SJTU SPEECH LAB

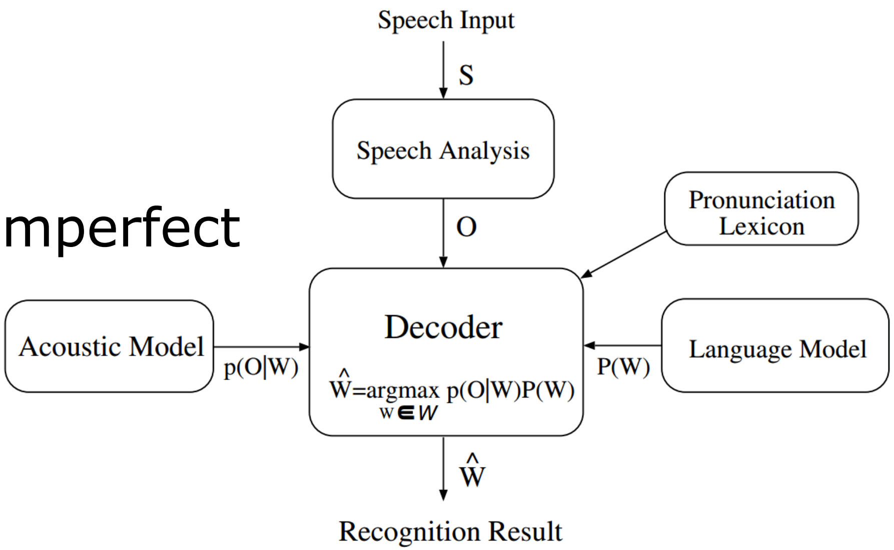
上海交通大學智能語音實驗室

Confidence Measures for CTC-based Phone Synchronous Decoding

Zhehuai Chen, Yimeng Zhuang, Kai Yu

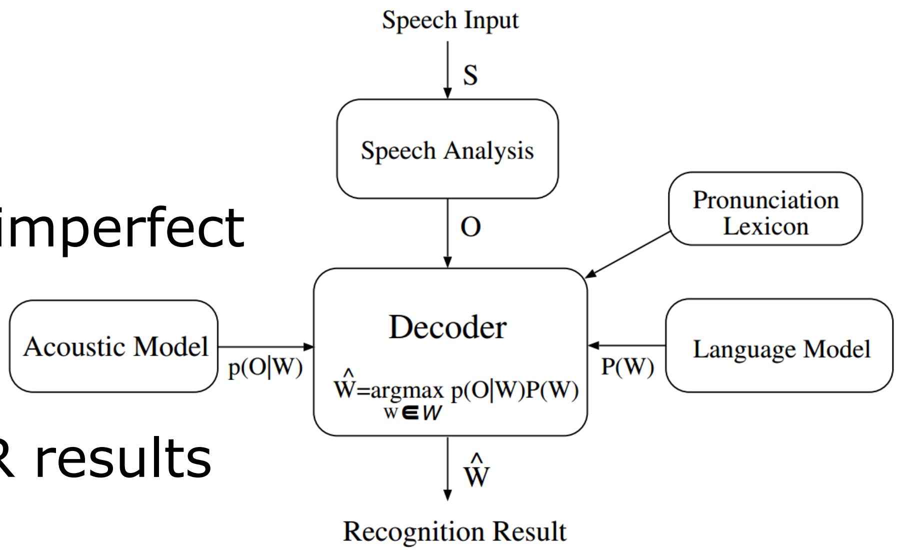
Introduction
ASR Decoding & Confidence Measure

- ASR Decoding
  - Inference by AM/LM/lex …
  - Model and search are both imperfect



Introduction
ASR Decoding & Confidence Measure

- ASR Decoding
  - Inference by AM/LM/lex …
  - Model and search are both imperfect
- Confidence Measure (CM)
  - Reliability evaluation of ASR results
  - Traditional CM
    - Predictor features based CM
      - Acoustic score, duration, entropy … (NOT ideal)
      - CRF, NN … (need training stage; train ≠ test)
    - Hypothesis Posterior based CM
      - Theoritically sounder



Introduction
Hypothesis Posterior based CM

- ASR as the *maximum a posterior* (MAP) decision process

$$\widehat{W} = \arg\max_{W \in \Sigma} p(W \mid X)$$
$$= \arg\max_{W \in \Sigma} \frac{p(X \mid W) \cdot p(W)}{p(X)}$$

$$p(X) = \sum_{H} p(X, H) = \sum_{H} p(H) \cdot p(X \mid H)$$

- H is from lattice/filler
  - Both imperfect

Introduction
Hypothesis Posterior based CM

- ASR as the *maximum a posterior* (MAP) decision process

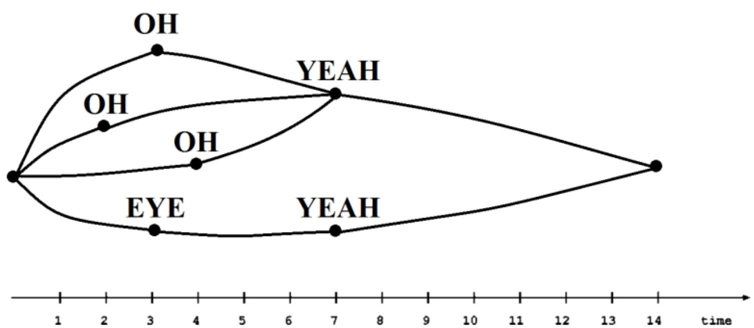
$$\widehat{W} = \arg\max_{W\in\Sigma} p(W \mid X)$$

