Automatic Speech Recognition.

Artem Sokolov, HSE 2019

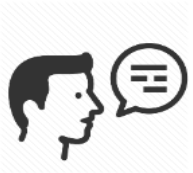
Automatic Speech Recognition (ASR)

- **Realtime stream speech recognition**
  - Do not have whole utterance to analyze
  - 1 sec of speech should take <= 1 sec of inference
  - Ex.: voice assistant

- **Recording recognition**
  - Have a recording of command for processing
  - Time limit is not strong
  - Ex.: transcription

Families

- **Hybrid ASR**



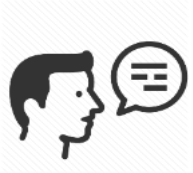
Feature extraction → Acoustic model → Pronunciation model → Language model → Text

MFCC → $o_3\ o_7\ o_7\ o_1\ o_9\ o_9\ o_9\ o_5$ → k æ t

k æ t → cat

the cat → “the cat”

- **End-to-end ASR**



End-to-end model

Acoustic model

Pronunciation model

Language model

→ Text

Models Overview

Model	Year	Architecture	Family	WER*, %
DeepSpeech2	2015	2 CNN + 7 BiRNN + 2 FC → CTC	e2e	5.33
CAPIO	2017	TDNN + TDNN-LSTM + CNN-bLSTM + Dense TDNN-LSTM	hybrid	3.19
ESPnet	2018	CNN (VGG)+N BiLSTM → CTC/CTC+Attention	e2e	4.0
Wav2Letter+	2018	Residual CNN + 19xCNN → CTC	e2e	3.44
RWTH ASR System	2019	HMM-DNN + LSTM LM + Transformer LM rescoring	hybrid	2.7
Jasper	2019	Residual Nx(CNN+BN+ReLU+Dropout) → CTC	e2e	2.95

***WER** – word error rate (the sum of insertions, deletions, substitutions divided by the total number of words)
Validated on LibriSpeech test-clean (corpus based on Public Domain audio books)

Language Model

Motivation

	Reference	E2E model output
Proper nouns	play the black **eyed peas** songs	play the black eye piece songs
Rare words	**echocardiography**	eco car biography
Accent and other	the **tail** of a dog	the tale of the dog

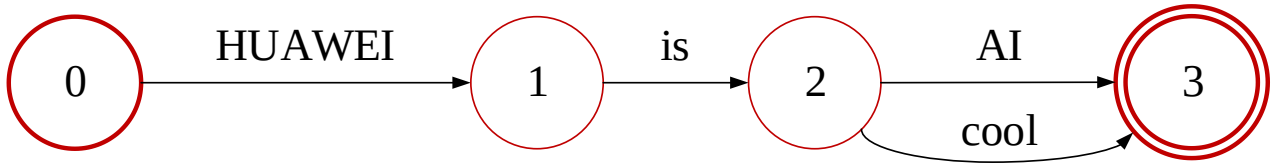
- E2E models requires ~**millions** audio-text utterances to train
- LM is trained on **billions** of sentences
- LM can bring contextual information to ASR

Approaches

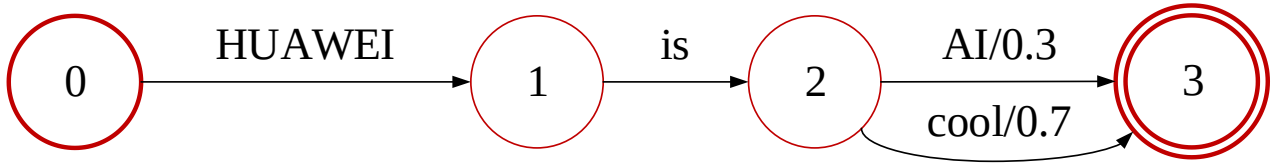
- **N-grams, Weighted Finite State Transducers (WFST)**
  - [Mohri et al., 2002] M. Mohri, F. Pereira, M. Riley, “Weighted Finite-State Transducers in Speech Recognition”
  - Most of ASR frameworks
- **RNN LM**
  - [Mikolov et al., 2010] T. Mikolov, M. Karafiat, L. Burget, “Recurrent neural network based language model”
  - [Kannan et al., 2017] A. Kannan, Y. Wu, P. Nguyen, “An analysis of incorporating an external language model into a sequence-to-sequence model”
  - ESPnet, EESEN end-to-end ASR frameworks
- **Transformers LM**
  - [Li et al., 2019] J. Li, V. Lavrukhin, B. Ginsburg, “Jasper: An End-to-End Convolutional Neural Acoustic Model”

