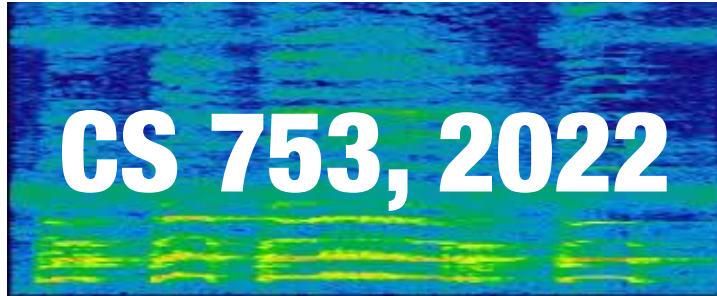


Pronunciation Models

Lecture 9c



CS 753, 2022

Instructor: Preethi Jyothi, IITB

Role-Playing Seminars

ROLE	Description	Evaluation/Due	Score
Talk Presentation	15 min talk + Q/A to clearly explain the main ideas in the paper	According to talk schedule	8
Scientific Reporter	Write a technical article about the paper for a non-specialist but general CS audience	Due towards the end of the semester	5
Peer Reviewer	Provide a full review of the paper	Reviews will be due later	4
Poster Presentation	Prepare a poster visualising the main highlights of the paper	GatherTown poster session at the end of the semester	5
Hacker	Implement a small part/simplified version of the paper	Code/demo submission after the final exams	8

30

Seminar Talk Schedule

DATE	Team 1 (Paper)	Team 2 (Paper)	Team 3 (Paper)	Team 4 (Paper)
Mar 8	CryptoFox (EMFORMER)	ASR (Blocksync Beam)	OK ASR! (Stream Trans ASR)	
Mar 11	ML_Monks (VoiceConv)	Audio Rhythm (SimulSpeech)	Tenacious Tomatoes (MetricGAN)	
Mar 12	Team Rocket (AutoVC)	waV1Vec (AV Dereverb)	The 3 Musketeers (Relthink ASR)	angry_nerds (Mask-CTC)
Mar 15	Unmuters (AudioAdversary)	Jediknights (Quaternion)	HMMmmmmm (VoiceSep)	
Mar 22	Binary Brains (Rhythm Transfer)	Vadati (BERT ASR)	Overload.ai (ASR Robustness)	
Mar 25	Pitch Perfect (wav2vec U)	Thunderstruck (Right2Talk)	Horn Please (NeuralSynth)	
Mar 26	235161 (SpeakQL)	VMV (SoundSep Video)	A-Team (Diarization)	DeepSpeech (Visual Voice)
Mar 29	ASRians (Stream Trans ASR 2)	Speech Hackers (Large Margin SV)	phoeniTIX (Efficient Vocalize)	
Apr 1	3 Idiots (Multiling ASR)	Runtime Terror (StyleTokens)	Glove_n_Caffe (ImpercepAdvASR)	
Apr 2	let it be X (Look2Speech)	Glorious Purpose (Clued AVVP)	Code black (GE2E)	Voice.AI (TriBERT)
Apr 5	J.A.R.V.I.S. (APS, Speech2SQL)	Audio-Panda (LatticeInputs)	ComplexASR (SpokenTranslation)	
Apr 8	Dhyana (Move2Hear)	Hustlers (RNNT Decoding)	Tri-Vachan (AVID)	
Apr 9	J.A.R.V.I.S (GS, Word-level E2E ASR)	Data Wave (Deepfakes)	Glitch (Imputer)	

Hacker (Paper Assignment)

