Interspeech 2018
End-to-End based ASR
Part 1-2

Lu Huang

2018-09-19

Papers for Mandarin Chinese ASR

- Alibaba: Acoustic Modeling with DFSMN-CTC and Joint CTC-CE Learning
- Bo Xu: Extending Recurrent Neural Aligner for Streaming End-to-End Speech Recognition in Mandarin
- Bo Xu: Syllable-Based Sequence-to-Sequence Speech Recognition with the Transformer in Mandarin Chinese

Acoustic Modeling with DFSMN-CTC and Joint CTC-CE Learning

Shiliang Zhang, Ming Lei

Machine Intelligence Technology, Alibaba Group

{sly.zsl, lm86501}@alibaba-inc.com

Motivations

- CTC based ASR system's latency
  - CTC: an output target is detected can be arbitrarily delayed
  - (B)LSTM: a huge amount of memory when the sequence is very long; BLSTM's latency
- How to:
  - Using DFSMN to replace (B)LSTM
  - Joint CTC-CE training to improve stability

CTC

$$\left.\begin{array}{l}\mathcal{F}(a,-,b,c,-,-)\\ \mathcal{F}(-,-,a,-,b,c)\\ \mathcal{F}(a,b,b,b,c,c)\\ \mathcal{F}(a,-,b,-,c,c)\end{array}\right\} => (a,b,c)$$

$$\mathbf{P}(\mathbf{z}|\mathbf{x}) = \sum_{\pi \in \Phi(\mathbf{z})} \mathbf{P}(\pi|\mathbf{x})$$

$$\mathcal{L}_{ctc}(\mathbf{x}) = -\log \mathbf{P}(\mathbf{z}|\mathbf{x})$$

CTC training

- CTC training is not stable
- How to:
  - by using **two output layers** with CTC and the conventional CE loss during the training
  - **initializing from a CE** loss pre-trained model.
- It is found that even with CE pre-trained networks as initialization, CTC training can sometime still fail to converge.
- CTC training with CI-Phones is more stable than CD-Phones.
  - The searching space of CD-Phones alignments is more huge than that of CI-Phones.

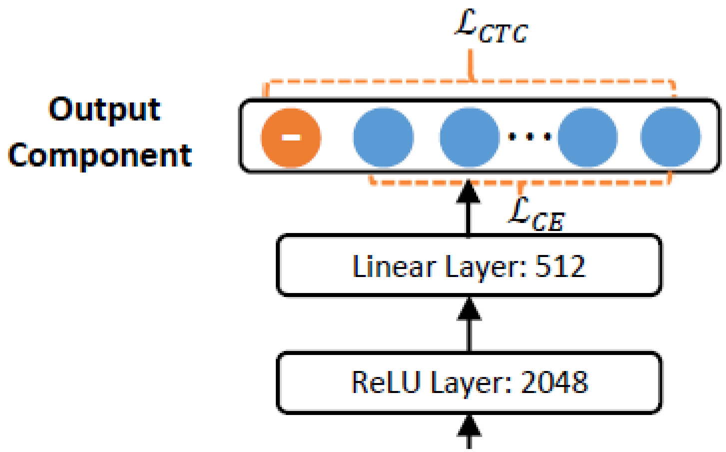
Joint CTC-CE Learning

- Difference between CTC CE:
  - loss function
  - additional CTC blank
- Joint CTC-CE
  - a single softmax output layer

$$\mathcal{L}_{ctcce}(\mathbf{x}) = \mathcal{L}_{ctc}(\mathbf{x}) + \alpha \cdot \mathcal{L}_{ce}(\mathbf{x})$$

$$\mathcal{L}_{ce}(\mathbf{x}) = -\sum_{i=2}^{K} (1 - p(y_1|\mathbf{x})) t_i \log p(y_i|\mathbf{x})$$

$\mathbf{T} = \{t_2, t_3, \cdots, t_K\}$ denotes the frame-level target labels.



- Need frame-level alignment
  - Still End-to-End?