Weighted Finite State Transducers

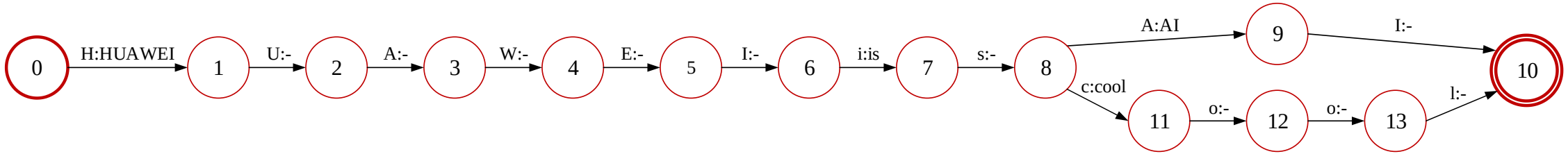
Finite State Acceptor:



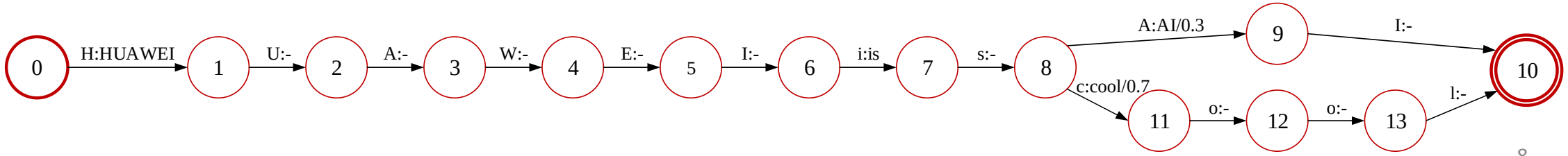
Weighted Finite State Acceptor:



Finite State Transducer:

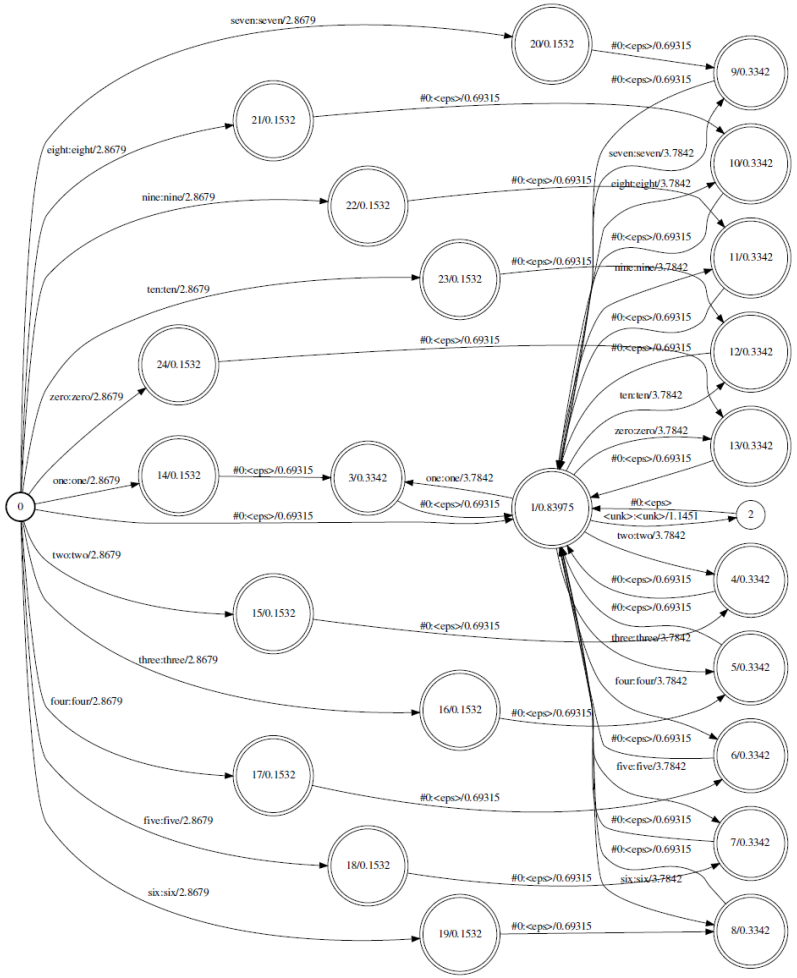


Weighted Finite State Transducer:

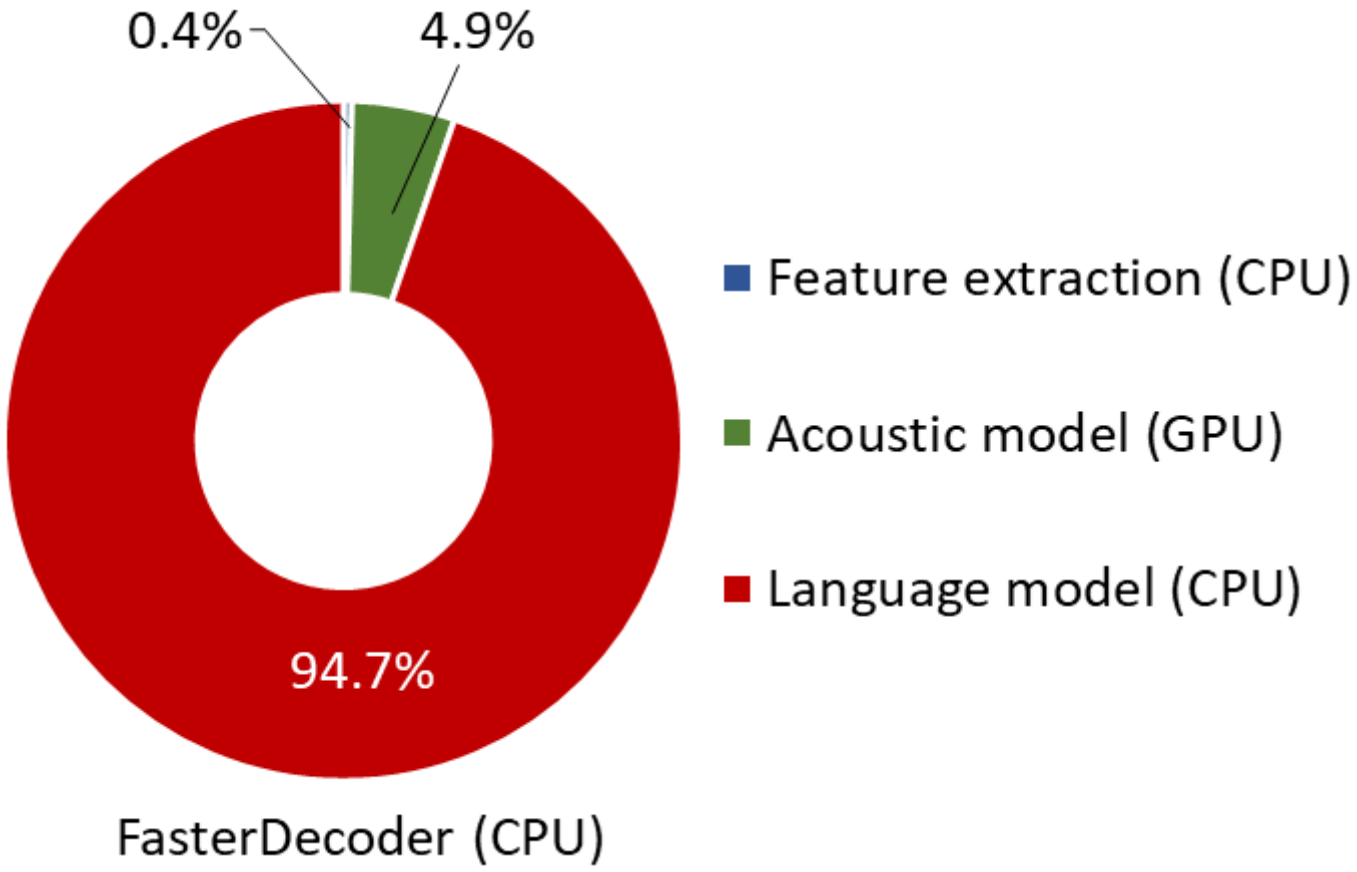


LM scoring is heavy

3-gram LM, 10 words



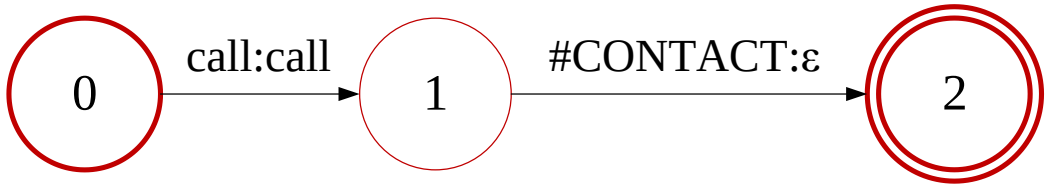
2018 GTC DC Nvidia, Hugo Braun, Justin Luitjens, Ryan Leary
"Accelerating automatic speech recognition"



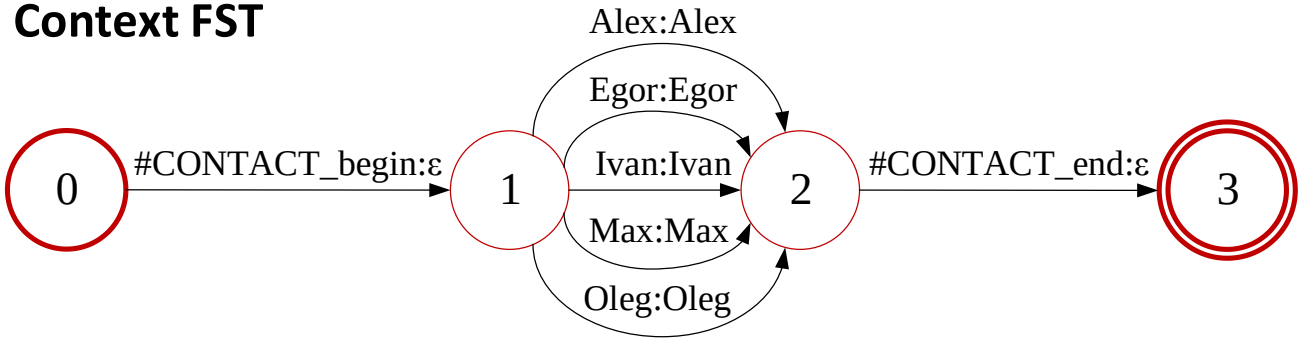
Biasing

"Bias" the priors to the speech models based on personal information (aka context).