$$= \arg\max_{W\in\Sigma} \frac{p(X \mid W)\cdot p(W)}{p(X)}$$

$$p(X) = \sum_H p(X,H) = \sum_H p(H)\cdot p(X \mid H)$$

- H is from lattice/filler
  - Both imperfect

- **Lattice quality** is the bottleneck
  - Not compact
    - Boundary unstable



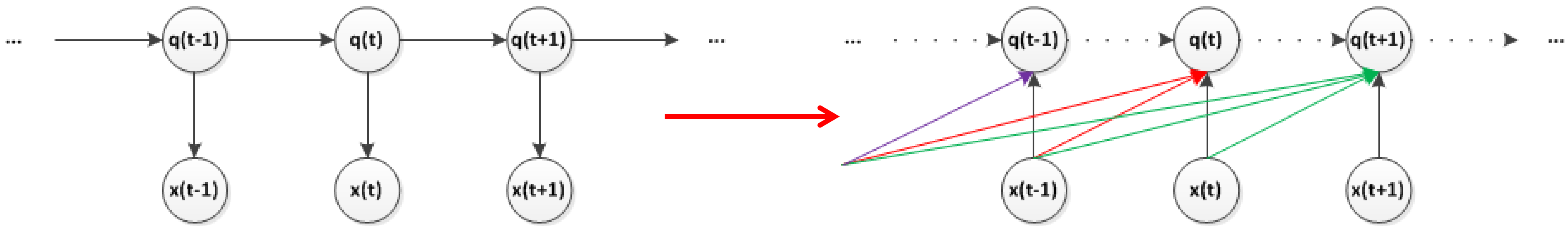
  - Not precise
    - Beam prune

WE NEED NEW MODEL !

上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

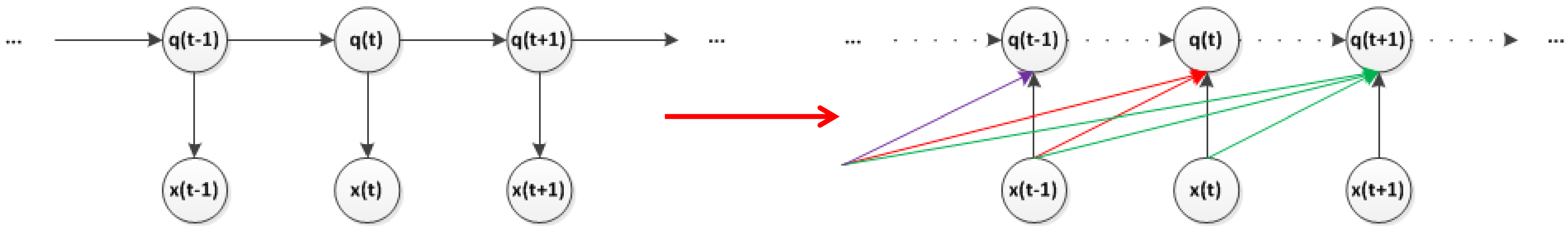
Introduction
from HMM to CTC acoustic model

- From HMM to CTC: do better in *sequential modeling*



Introduction
from HMM to CTC acoustic model

- From HMM to CTC: do better in *sequential modeling*



- CTC model: learn the many-to-one function of $\mathcal{B}$

$$P(\mathbf{l}|\mathbf{x}) = \sum_{\boldsymbol{\pi} \in \mathcal{B}^{-1}(\mathbf{l})} P(\boldsymbol{\pi}|\mathbf{x}) = \sum_{\pi : \pi \in L', \mathcal{B}(\boldsymbol{\pi}_{1:T}) = \mathbf{l}} \prod_{t=1}^{T} y_{\pi_t}^{t}$$

$$\mathcal{B} : L' \mapsto L$$

$$L' = L \cup \{\texttt{blank}\}$$

Introduction
from HMM to CTC acoustic model

- From HMM to CTC: do better in *sequential modeling*

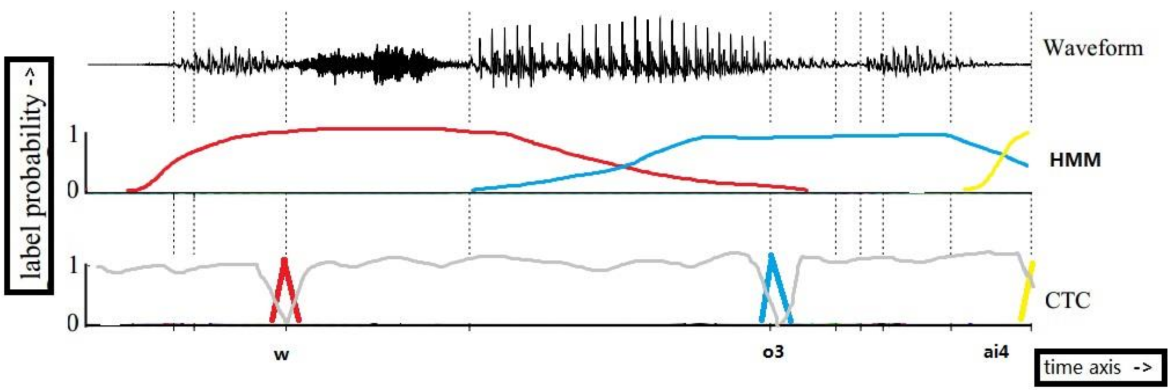
- CTC model: learn the many-to-one function of $\mathcal{B}$