231561 (Imputer)	3 Idiots (PARP)	A Team (BERT ASR)
angry_nerds (Word-level E2E ASR)	ASR (TriBERT)	ASRians (EMFORMER)
Audio Rhythm (Prosody Repr)	Audio-Panda (Quaternion)	Binary Brains (CrossAtt Trans)
Code black (CIF ASR)	ComplexASR (SimulSpeech)	CryptoFox (Rhythm Transfer)
Data Wave (Blocksync Beam)	DeepSpeech (MetricGAN)	Dhyana (Large Margin SV)
Glorious Purpose (Diarization)	Glove_n_Caffe (AudioAdversary)	HMMmmmmm (wav2vec U)
Horn Please (Move2Hear)	Hustlers (Deepfakes)	J.A.R.V.I.S. (APS, Clued AVVP)
J.A.R.V.I.S. (GS, SoundSep Video)	Jediknights (AVVP)	let it be X (Right2Talk)
ML_Monks (Mask-CTC)	OK ASR! (Disent Repr)	Overload.ai (VoiceConv)
phOeniTiX (NeuralSynth)	Pitch perfect (Stream Trans ASR 2)	Runtime Terror (SpeakQL)
Speech Hackers (LatticeInputs)	Team Rocket (StyleTokens)	Tenacious Tomatoes (VoiceSep)
The 3 Musketeers (ASR Robustness)	Thunderstruck (RNNT decoding)	Tri-Vachan (AutoVC)
Unmuters (Multiscale ASR)	Vadati (Stream Trans ASR)	VMV (VisualVoice)
Voice.AI (GE2E)	wav1Vec (Speech2SQL)	Glitch (AVID)

All Role Assignments

Team	Seminar	Hacker	Reviewer	Poster	Scientific Report
CryptoFox	Emformer	Rhythm Transfer	AutoVC	Diarization	Deepfakes
ASR	Blocksync Beam	TriBERT	PARP	Disent Repr	AudioAdversary
OK ASR!	Stream Trans ASR	Disent Repr	Large Margin SV	CrossAtt Trans	Move2Hear
ML_Monks	VoiceConv	Mask-CTC	LatticeInputs	Prosody Repr	SpeakQL
Audio Rhythm	SimulSpeech	Prosody Repr	Quaternion	Right2Talk	GE2E
Tenacious Tomatoes	MetricGAN	Stream Trans ASR 2	Speech2SQL	Large Margin SV	VoiceSep
Team Rocket	AutoVC	StyleTokens	MetricGAN	VoiceSep	RelThink ASR
waV1Vec	AV Dereverb	Speech2SQL	VisualVoice	Word-level E2E ASR	Blocksync Beam
The 3 Musketeers	Relthink ASR	ASR Robustness	Multiscale ASR	EMFORMER	Quaternion
angry_nerds	Mask-CTC	Word-level E2E ASR	Stream Trans ASR 2	AVVP	MetricGAN
Unmuters	AudioAdversary	Multiscale ASR	EMFORMER	CIF ASR	AV Dereverb
Jediknights	Quaternion	AVVP	Prosody Repr	SpeakQL	RNNT Decoding
HMMmmmmm	SpeakQL	wav2vec U	VoiceConv	StyleTokens	Efficient Vocalize
Binary Brains	Rhythm Transfer	CrossAtt Trans	APS, Clued AVVP	MetricGAN	Speech2SQL
Vadati	BERT ASR	Stream Trans ASR	CrossAtt Trans	Rhythm Transfer	SoundSep Video
Overload.ai	ASR Robustness	VoiceConv	CIF ASR	Deepfakes	BERT ASR
Pitch Perfect	wav2vec U	VoiceSep	AVVP	Stream Trans ASR 2	Mask-CTC
Thunderstruck	Right2Talk	RNNT decoding	Deepfakes	PARP	AutoVC
Horn Please	NeuralSynth	Move2Hear	Right2Talk	Quaternion	Multiling ASR
235161	VoiceSep	Imputer	SpeakQL	BERT ASR	Right2Talk
VMV	SoundSep Video	VisualVoice	RNNT decoding	AVID	SpokenTranslation
A-Team	Diarization	BERT ASR	TriBERT	AudioAdversary	LatticeInputs
DeepSpeech	Visual Voice	MetricGAN	Diarization	GS, SoundSep Video	Stream Trans ASR
ASRians	Stream Trans ASR 2	EMFORMER	Blocksync Beam	VisualVoice	VoiceConv
Speech Hackers	Large Margin SV	LatticeInputs	Rhythm Transfer	Imputer	Look2Speech

Pronunciation Dictionary or Lexicon

- Pronunciation model/dictionary/lexicon: Lists one or more pronunciations for a word
- Typically derived from language experts: Sequence of phones written down for each word
- Dictionary construction involves:
 1. Selecting what words to include in the dictionary
 2. Pronunciation of each word (also, check for multiple pronunciations)

