aa/
all	/aa l/	toy	/t oy/	lie	/l ay/	sir	/s er/
saw	/s aa/	pa/paw	/p aa/	i'll	/ay l/	aid	/ey d/
off	/aa f/	may	/m ey/	'em	/ah m/	it	/ih t/
oil	/oy l/	my	/m ay/	sigh	/s ay/	i'd	/ay d/
lie	/l ay/	pit	/p ih t/	low	/l ow/	ate	/ey t/
eight	/ey t/	they	/dh ey/	pie	/p ay/	zoo	/z uw/
pull	/p uh l/	hull	/hh ah l/	mine	/m ay n/	eyes	/ay z/
pause	/p aa z/	us	/ah s/	sigh	/s ay/	ease	/iy z/
raw	/r aa/						

- I have not tried any objective evaluation for word segmentation yet.
- Some basic syllables seemed to be extracted (?)
- So many errors were included in this conversion
  - Errors in unsupervised acoustic and language modeling
  - Token to phoneme mapping
- Feed back from ULM to UAM currently did not improve performance

Summary

- Unfortunately the approach did not work well due to bugs? mistakes?, but it would be a good starting point for us toward language acquisition from speech
- I don't use any label information at all, but use some knowledge
  - Average phoneme length, number of phonemes, ASR knowledge
- Unsupervised acoustic modeling should be improved
  - Feature, model, training method
- Feed back part is not tightly integrated (not fst)
- Evaluation measure
  - DER would not be a good measure
- Model complexity control in acoustic modeling
  - Nonparameteric Bayes by Lee and Glass (2012) or Tawara et al (2012)

Tools

- Kaldi (ASR tool)

  http://kaldi.sourceforge.net/index.html
- Openfst

  http://www.openfst.org/twiki/bin/view/FST/WebHome
- Latticelm

  https://github.com/neubig/latticelm
- Diarization error rate

  http://www.itl.nist.gov/iad/mig/tests/rt/2006-spring/code/md-eval-v21.pl

My final goal

Unsupervised generative models of whole speech communication

- Who speak when, what?
- What is a topic?
- What kind of room environment, atmosphere, role, emotion

Learning generative models involving everything related to speech communication

Thank you for your attention