Introduction

1. **Objective**

   To investigate the effectiveness of incorporating phone deletions explicitly in whole word acoustic models.

2. **Motivations**

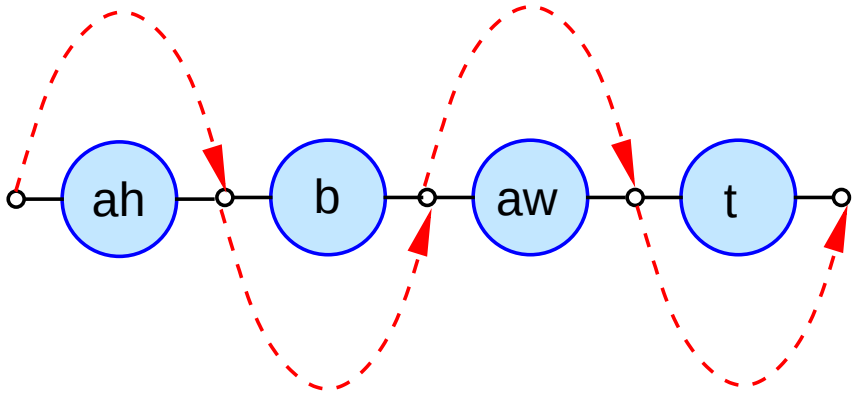
   - Phone deletion rate is about 12% in Switchboard.
     [Greenberg, "Speaking in shorthand — a syllable-centric perspective for understanding pronunciation variation," ESCA Workshop 1998]
   - Phone deletions cannot be modeled well by triphone training.
     [Jurafsky, "What kind of pronunciation variation is hard for triphones to model?" ICASSP 2001]

3. **Proposal**

   Context-dependent fragmented word models (CD-FWM) bootstrapped from tied-state cross-word triphones.

Explicit Modeling of Phone Deletions

Conceptually, we may explicitly model phone deletions by adding skip arcs.



- Practically, it requires a unit bigger than a phone to implement the skip arcs.
- Problems (assume a vocab size of W and a phone set of N):
  - large number of context-dependent models: W^3 vs. N^3
  - training data sparsity
  - need to develop new state-tying rules for the new units

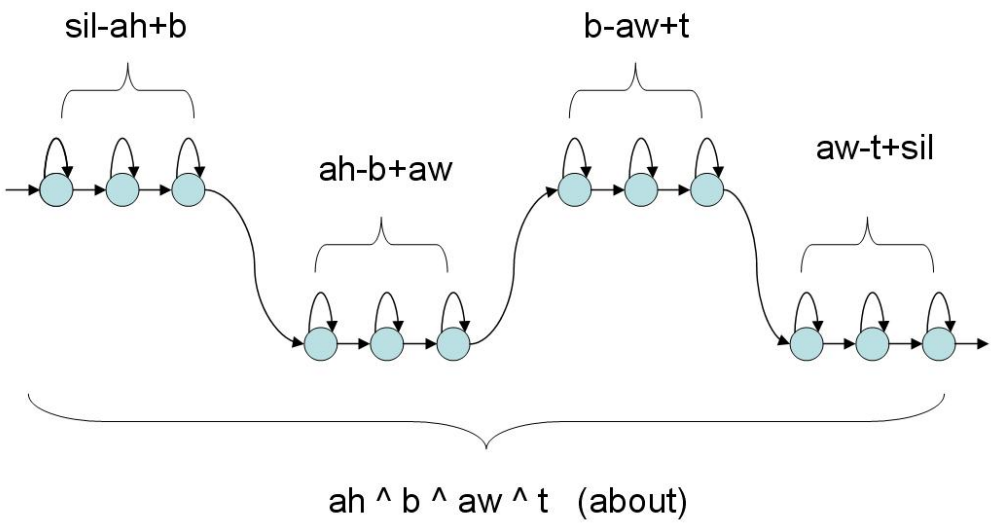
Context-dependent Fragmented Word Model (1)

Bootstrap word models from tied-state cross-word triphone models

$\Rightarrow$ solve the problem of data sparsity, and

$\Rightarrow$ reuse the existing state-tying rules.

Word	Phonemic Transcription	Construction from Triphones
"about"	ah b aw t	sil-ah+b ah-b+aw b-aw-t aw-t+sil



Context-dependent Fragmented Word Model (2)

Fragmented word models consist of word fragments:

- middle fragments are multi-phone sub-word units (MP-SWU)
- beginning and ending fragments are single phones

$\Rightarrow$ cross-word contexts do not affect the middle MP-SWUs

$\Rightarrow$ great reduction in the number of context-dependent fragments.

Word	Phonemic Transcription	Construction from Multi-phone Sub-word Units
"consider"	k ah n s ih d er	k ah n^s^ih d er

CD-FWM				
?-k+ah	k-ah+n^s^ih	ah-n^s^ih+d	n^s^ih-d+er	d-er+?