$$P(\mathbf{l}|\mathbf{x}) = \sum_{\boldsymbol{\pi}\in\mathcal{B}^{-1}(\mathbf{l})} P(\boldsymbol{\pi}|\mathbf{x}) = \sum_{\pi:\pi\in L',\mathcal{B}(\boldsymbol{\pi}_{1:T})=\mathbf{l}} \prod_{t=1}^{T} y_{\pi_t}^{t}$$

$$\mathcal{B} : L' \mapsto L$$

$$L' = L \cup \{\texttt{blank}\}$$

- peaky distribution and concentrated information output



Frame Sync. to Phone Sync.

- **frame synchronous Viterbi beam search in CTC**

$$\mathbf{w}^* = \operatorname*{argmax}_{\mathbf{w}}\{P(\mathbf{w})p(\mathbf{x}|\mathbf{w})\} = \operatorname*{argmax}_{\mathbf{w}}\{P(\mathbf{w})p(\mathbf{x}|\mathbf{l_w})\} \quad (1)$$

$$= \operatorname*{argmax}_{\mathbf{w}}\left\{P(\mathbf{w})\max_{\mathbf{l_w}}\frac{P(\mathbf{l_w}|\mathbf{x})}{P(\mathbf{l_w})}\right\} \quad (2)$$

$$\cong \operatorname*{argmax}_{\mathbf{w}}\left\{P(\mathbf{w})\max_{\pi:\pi\in L',\mathcal{B}(\pi_{1:T})=\mathbf{l_w}}\frac{1}{P(\mathbf{l_w})}\prod_{t=1}^{T}y_{\pi_t}^{t}\right\} \quad (3)$$

Frame Sync. to Phone Sync.

- **frame synchronous Viterbi beam search in CTC**

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmax}}\{P(\mathbf{w})p(\mathbf{x}|\mathbf{w})\} = \underset{\mathbf{w}}{\operatorname{argmax}}\{P(\mathbf{w})p(\mathbf{x}|\mathbf{l_w})\} \qquad (1)$$

$$= \underset{\mathbf{w}}{\operatorname{argmax}}\left\{P(\mathbf{w})\max_{\mathbf{l_w}}\frac{P(\mathbf{l_w}|\mathbf{x})}{P(\mathbf{l_w})}\right\} \qquad (2)$$

$$\cong \underset{\mathbf{w}}{\operatorname{argmax}}\left\{P(\mathbf{w})\max_{\pi:\pi\in L',\mathcal{B}(\pi_{1:T})=\mathbf{l_w}}\frac{1}{P(\mathbf{l_w})}\boxed{\prod_{t=1}^{T}} y_{\pi_t}^{t}\right\} \qquad (3)$$

- **frame sync. to phone synchronous decoding**

$$\mathbf{w}^* \cong \underset{\mathbf{w}}{\operatorname{argmax}}\left\{P(\mathbf{w})\max_{\pi:\pi\in L',\mathcal{B}(\pi_{1:T})=\mathbf{l_w}}\frac{1}{P(\mathbf{l_w})}\left\{\prod_{t\notin U} y_{\pi_t}^{t}\cdot\prod_{t\in U} y_{\mathtt{blank}}^{t}\right\}\right\} \qquad (4)$$