Experiments

- Data: 1k, 4k, 20k hours
  - a normal test set and a fast speed test set
- Feature: 80-dim FBK
  - stack the consecutive frames(± 5)
  - Subsample with 3

Results

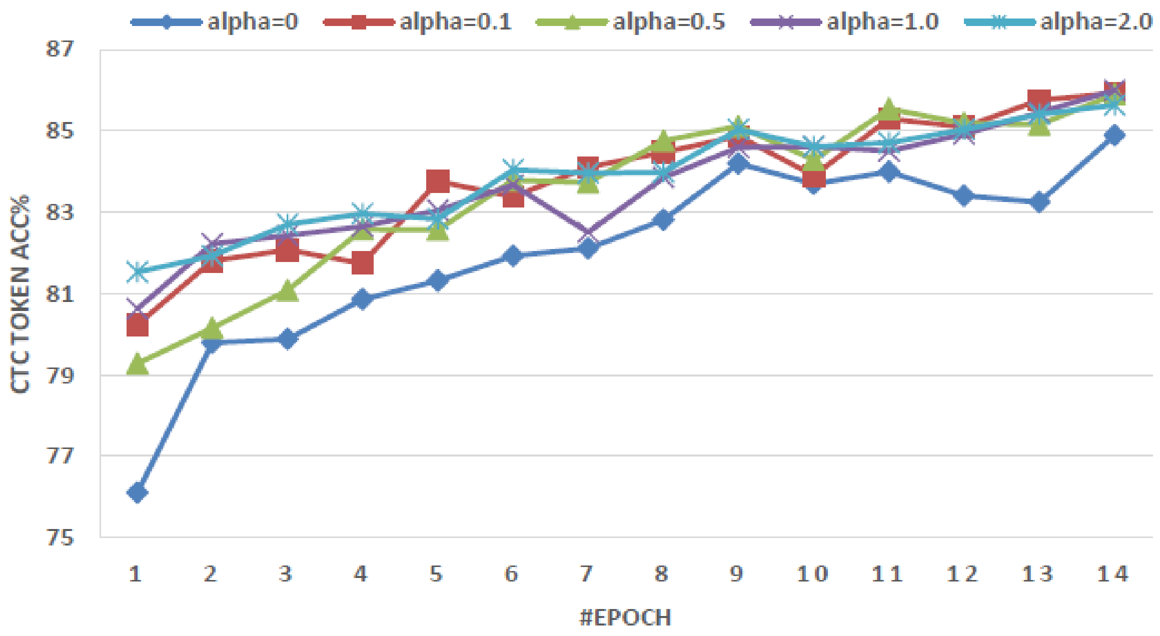
- Baseline

Method	Label	Model Size (MB)	Time/Epoch (Hours)
BLSTM-CE	CD-Phone	155	3.67
DFSMN-CE	CD-Phone	114	0.50
DFSMN-CTC	CD-Phone	114	0.58
DFSMN-CTC	CI-Phone	97	0.43

Method	Label	Data (Hours)	Test set (WER %)	
			Normal	Fast
BLSTM-CE	CD-Phone	1k	19.77	47.56
		4k	16.53	37.17
		20k	13.97	31.71
DFSMN-CE	CD-Phone	1k	18.19	44.25
		4k	14.24	33.92
		20k	12.10	29.79
DFSMN-CTC	CI-Phone	1k	17.82	43.22
		4k	13.82	32.15
		20k	11.46	26.84
DFSMN-CTC	CD-Phone	1k	16.95	40.27
		4k	13.13	26.70
		20k	11.71	24.04