Pronunciations

- Same word can have multiple pronunciations
 - Accent: E.g., *route* /R UW T/ vs. /R AW T/
 - Part-of-speech: E.g., *conduct* (*verb*) /C UH N D AH K T/ vs. *conduct* (*noun*) /C AA N D AH K T/
 - Conversational effects: E.g., *probably* /P R AA B L IY/
- Most dictionaries only have words with a single pronunciation

Out-of-vocabulary Problem

- Encountering new (or *out-of-vocabulary*, *OOV*) words during test time that never appeared during training
- OOV Rate: Percentage of tokens in test data that do not appear in the word vocabulary of the system
- High OOV rates are not desirable
- More challenging for morphologically rich languages

Grapheme to phoneme (G2P) conversion

- Produce a pronunciation (phoneme sequence) given a written word (grapheme sequence)
- Learn G2P mappings from a pronunciation dictionary
- Useful for:
 - ASR systems in languages with no pre-built lexicons
 - Speech synthesis systems
 - Deriving pronunciations for out-of-vocabulary (OOV) words

G2P conversion (I)

- One popular paradigm: Joint sequence models [BN12]
- Grapheme and phoneme sequences are first aligned using EM-based algorithm
- Results in a sequence of graphones (joint G-P tokens)
- Ngram models trained on these graphone sequences
- WFST-based implementation of such a joint graphone model [Phonetisaurus]

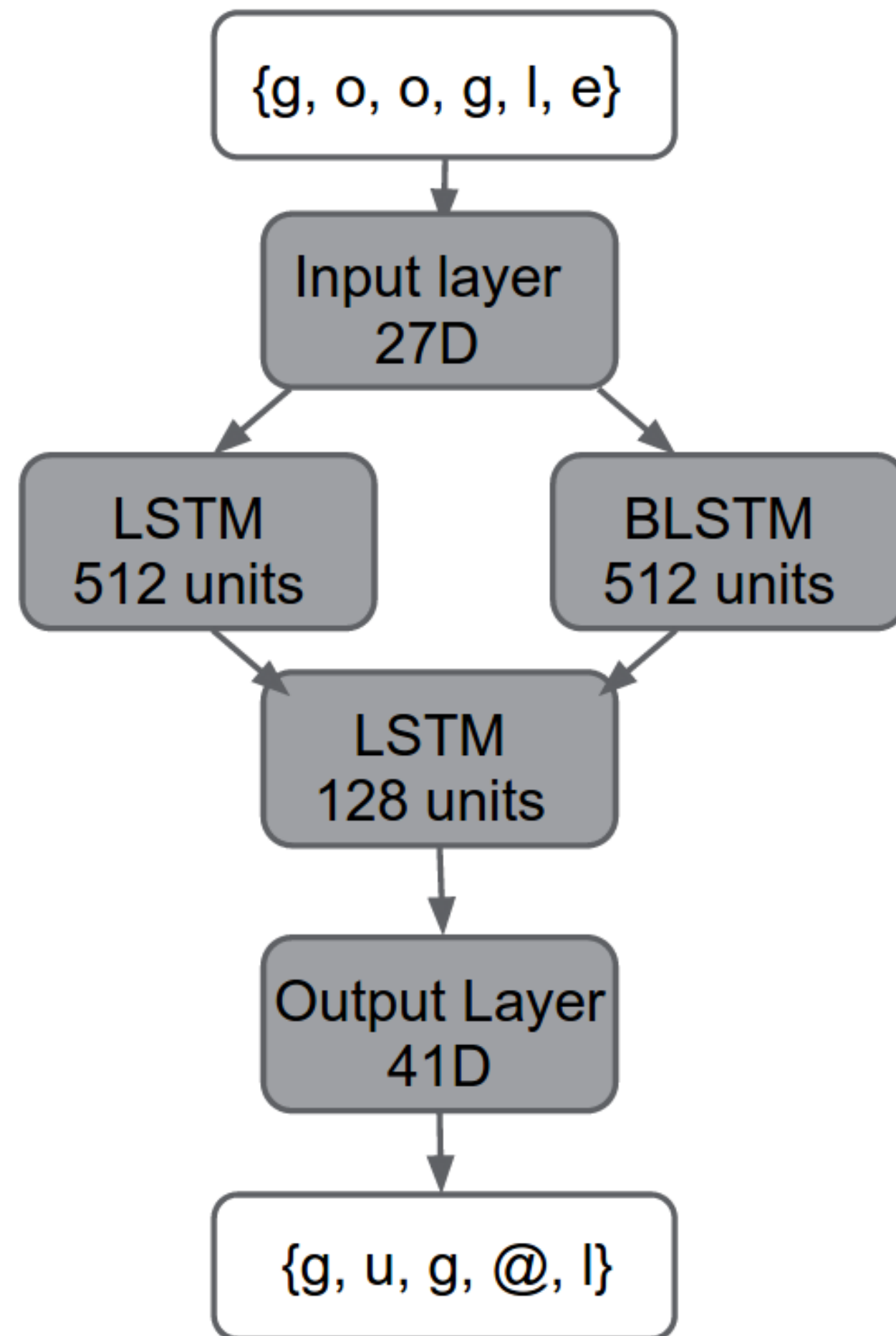
[BN12]:Bisani & Ney , “Joint sequence models for grapheme-to-phoneme conversion”,Speccom 2012