- In the "consider" example
  - 1st, 2nd, 4th, 5th fragments are simply single phones
  - 3rd middle fragment is a MP-SWU.

Fragmentation Reduces the Number of CD Models

Observation: multi-phone sub-word units are almost unique.

$$\begin{aligned} \text{vocabulary size} &= 5000 \\ \#\ \text{base phonemes} &= 40 \end{aligned}$$

	Not Fragmented	Fragmented
CI mono-units	ah^b^aw^t	ah b^aw t
CD tri-units	?-ah^b^aw^t+?	?-ah+b^aw ah-b^aw+t b^aw-t+?
#Models	$40 \times 5K \times 40 = 8M$	$40 \times 5K + 5K + 5K \times 40 = 0.4M$

Context-dependent Fragmented Word Model (3)

Fragmentation scheme:

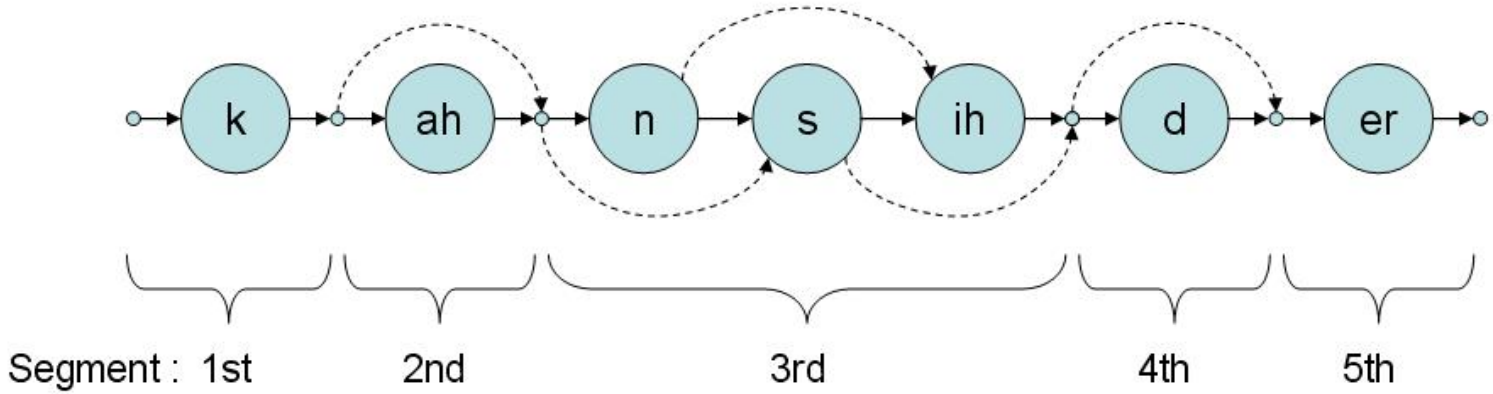
#fragments depends on L = #phonemes in a word.

- $L \leq 3$: a word is represented by its cross-word triphones; no phone deletions are allowed.
- $L = 4$ or 5: a word is split into 3 fragments
  - 1st and 3rd fragments are simply single phones
  - 2nd fragment is a MP-SWU.
- $L \geq 6$: a word is split into 5 fragments
  - 1st, 2nd, 4th, 5th fragments are simply single phones
  - 3rd fragment is a MP-SWU.

Context-dependent Fragmented Word Model (4)

Skip arcs are added with the following restrictions

- the first phone is not skipped
  $\Leftarrow$ syllable onset is well preserved.
- the last phone is not skipped
  $\Leftarrow$ to reduce the number of additional cross-word models.
- 2 successive phones in a MP-SWU are not skipped
  $\Leftarrow$ some technical reason.



Derivation and Training of CD-FWM

1. The canonical pronunciations in the dictionary are replaced by the corresponding FWM fragments.
2. The required models in the CD-FWM system: cross-word triphones, additional CD phones, CD MP-SWUs are constructed from the baseline tied-state cross-word triphones.
3. Skip arcs are added to the additional CD phones and CD MP-SWUs to allow phone deletions in words (with some restrictions).
4. The new CD-FWMs with skip arcs are re-trained for 4 EM iterations.
   - all model parameters and skip arc probabilities are re-estimated.