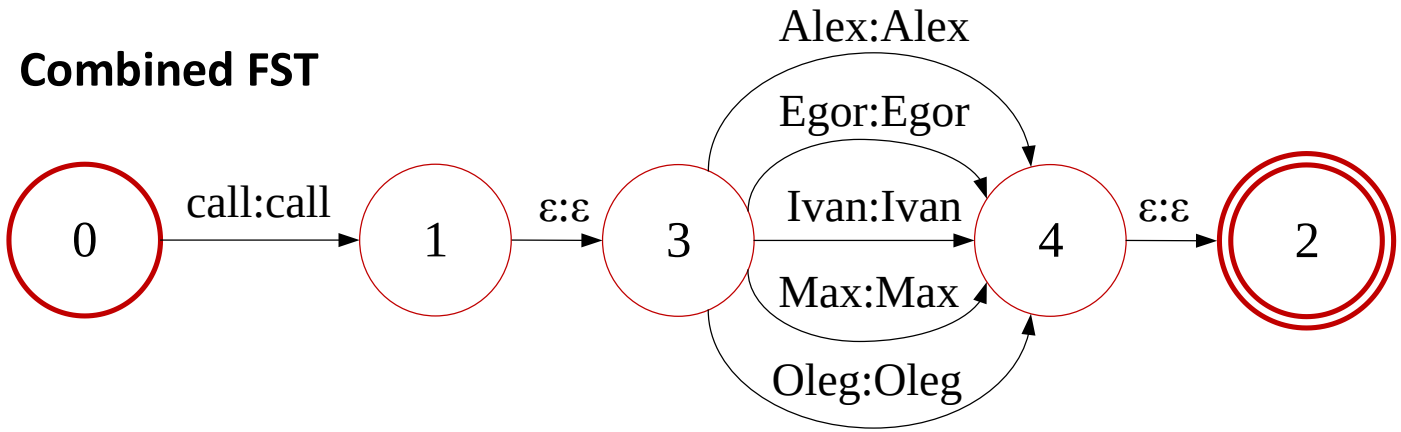
Original FST



Context FST



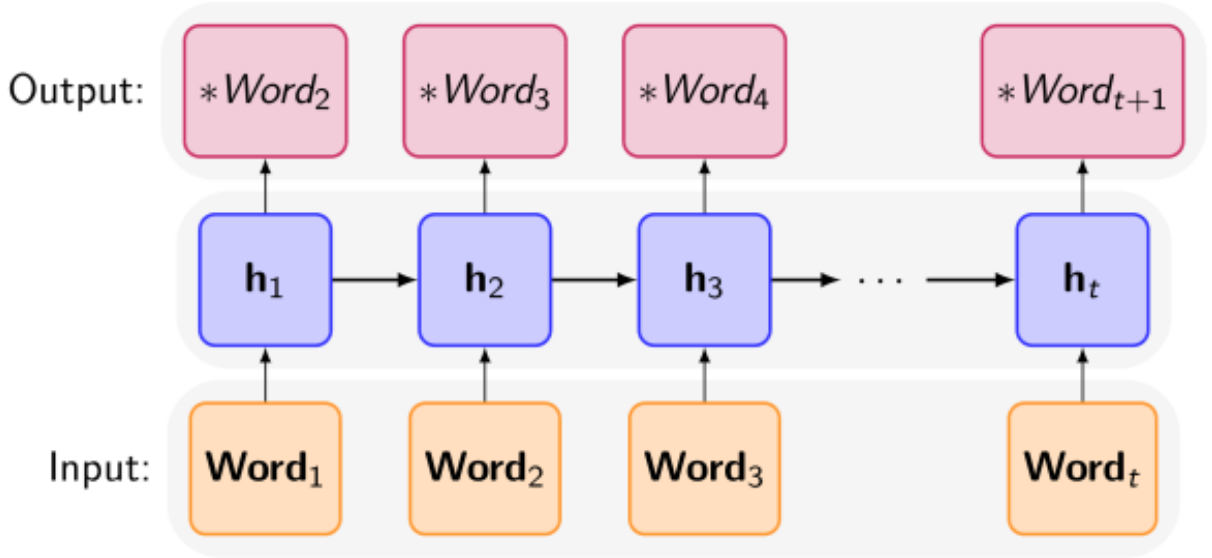
Combined FST



[Prabhavalkar et al.] Proc. of Interspeech 2018

Test Set	WER, No Biasing (%)	WER, Biasing (%)
Contacts	15.0	2.8
Numeric	11.0	4.7
Yes-No-Cancel	18.8	10.4

RNN LM



[Kannan et al., 2018]