$$U = \{u : y_{\mathtt{blank}}^{u} \simeq 1\} \qquad (5)$$

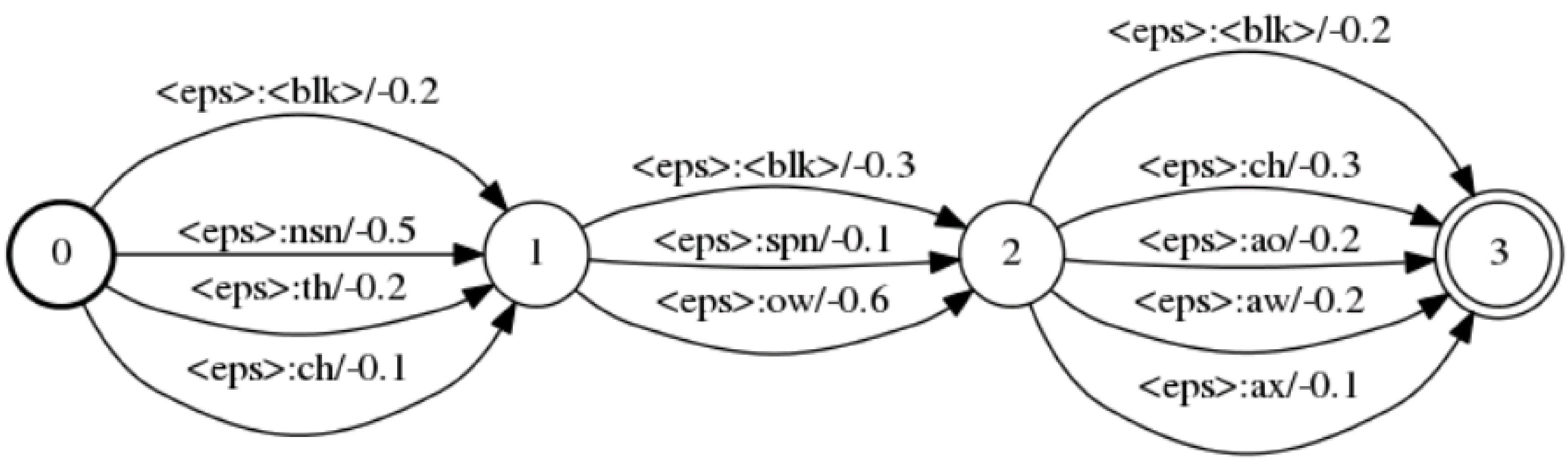
$$= \underset{\mathbf{w}}{\operatorname{argmax}}\left\{P(\mathbf{w})\max_{\pi':\pi'\in L,\mathcal{B}(\pi'_{1:J})=\mathbf{l_w}}\frac{1}{P(\mathbf{l_w})}\boxed{\prod_{j=1}^{J}} y_{\pi'_j}^{t_j}\right\} \qquad (6)$$

$$J = T - |U| \qquad (7)$$

CTC Lattice

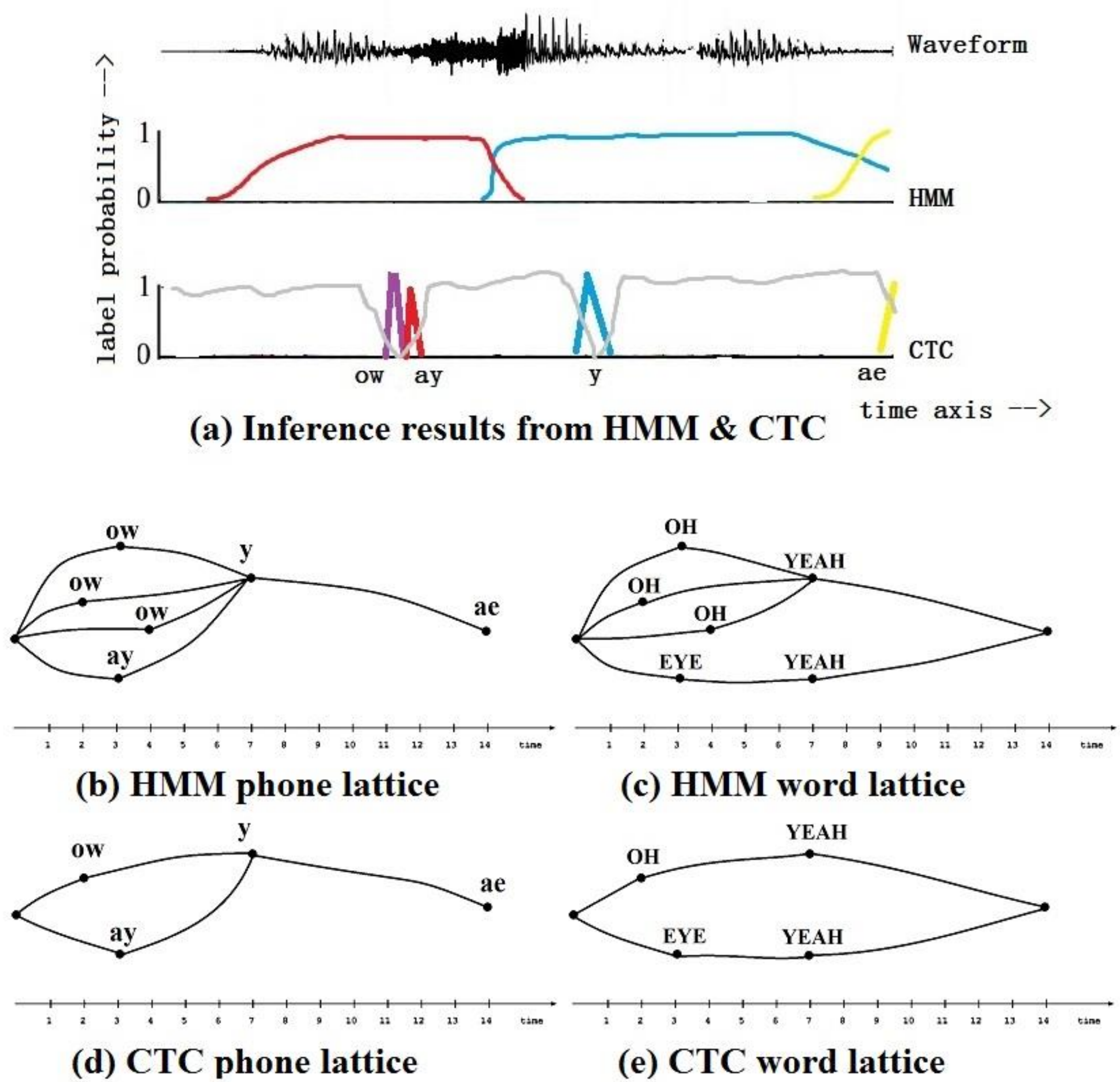
- **CTC Lattice - Extremely Compact Acoustic Information Preserver**