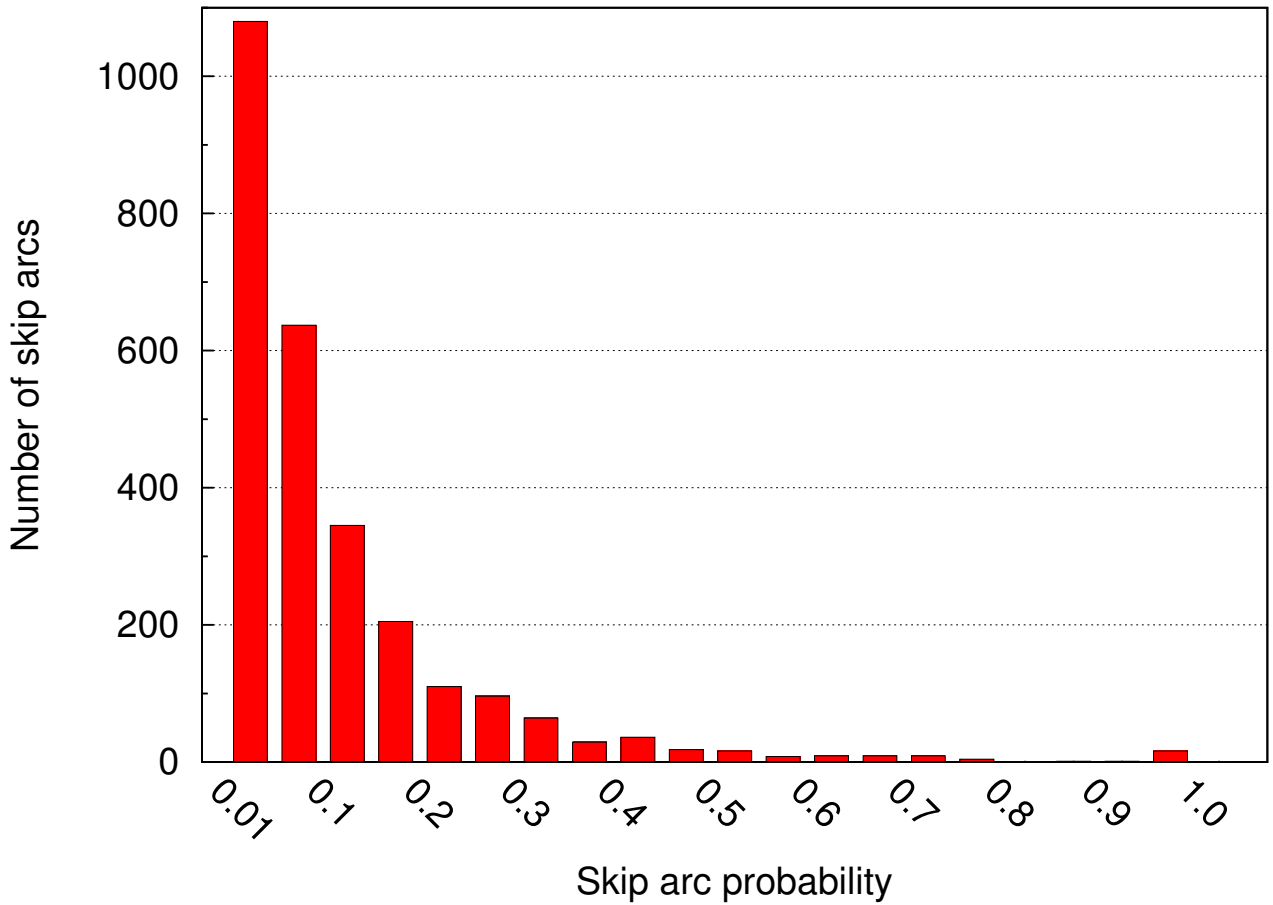
WSJ1 Evaluation

- Training Set : WSJ0 + WSJ1 (46995 utterances), about 44 hours of read speech, 302 speakers
- Dev. Set :
  - WSJ0 Nov92 Evaluation Set (330 utterances)
  - WSJ 5k development set (496 utterances)
- Test Set : WSJ1 Nov93 Evaluation Set (205 utterances)
- #Base Phonemes : 39
- #Triphones : 62402
- #HMM Tied States : 5864
- #Gaussian/State : 16
- #State/Phone : 3
- Language Model : bigram
- Acoustic Feature : 39-dimensional MFCC vector

Empirical Results

Model	#CD Phones	#CD MP-SWUs	Word Acc.
cross-word triphones	62,402	0	91.53%
CD-FWM for $L \geq 6$:			
no phone deletion	79,767	9,117	91.55%
+ phone deletion	79,767	9,117	92.10%
CD-FWM for $L \geq 4$:			
no phone deletion	380,341	13,907	91.58%
+ phone deletion	380,341	13,907	92.05%

Distribution of the Phone Deletion Probabilities



- System: CD-FWMs with $L \geq 6$.
- Those with a probability ≤ 0.01 are removed from the plot.
- Out of the total 52,507 phone deletion skip arcs, 49,814 (about 95%) of them have a probability ≤ 0.01.

Summary & Future Work

- We proposed a method of modeling pronunciation variations from the acoustic modeling perspective.
- The pronunciation weights are captured naturally by the skip arc probabilities in the context-dependent fragmented word models (CD-FWM).
- During the re-estimation of the model parameters of the CD-FWMs, some word-specific information may have been captured.
- At this moment, since states are tied across all acoustic units, word-specific information are not captured by the re-estimated state distributions.
  $\Rightarrow$ Future work will consider how to untie some states.
- Right now, we did not delete the last phoneme in a word.
  $\Rightarrow$ Future work will deal with this limitation as well.