System	Dev	Test
LAS-G	13.0	10.3
LAS-G + 20-GRAM-G	10.3	7.7
LAS-G + 3-GRAM-W	10.0	7.6
LAS-G + RNN-G	9.3	6.9

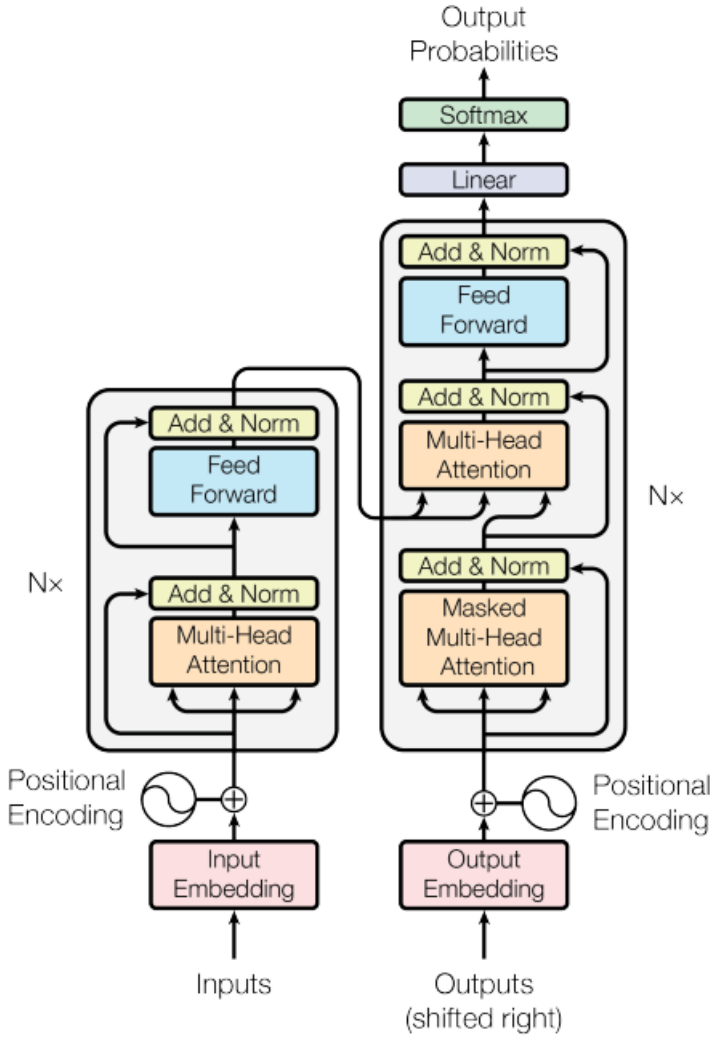
Table 1: WER of LAS-G fused with various LMs. While word constraints do help the n-gram LM, RNN-G performs even better.

System	Dev	Test	LM size
LAS-G	9.5	7.7	0GB
LAS-WP	9.2	7.7	0GB
LAS-WP + PRODLM1	8.8	7.4	2GB
LAS-WP + PRODLM2	8.7	7.2	80GB
LAS-WP + RNN-WP	8.4	7.0	1.1GB
LAS-WP + RNN-WP + PRODLM2	8.4	7.0	81.1GB