Time	phone label : acoustic score
0.4s	$< blk >: 0.2 \quad nsn : 0.5 \quad th : 0.2 \quad ch : 0.1$
0.9s	$< blk >: 0.3 \quad ow : 0.6 \quad spn : 0.1$
1.5s	$< blk >: 0.2 \quad ch : 0.3 \quad ao : 0.2 \quad aw : 0.2 \quad ax : 0.1$



Hypothesis Posterior CM

- Procedule:
  - Phone level CTC lattice
  - Word Lattice
  - Confusion Network
  - Confidence Measure



Experiment

- Setup
  - Swb 300h, 2-2.5M parameters, NIST hub5e-swb subset
  - details on our paper
- Baseline WER performance

Model Unit	AM	Decoding	WER
CD-state	DNN-HMM	FSD	16.7
CI-phone	LSTM-CTC	FSD	18.7
		PSD	18.8

- CM Evaluation: Normalised Cross Entropy (NCE)

$$NCE = \frac{H(\mathbf{C}) - H(\mathbf{C}|\mathbf{x})}{H(\mathbf{C})}$$

$H(\mathbf{C})$ corresponds to the entropy of the tag sequence, $H(\mathbf{C}|\mathbf{x})$ is the entropy of the confidence score sequence

  - The higher the better

Experiment

- Hypothesis Posterior CM [1]

AM	Decoding	CM	NCE
DNN-HMM	FSD	CN	0.172
LSTM-CTC	FSD	CN	0.019
	PSD	CN	$\boxed{0.224}$
		AC+CN	0.230

  - CN hypothesis posterior CM can't be directly applied to CI-phone-CTC model
    - Blank allocation problem:
      - e.g., ow <blk> ch <blk> <blk> <blk> ao <blk>

[1] We also derive a PSD version of predictor based CM, detail comparison can be referred to our paper.

Experiment

- Hypothesis Posterior CM [1]

AM	Decoding	CM	NCE
DNN-HMM	FSD	CN	0.172
LSTM-CTC	FSD	CN	0.019
	PSD	CN	0.224
		AC+CN	0.230

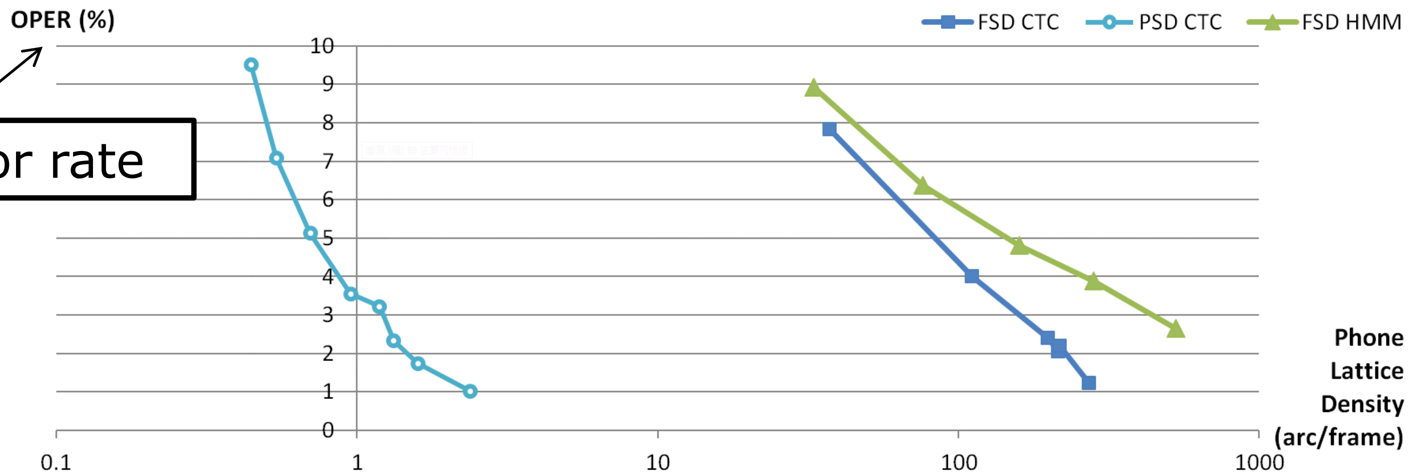
- CN hypothesis posterior CM can't be directly applied to CI-phone-CTC model
  - Blank allocation problem:
    - e.g., ow <blk> ch <blk> <blk> <blk> ao <blk>
- In PSD, CN hypothesis posterior CM can be successfully applied
- Even with significantly better NCE: 0.224 → 0.172

[1] We also derive a PSD version of predictor based CM, detail comparison can be referred to our paper.