[Phonetisaurus] J. Novak, Phonetisaurus Toolkit

G2P conversion (II)

- Neural network based methods are the new state-of-the-art for G2P
- Bidirectional LSTM-based networks using a CTC output layer. Comparable to Ngram models.
- Incorporate alignment information [Yao15]. Beats Ngram models.
- No alignment. Encoder-decoder with attention. Beats the above systems [Toshniwal16].

LSTM + CTC for G2P conversion [Rao15]



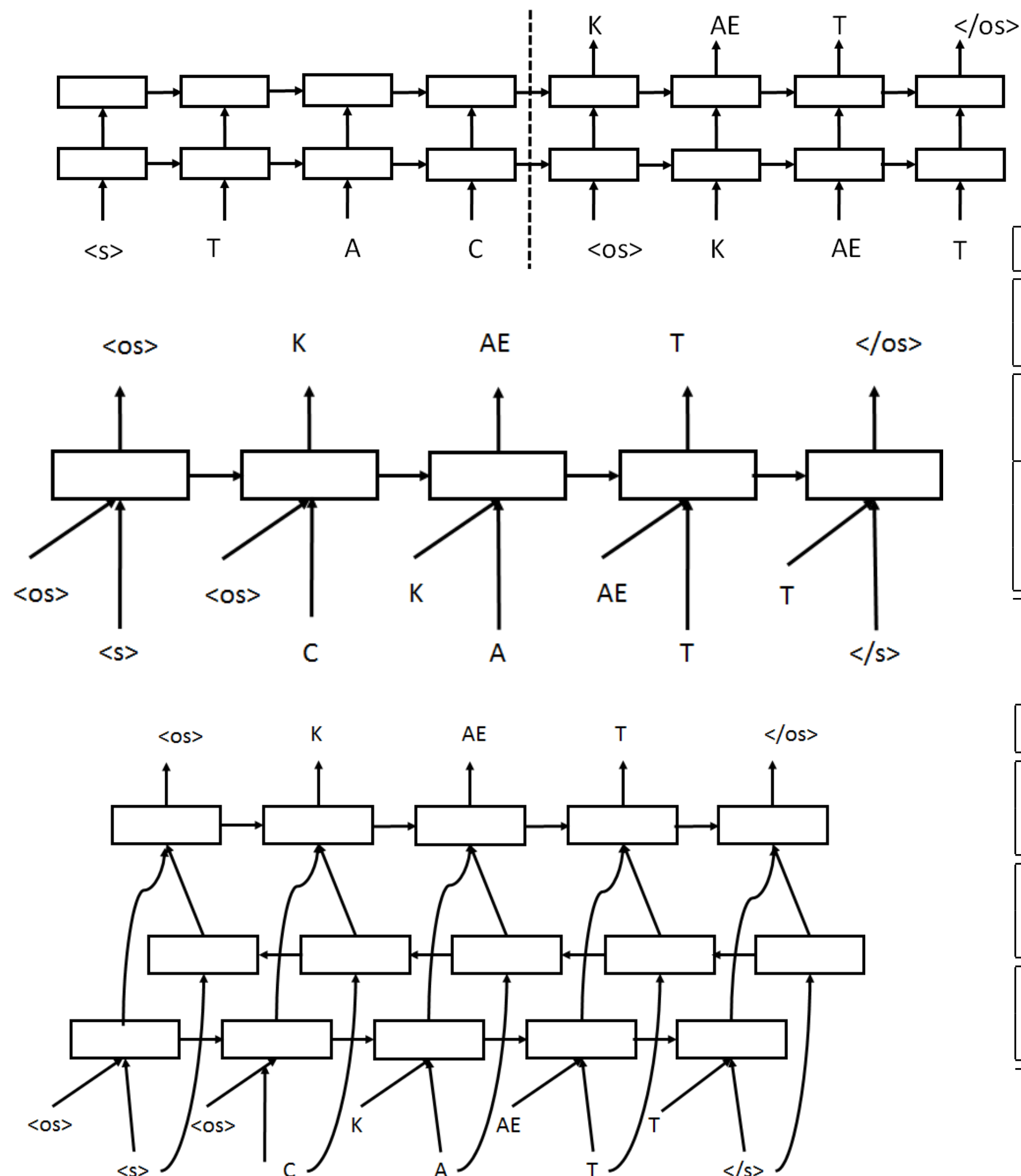
Model	Word Error Rate (%)
Galescu and Allen [4]	28.5
Chen [7]	24.7
Bisani and Ney [2]	24.5
Novak et al. [6]	24.4
Wu et al. [12]	23.4
5-gram FST	27.2
8-gram FST	26.5
Unidirectional LSTM with Full-delay	30.1
DBLSTM-CTC 128 Units	27.9
DBLSTM-CTC 512 Units	25.8
DBLSTM-CTC 512 + 5-gram FST	21.3