Results

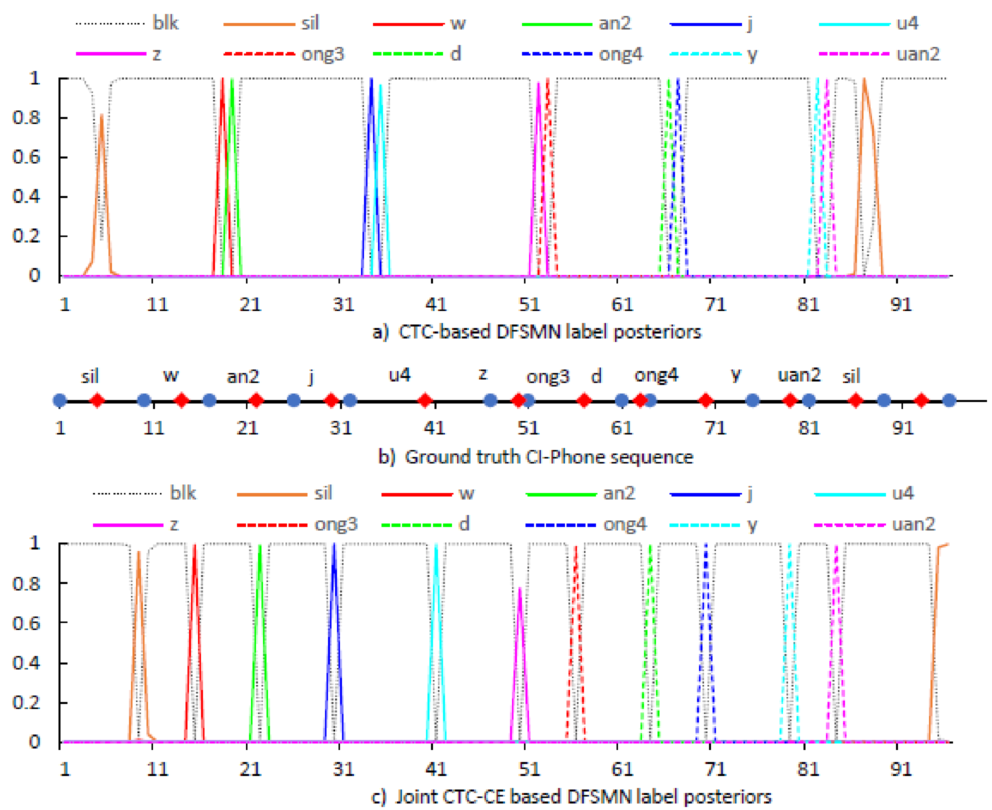
- Joint CTC-CE
  - CD-Phone



Method	Alpha	Test set (WER %)			
		Normal	Gain	Fast	Gain
CE	-	12.10	-	29.79	-
CTC	-	11.71	3.2%	24.04	19.3%
Joint CTC CE	0.1	10.92	9.8%	21.68	27.2%
	0.5	10.67	11.8%	21.98	26.2%
	1.0	10.77	11.0%	20.80	30.1%
	2.0	11.03	8.8%	22.86	23.3%

Results

- Joint CTC-CE
  - accurate alignment



Extending Recurrent Neural Aligner for Streaming End-to-End Speech Recognition in Mandarin

Linhao Dong[1,2], *Shiyu Zhou*[1,2], *Wei Chen*[1], *Bo Xu*[1]

[1]Institute of Automation, Chinese Academy of Sciences, China
[2]University of Chinese Academy of Sciences, China

Motivations

- English->Chinese
- Recurrent Neural Aligner(RNA)
  - streaming recognition
- Improve by:
  - redesign the temporal down-sampling and introduce a powerful convolutional structure.
  - In the decoder, we utilize a regularizer to smooth the output distribution and conduct joint training with a language model.

RNA

- e_u is the and encoded vector of z_u
- Diff with CTC in
  - the conditional distribution

    $$p(\mathbf{z}|\mathbf{x}) \;=\; \prod_u p(z_u|z_1^{u-1}, \mathbf{x})$$

    $$p(\mathbf{z}|\mathbf{x}) \;=\; \prod_u p(z_u|\mathbf{x})$$

  - RNA obtains the predicted output sequence by simply removing the blanks from alignment, while the CTC model needs to remove first the repeated labels and then the blanks