Experiment

- Reason of Better CM
  - Better lattice

Phone



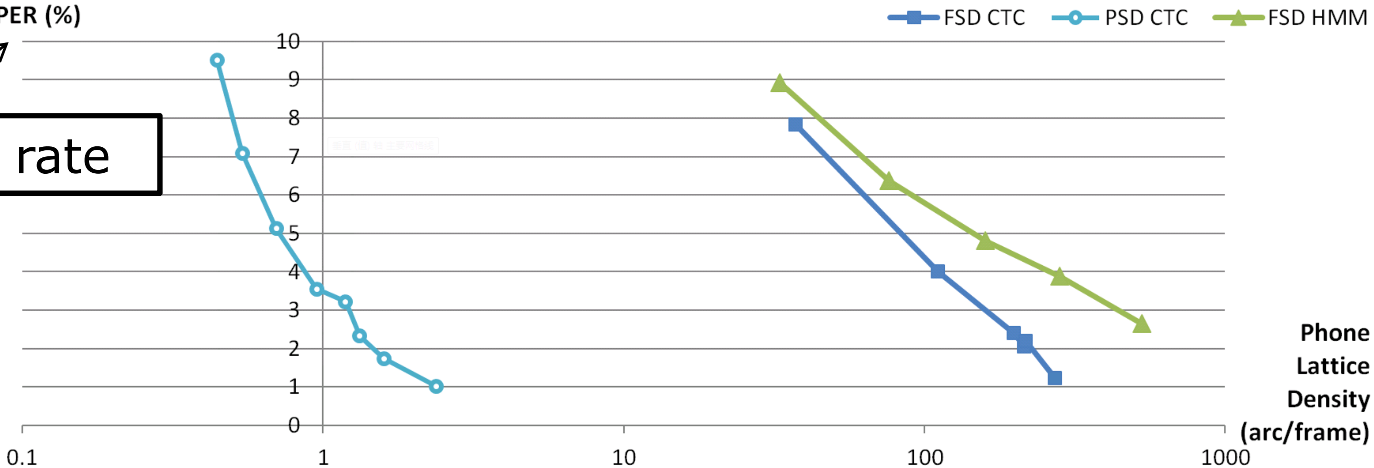
Oracle phone error rate

Experiment

- Reason of Better CM
  - Better lattice

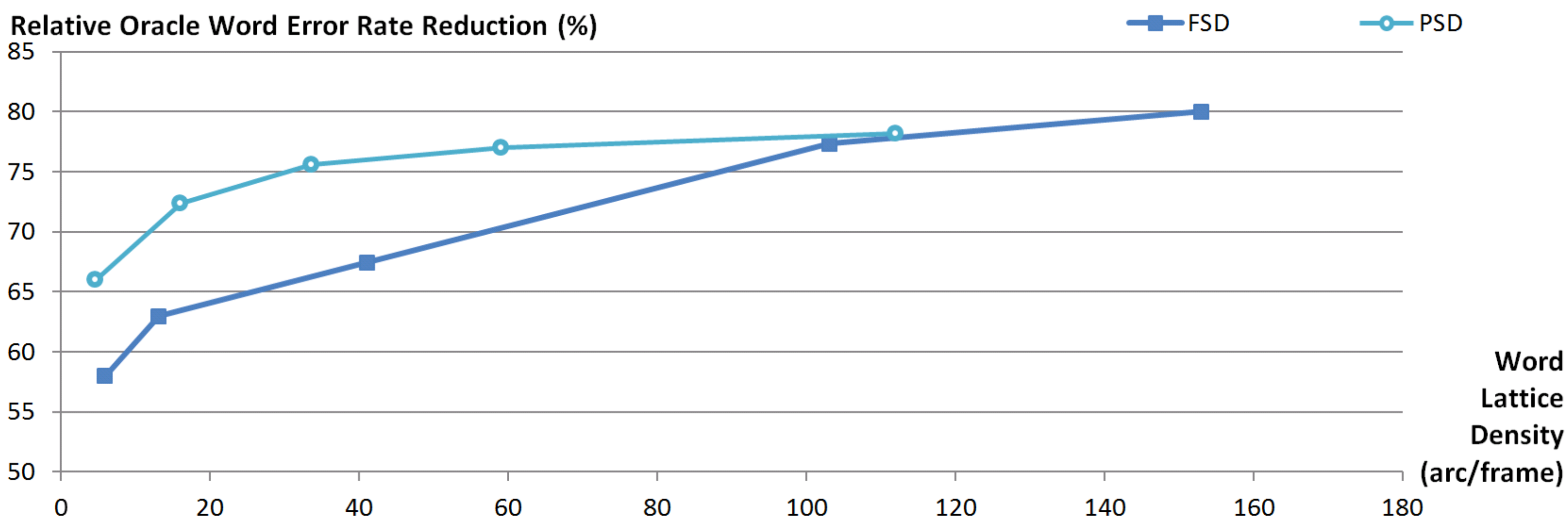
Phone