G2P conversion (II)

- Neural network based methods are the new state-of-the-art for G2P
 - Bidirectional LSTM-based networks using a CTC output layer [Rao15]. Comparable to Ngram models.
 - Incorporate alignment information [Yao15]. Beats Ngram models.
 - No alignment. Encoder-decoder with attention. Beats the above systems [Toshniwal16].

Seq2seq models

(with alignment information [Yao15])



Method	PER (%)	WER (%)
encoder-decoder LSTM	7.53	29.21
encoder-decoder LSTM (2 layers)	7.63	28.61
uni-directional LSTM	8.22	32.64
uni-directional LSTM (window size 6)	6.58	28.56
bi-directional LSTM	5.98	25.72
bi-directional LSTM (2 layers)	5.84	25.02
bi-directional LSTM (3 layers)	5.45	23.55

Data	Method	PER (%)	WER (%)
CMUDict	past results [20]	5.88	24.53
	bi-directional LSTM	5.45	23.55
NetTalk	past results [20]	8.26	33.67
	bi-directional LSTM	7.38	30.77
Pronlex	past results [20,21]	6.78	27.33
	bi-directional LSTM	6.51	26.69

G2P conversion (II)

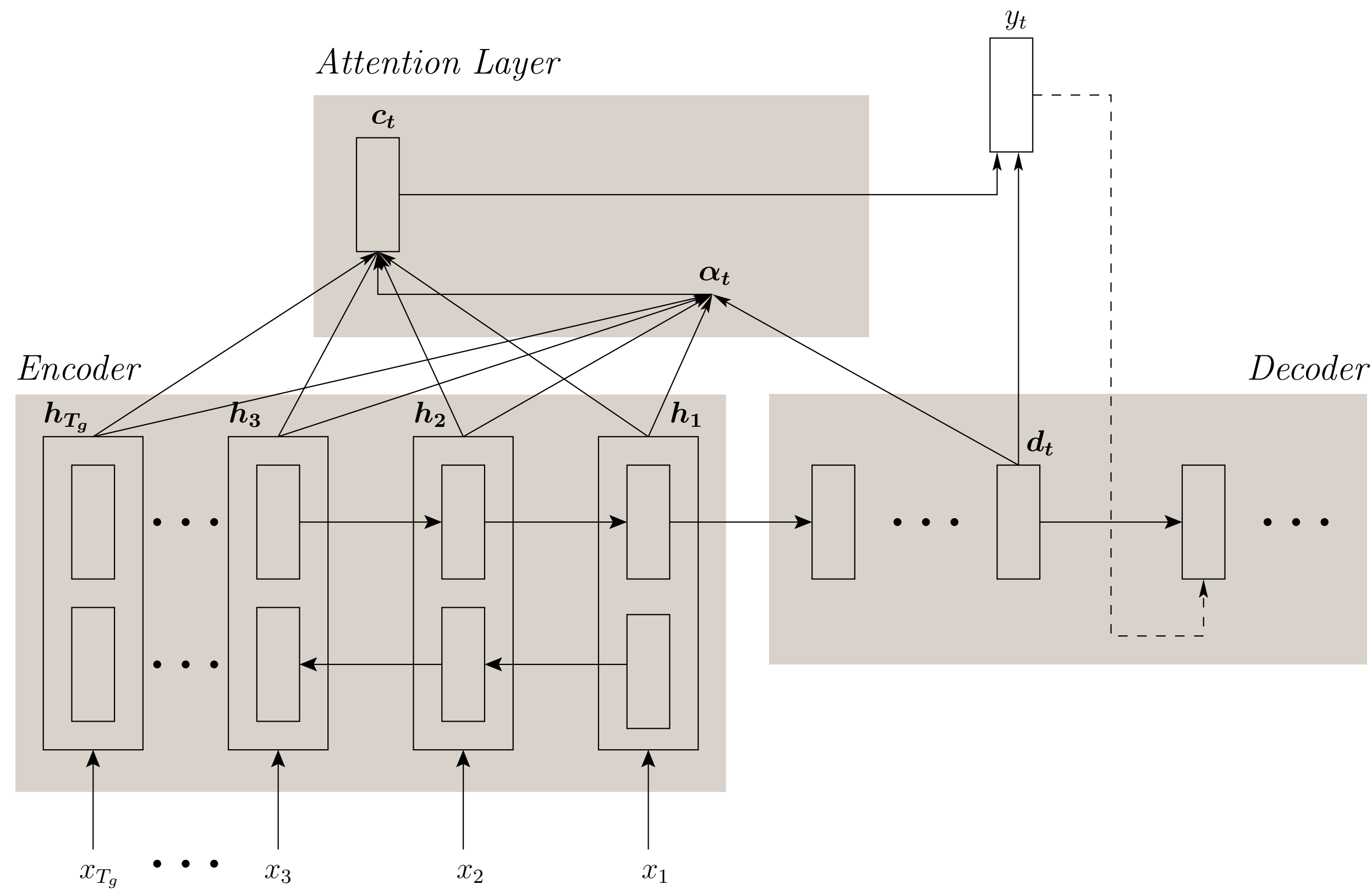
- Neural network based methods are the new state-of-the-art for G2P
- Bidirectional LSTM-based networks using a CTC output layer [Rao15]. Comparable to Ngram models.
- Incorporate alignment information [Yao15]. Beats Ngram models.
- No alignment. Encoder-decoder with attention. Beats the above systems [Toshniwal16].

