Phone Synchronous Decoding with CTC Lattice

SJTU Speech Lab 上海交通大學智能語音實驗室

AISPEECH 思必驰

Zehuai Chen, Wei Deng, Tao Xu, Kai Yu

Key Lab. of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering

Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

AISpeech Ltd.

Overview

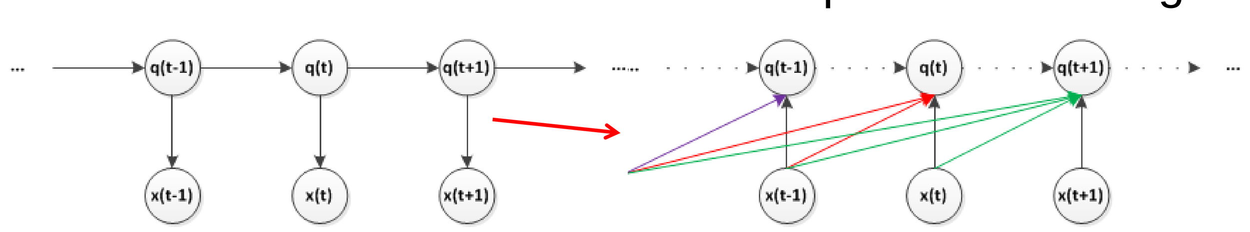
- **Motivation: CTC model shows peaky posterior property and ignoring *blank* frames will not introduce additional search errors.**
- **Approach: A novel *phone synchronous decoding* framework and compact acoustic space representation, *CTC lattice* are proposed.**
- **Experiments & Discussion: Experiments on both English and Mandarin show an extra 2-3 times speed up compared to the traditional frame synchronous CTC implementation.**

ASR Decoding & its Weakness

- **Difference in model granularity→Decoder**
- AM, LM, HMM, Lexicon…
- **Prior arts of decoding**
- Offline WFST based optimization and online viterbi search and beam prune
- *Variable frame rate* (VFR) : from equal interval search to unequal (by feature analysis)
- **Weakness**
- Huge search space
- Search errors from pruning
- Feature level VFR shows limited improvement

From HMM to CTC model

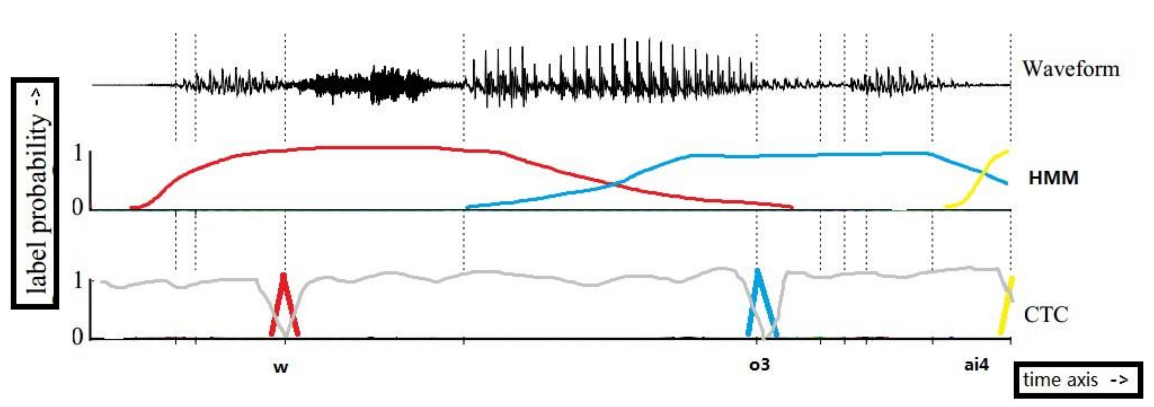
- From HMM to CTC: do better in sequential modeling



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

$$P(\mathbf{l}|\mathbf{x}) = \sum_{\pi\in\mathcal{B}^{-1}(\mathbf{l})} P(\pi|\mathbf{x}) = \sum_{\pi:\pi\in L',\mathcal{B}(\pi_{1:T})=\mathbf{l}} \prod_{t=1}^{T} y^t_{\pi_t} \qquad \begin{aligned}\mathcal{B} &: L' \mapsto L\\ L' &= L\cup\{\texttt{blank}\}\end{aligned}$$

- peaky distribution and concentrated information output



Frame Sync. To Phone Sync.