Oracle phone error rate



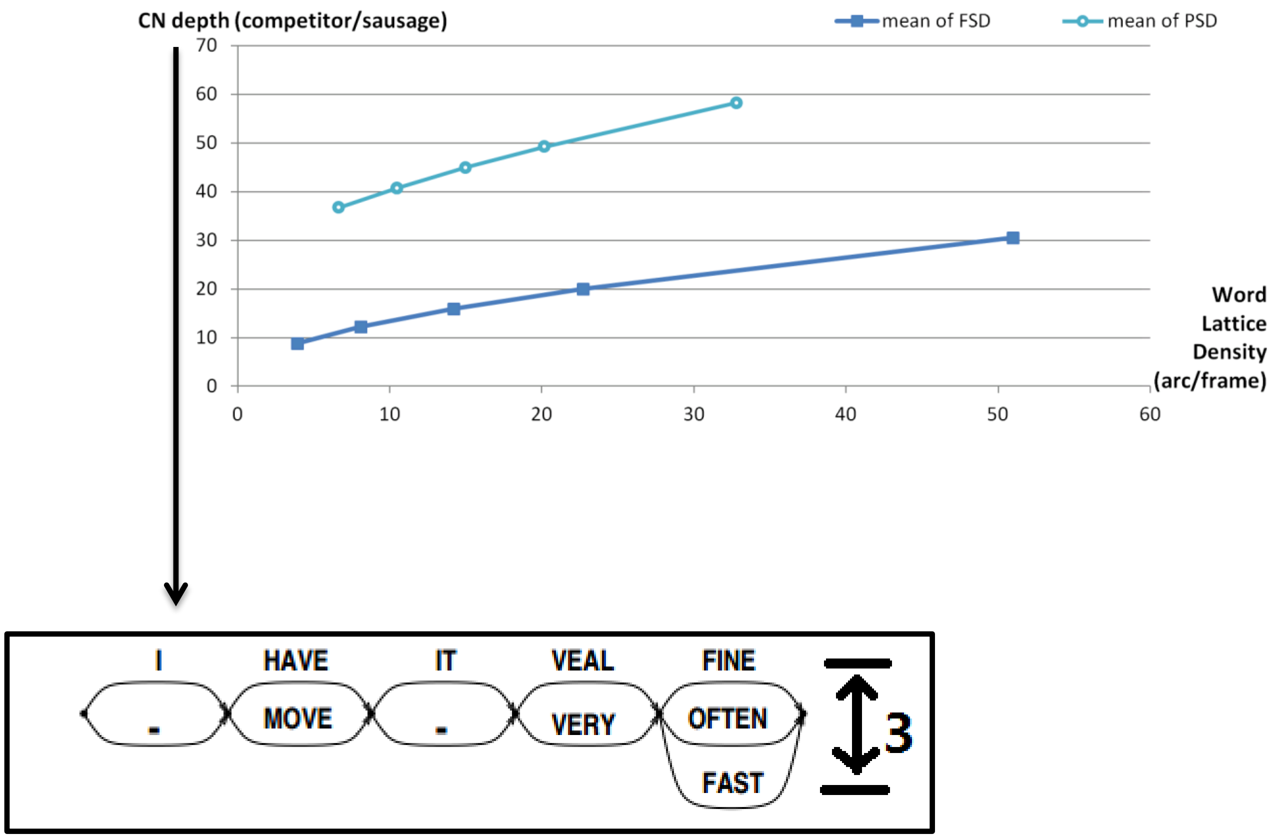
Word

$$1 - \frac{OWER}{WER}$$



Experiment

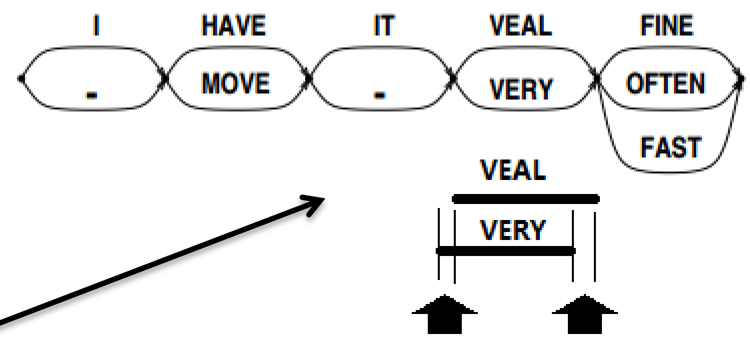
- Reason of Better CM
  - Larger CN depth → more competing information