[Rao15] Grapheme-to-phoneme conversion using LSTM RNNs, ICASSP 2015

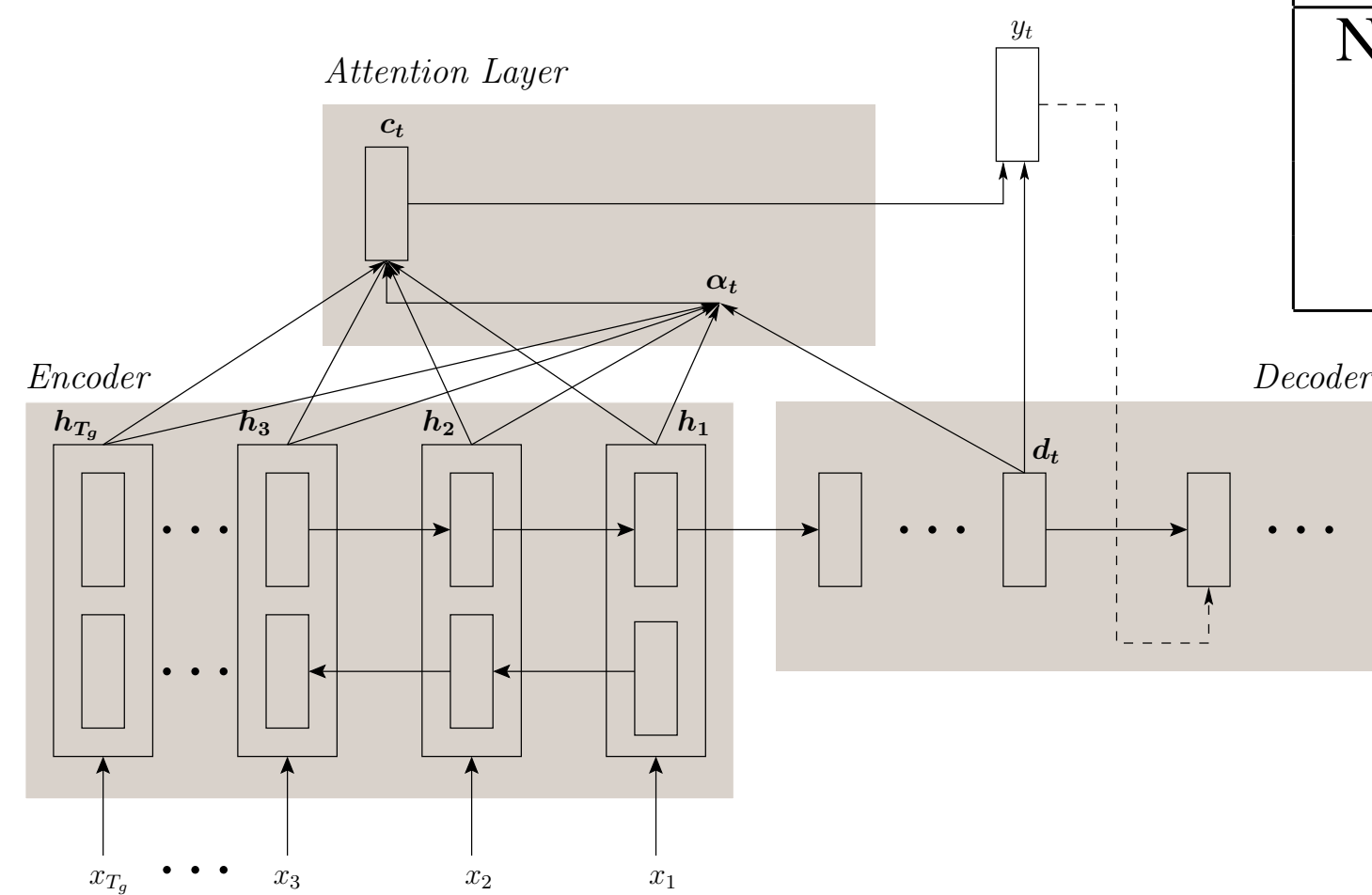
[Yao15] Sequence-to-sequence neural net models for G2P conversion, Interspeech 2015

[Toshniwal16] Jointly learning to align and convert graphemes to phonemes with neural attention models, SLT 2016.

Encoder-decoder + attention for G2P [Toshniwal16]



Encoder-decoder + attention for G2P [Toshniwal16]



Data	Method	PER (%)
CMUDict	BiDir LSTM + Alignment [6]	5.45
	DBLSTM-CTC [5]	-
	DBLSTM-CTC + 5-gram model [5]	-
	Encoder-decoder + global attn	5.04 ± 0.03
	Encoder-decoder + local- m attn	5.11 ± 0.03
	Encoder-decoder + local- p attn	5.39 ± 0.04
	Ensemble of 5 [Encoder-decoder + global attn] models	4.69
Pronlex	BiDir LSTM + Alignment [6]	6.51
	Encoder-decoder + global attn	6.24 ± 0.1
	Encoder-decoder + local- m attn	5.99 ± 0.11
	Encoder-decoder + local- p attn	6.49 ± 0.06
NetTalk	BiDir LSTM + Alignment [6]	7.38
	Encoder-decoder + global attn	7.14 ± 0.72
	Encoder-decoder + local- m attn	7.13 ± 0.11
	Encoder-decoder + local- p attn	8.41 ± 0.19

Related problem of interest: Transliteration

- Converting a sequence of graphemes from one script to another
 - adhyapak → अध्यापक
 - तीसरा → teesara
- Of relevance to end-to-end ASR systems dealing with code-switched speech, multilingual speech, etc.