$$\mathbf{h} = \mathrm{encoder}(\mathbf{x})$$

$$z_u = \arg\max_{l \in [1, L+1]}(\mathrm{decoder}(h_u, e_{u-1}))$$

$$p(\mathbf{y}|\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{x})$$

Temporal down-sampling

- Pooling between LSTMs
- Strided convolutional layers

Multiplicative Units

$$\begin{aligned}
\mathbf{g_1} &= \sigma(\mathbf{W_1} * \mathbf{I} + \mathbf{b_1}) \\
\mathbf{g_2} &= \sigma(\mathbf{W_2} * \mathbf{I} + \mathbf{b_2}) \\
\mathbf{g_3} &= \sigma(\mathbf{W_3} * \mathbf{I} + \mathbf{b_3}) \\
\mathbf{u} &= \tanh(\mathbf{W_4} * \mathbf{I} + \mathbf{b_4}) \\
\mathrm{MU}(\mathbf{h}; \mathbf{W}) &= \mathbf{g_1} \odot \tanh(\mathbf{g_2} \odot \mathbf{h} + \mathbf{g_3} \odot \mathbf{u} + \mathbf{b_5})
\end{aligned}$$

Confidence Penalty

- Label Smoothing
- Obtain better generalization

$$H(p_\theta(\mathbf{z}|\mathbf{x})) = -\sum_{u\in[1,U]}\sum_{z_u\in[1,L+1]} p_\theta(z_u|\mathbf{x})\log(p_\theta(z_u|\mathbf{x}))$$

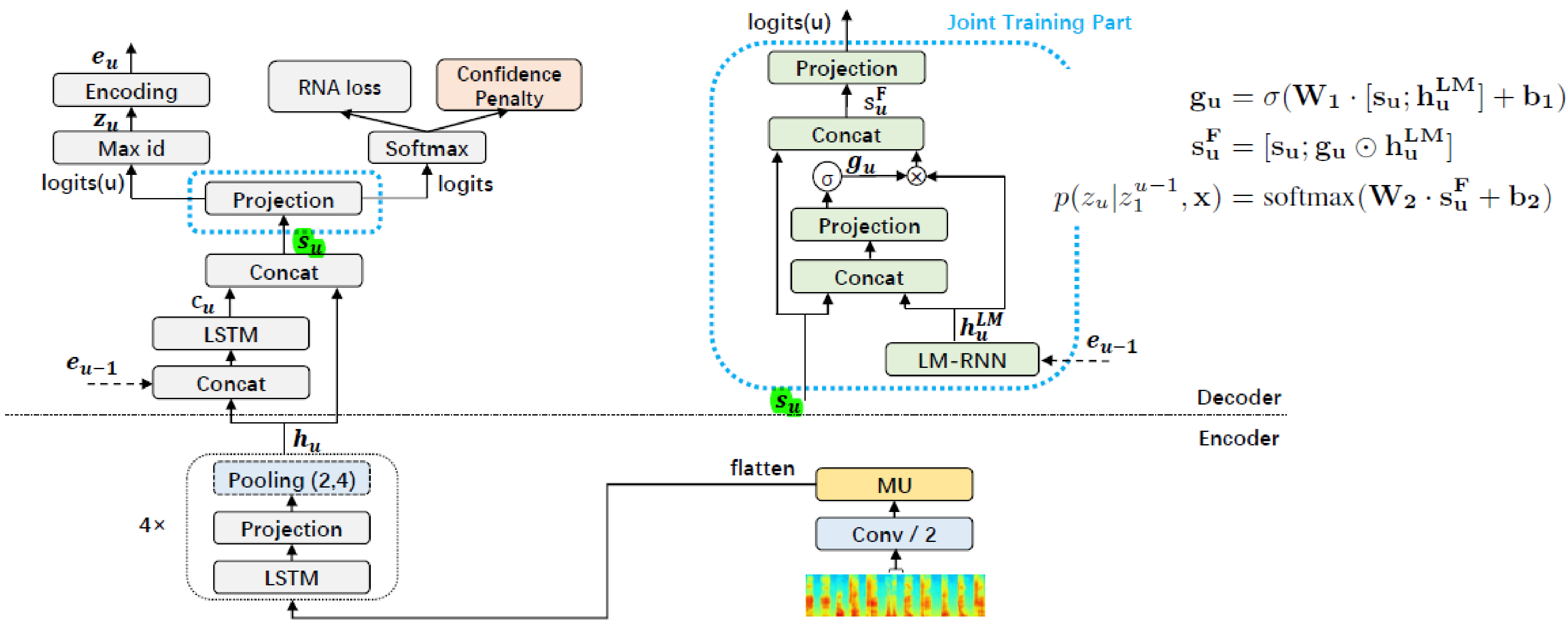
$$L(\theta) = \sum_{(\mathbf{x},\mathbf{y})} -\log(p_\theta(\mathbf{y}|\mathbf{x})) - \lambda\sum_{\mathbf{x}} H(p_\theta(\mathbf{z}|\mathbf{x}))$$