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

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

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

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

$\pi_{1:T} = (\pi_1,\ldots,\pi_T)$ is the frame-wise decoding *path*

$\mathbf{l}_\mathbf{w}$ is phone sequence corresponding to w in dictionary

$l\in L$ and L is the phone set

$\pi\in L'$ and $L' = L\cup\{\texttt{blank}\}$

- **Frame synchronous to phone synchronous decoding**

$$\mathbf{w}^* \cong \operatorname*{argmax}_{\mathbf{w}}\left\{P(\mathbf{w})\max_{\pi:\pi\in L',\mathcal{B}(\pi_{1:T})=\mathbf{l}_\mathbf{w}}\frac{1}{P(\mathbf{l}_\mathbf{w})}\left\{\prod_{t\notin U}y^t_{\pi_t}\cdot\prod_{t\in U}y^t_{\texttt{blank}}\right\}\right\}$$

$$= \operatorname*{argmax}_{\mathbf{w}}\left\{P(\mathbf{w})\max_{\pi':\pi'\in L,\mathcal{B}(\pi'_{1:J})=\mathbf{l}_\mathbf{w}}\frac{1}{P(\mathbf{l}_\mathbf{w})}\prod_{j=1}^{J}y^{t_j}_{\pi'_j}\right\}$$

$U = \{u : y^u_{\texttt{blank}} \simeq 1\}$ is the set of common *blank* time indexes

$J = T - |U|$ is the number of output phone labels

- **Different information rate**
  - Acoustic information processing: frame by frame
  - Linguistic information processing: phone by phone
- **Adjustable search interval**
  - WFST search interval is self-adjusted but not equal interval
- **Compared with VFR**
  - Frame rate analysis on model rather than feature level

- **Analysis on Search Space Compression**
- Network Traversal Reduction $\quad \lambda = \frac{1}{N}\sum_{n=1}^{N}\frac{\#\{U^{(n)}\}}{T^{(n)}}$

  λ is the average of blank frame percentages of test utterances
- Theoretical Compression Rate $\quad R = 1-(1-\lambda)\times\beta$

  β is the percentage of active phones with respect to all phones for a given set of test utterances

  R is the overall measure of the search space compression yielded by PSD

Experiments

- **Experimental Setup**
- Training stage
  - English: SWB 300h 3-gram LM from SWB without interpolation
  - Mandarin: 300h & 5000h 3-gram LM (1.7GB with 118K words)
  - Procedure similar to *EESEN* (miao et al. 2015)
  - 2-3M parameters
- Test stage
  - On *Intel(R) Xeon(R) CPU E5-2690 v2 @ 3.00GHz*.
  - *Hub5e00* testset from Switchboard and a Mandarin testset, *CellPhone*, is used, which is recorded in several speech scenarios and with about 25 hours

- **Baseline CER/WER & RTF performance**

Task	Context Dependency	Acoustic Model	CER / WER	RTF
Switchboard	CD	dnn-hmm	18.3	0.27
	CI	lstm-ctc	20.7	0.044
CellPhone	CD	dnn-hmm	13.30	0.32
	CI	lstm-ctc	10.20	0.044

- With 300 hours, CI-phone-CTC and CD-state-HMM are similar
- With 5000 hours, CI-phone-CTC outperforms CD-state-HMM
- CTC is faster than HMM by 7 times
- **Search Space Compression**
- All gotten by force-aligned CTC paths

testset	$\lambda(\%)$	$\beta(\%)$	$R(\%)$
Switchboard	88	5	99.4
CellPhone	87	11	98.6

- Phone synchronous decoding remaining 10% network traversal in WFST search
- CTC lattice remaining 1% acoustic information from acoustic posterior distribution