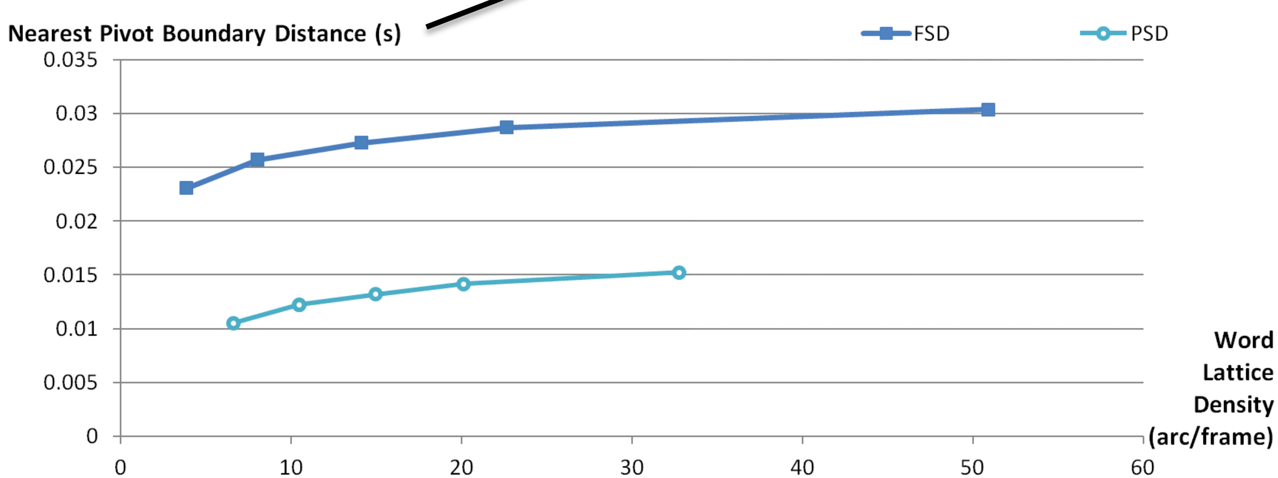
SHANGHAI JIAO TONG UNIVERSITY

Experiment

- Reason of Better CM
  - Larger CN depth → more competing information

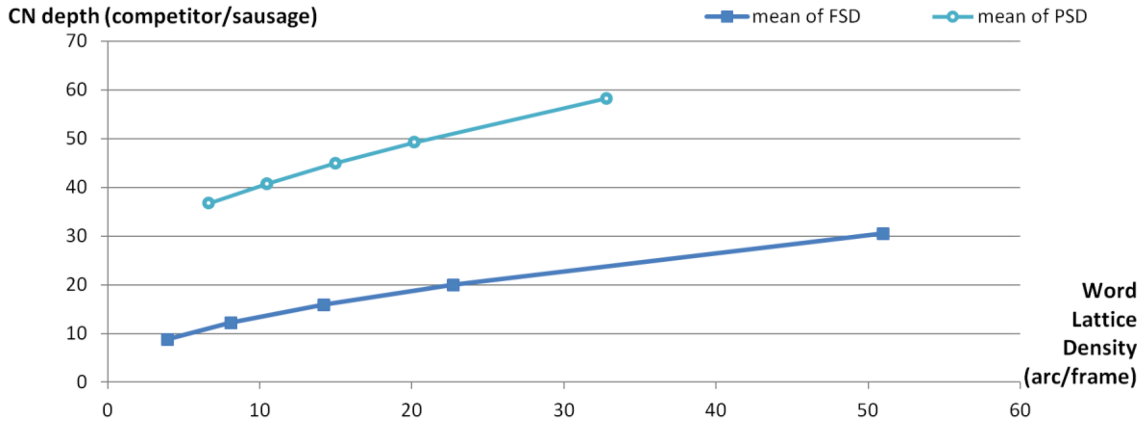


  - Because of more stable boundary

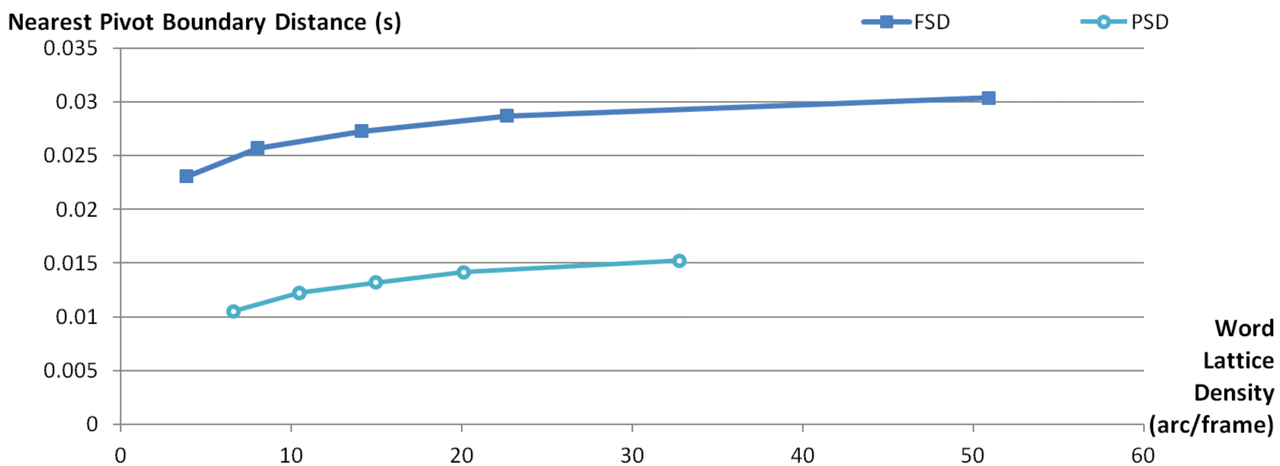


Experiment

- Reason of Better CM
  - Larger CN depth → more competing information



  - Because of more stable boundary



Summary

- The potential of **compact** and **precise** PSD CTC lattice in preserving acoustic information was utilized to form better CMs
- PSD version of predictor based CM was proposed with elaborate phonemic normalization and blank info (in paper)
- The characteristics of **lattice** and **confusion network** generated from **PSD** framework were carefully investigated, and CN hypothesis posterior CM was proposed
- The two types of CMs can be combined together as a pair of complements

- Future work: applying proposed CMs as predictors in model training framework