Joint training with RNN-LM

- Difficult:
  - If we use the shallow fusion in, it’s hard to obtain accurate alignments containing blank for training the LM.
  - If we use the mechanism of joint training with RNN-LM, the blank label hampers the synchronism between the outputs of RNA and the RNN-LM

Joint training with RNN-LM

- Let h^{LM}_u represents the LM state
  - uses the current output of LM-RNN if z_{u−1} is non-blank
  - uses the previous output of LM-RNN if z_{u−1} is blank



Exp

- HKUST

Exp

- Temporal down-sampling

Down-sampling mechanism	Rate	CER
frame stacking and sub-sampling [5]	1/3	43.19
pooling{2,4}-width{2,2}	1/4	39.80
pooling{2,4}-width{3,2}	1/6	34.07
pooling{1,2,4}-width{2,2,2}	1/8	**31.94**
pooling{1,2,4}-width{3,2,2}	1/12	33.53
pooling{1,2,3,4}-width{2,2,2,2}	1/16	36.63
conv-stride{2,2,2}	1/8	34.78
conv-stride{2,2} + pooling{2}-width{2}	1/8	32.62
conv-stride{2} + pooling{2,4}-width{2,2}	1/8	**30.86**

Exp

- further extensions on RNA

Model-ID	Model	CER
$M1$	RNA with the best down-sampling	30.86
$M2$	$M1$ + 1 * MU	29.89
-	$M1$ + 1 * ConvLSTM	30.55
-	$M1$ + 1 * GLU	30.36
$M3$	$M2$ + Confidence Penalty (λ = 0.2)	29.06
$M4$	$M3$ + Joint training with RNN-LM	28.32

Syllable-Based Sequence-to-Sequence Speech Recognition with the Transformer in Mandarin Chinese

Shiyu Zhou[1,2], *Linhao Dong*[1,2], *Shuang Xu*[1], *Bo Xu*[1]

[1]Institute of Automation, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences

Motivation

- Transformer achieves a state-of-the-art BLEU on NMT
- Extend it to speech as the basic architecture of sequence-to-sequence attention-based model on Mandarin Chinese ASR
- Investigate a comparison between syllable based model and context-independent phone based model
    - syllables have the advantage of avoiding OOV problem
- A greedy cascading decoder with the Transformer is proposed for mapping CI-phoneme sequences and syllable sequences into word sequences

Transformer model

- the same as sequence-to-sequence attention-based models except relying entirely on self-attention and position-wise
  - Encoder
  - Decoder
  - Multi-head attention

$$Attention(Q, K, V) = softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V$$

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O$$
$$\text{where } head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

Transformer model

- MHA and position-wise, fully connected layers for both the encode and decoder
- positional encodings