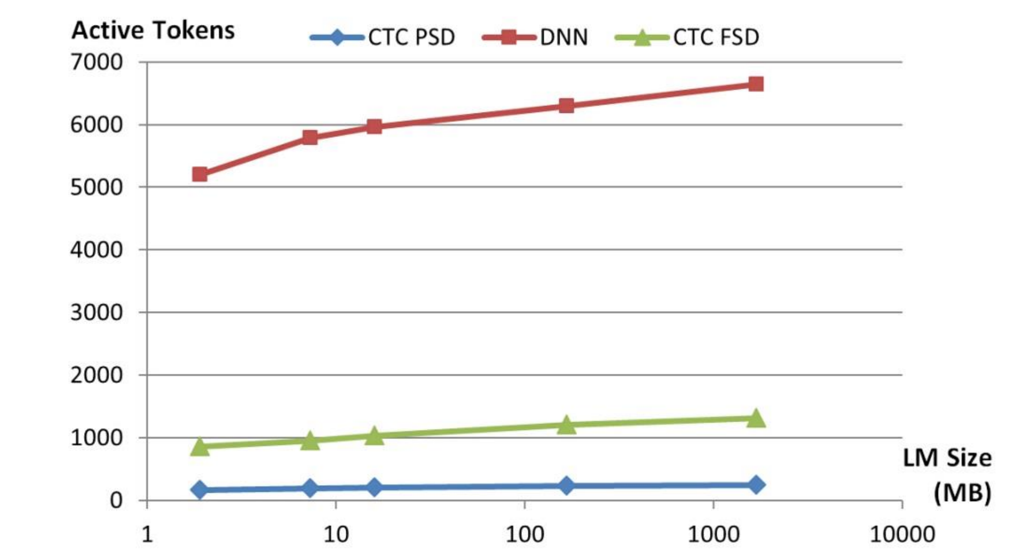
•**Decoding Speed-up**

model	search step	CER	RTF
HMM	frame	13.3	0.32
CTC	frame	10.2	0.044**(7.3X)**
	phone	10.1	0.016**(20X)**

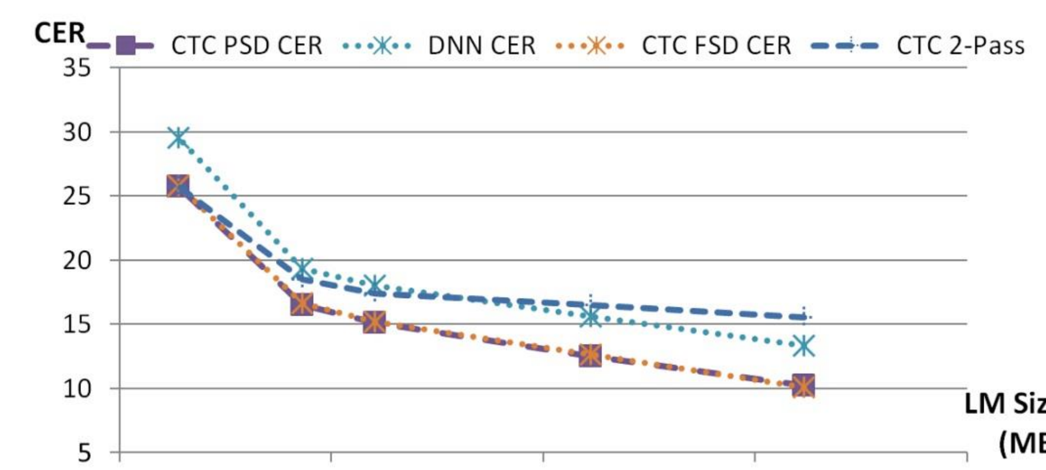
•**3X** speed-ups with no CER deterioration
(similar speedup rate in CD-phone-CTC in our recent work)

•Result of English corpus is similar and listed in our paper

- **Speed robustness**
- Extendibility of more complex linguistic search space
  - LM size ↑ → linguistic search space ↑
  - Active Tokens ↑ → RTF ↑ → Speed ↓



- Extendibility: CTC PSD > CTC FSD >> DNN FSD



- LM size ↑ → CER ↓
- CTC PSD is suitable for combining with complex linguistic search space

Conclusions

- **Frame synchronous decoding was transformed into phone synchronous decoding**
  - Self-adjusting decoding interval
  - Model level variable frame rate
  - Removing tremendous search redundancy
- **CTC lattice was proposed**
  - Extremely compact acoustic information preserver
  - Extensibility of combining with other knowledge sources