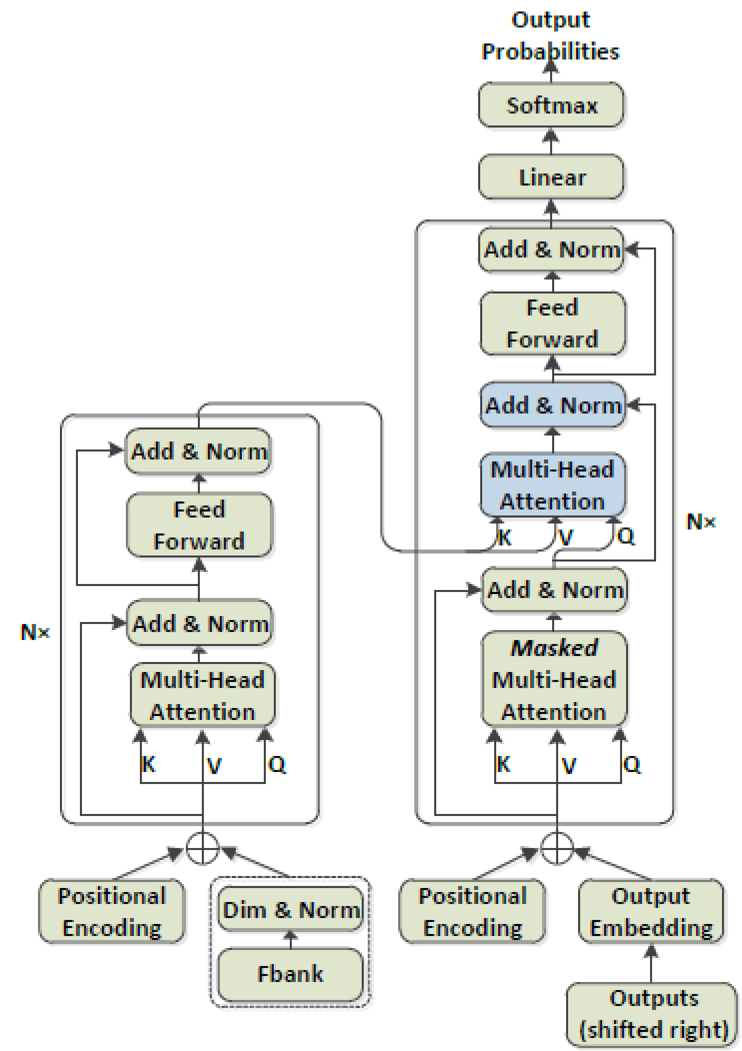


Figure 1: *The architecture of the ASR Transformer.*

Greedy cascading decoder

- First, the best sub-word unit sequence s is calculated by the Transformer from observation X to sub-word unit sequence with beam size β.
- Then, the best word sequence W is chosen by the Transformer from sub-word unit sequence to word sequence with beam size γ.

$$
\begin{aligned}
\tilde{W} &= \underset{W}{\operatorname{argmax}}\, Pr(W|X) \\
&= \underset{W}{\operatorname{argmax}} \sum_{s} Pr(W|s) Pr(s|X) \\
&\approx \underset{W}{\operatorname{argmax}}\, Pr(W|s) Pr(s|X)
\end{aligned}
$$

Exp

- HKUST
- CI-phoneme: 122
- syllable: 1388

Exp

- CI-phoneme and syllable based model

Table 2: *Comparison of CI-phoneme and syllable based model with the Transformer on HKUST datasets in CER (%).*

sub-word unit	model	CER
CI-phonemes	D512-H8	32.94
	D1024-H16	**30.65**
	D1024-H16 (speed perturb)	30.72
Syllables	D512-H8	31.80
	D1024-H16	29.87
	D1024-H16 (speed perturb)	**28.77**

Exp

- Comparison with previous works

Table 3: *CER (%) on HKUST datasets compared to previous works.*

model	CER
LSTMP-9×800P512-F444 [24]	30.79
CTC-attention+joint dec. (speed perturb., one-pass)	
+VGG net	28.9
+RNN-LM (separate) [9]	**28.0**
CI-phonemes-D1024-H16	30.65
Syllables-D1024-H16 (speed perturb)	**28.77**

Exp

- Comparison of different frame rates

Table 4: *Comparison of different frame rates on HKUST datasets in CER (%).*

model	frame rate	CER
CI-phonemes-D1024-H16 (speed perturb)	$30ms$	**30.72**
	$50ms$	31.68
	$70ms$	33.96
Syllables-D1024-H16 (speed perturb)	$30ms$	**28.77**
	$50ms$	29.36
	$70ms$	32.22