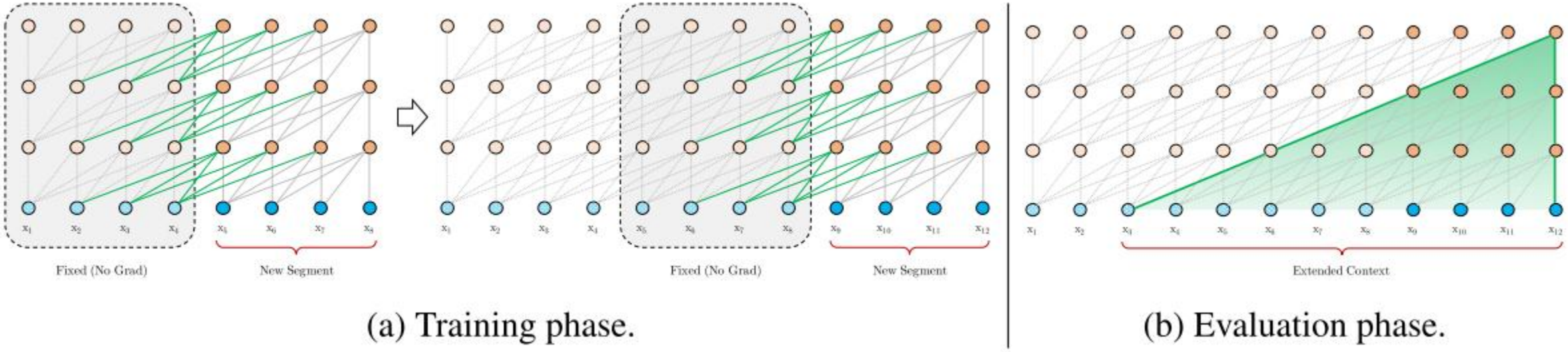
Table 3: WER of shallow fusion of LAS with production n-gram LMs and an RNN LM. The RNN LM captures all the benefits of PRODLM2 in a compact form.

Transformers

- Transformers came from translation tasks
- Transformer uses encoder-decoder architecture
- Do not use recurrent layers
- Use additional positional encoding instead of memory of RNN layers



Transformer-XL LM



Transformer-XL – training and evaluation phase

[Li et al., 2019] J. Li, V. Lavrukhin, B. Ginsburg, "Jasper: An End-to-End Convolutional Neural Acoustic Model", April 2019

Table 5: *LibriSpeech, WER (%)*

Model	E2E	LM	dev-clean	dev-other	test-clean	test-other
Jasper DR 10x5	Y	-	3.64	11.89	3.86	11.95
Jasper DR 10x5	Y	6-gram	2.89	9.53	3.34	9.62
Jasper DR 10x5	Y	Transformer-XL	2.68	8.62	**2.95**	8.79

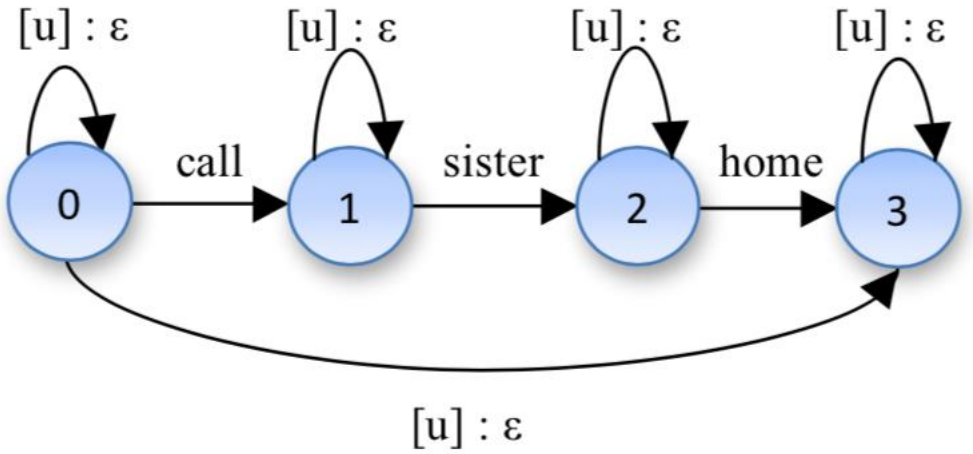
Table 6: *WSJ End-to-End Models, WER (%)*

Model	LM	nov93	nov92
Jasper 10x3	-	16.1	13.3
Jasper 10x3	4-gram	9.9	7.1
Jasper 10x3	Transformer-XL	9.3	6.9

Experiments

Context Depended Grammar

- Hand crafted rules
- Automatically align some variants to one pattern.
- Could include with biasing domains



Data

Framework:

- Kaldi ASpiRE receipt
- TDNN, BiLSTM models

Speakers description:

- 1 male 1 female
- English is a second language of each

Example:

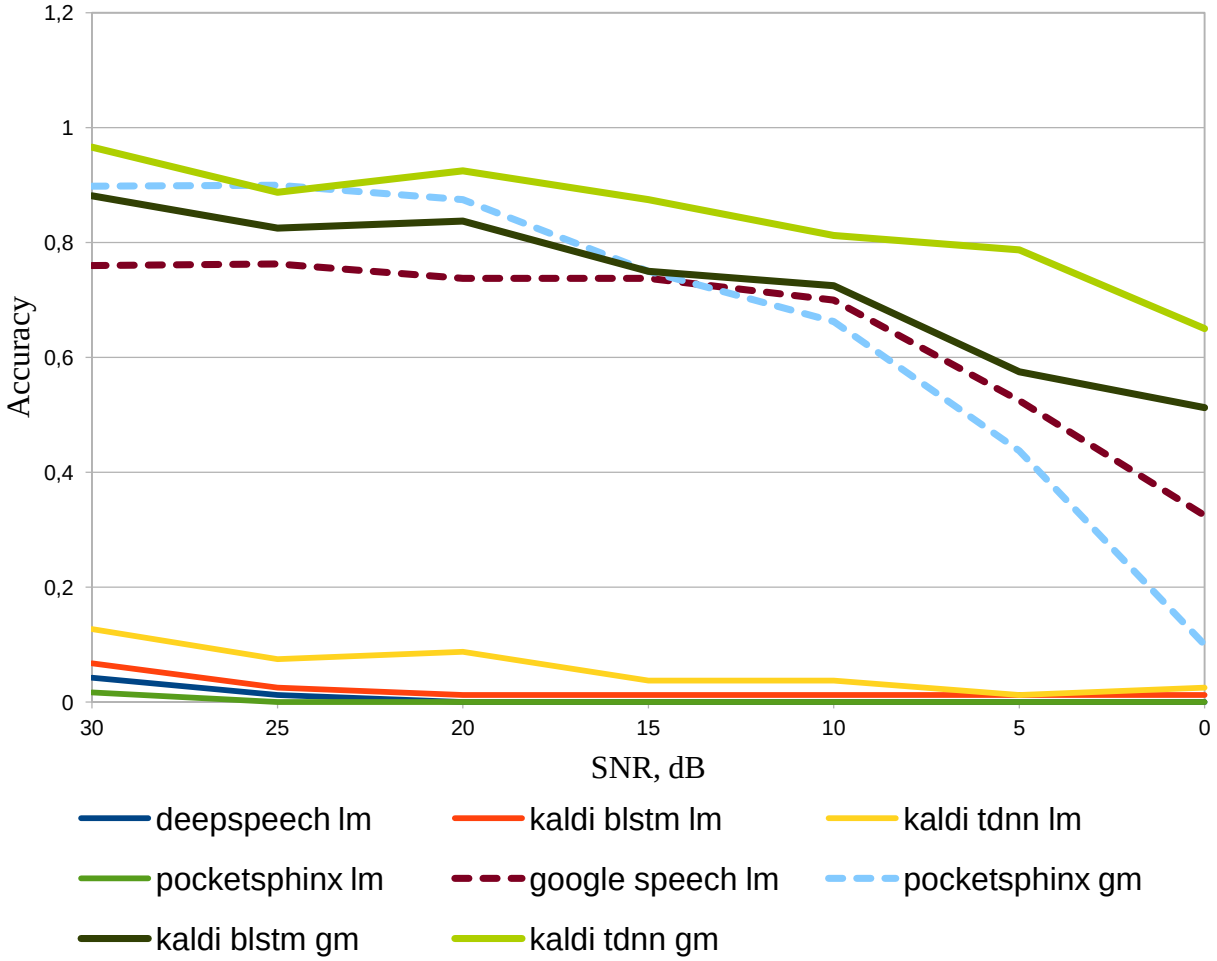
- *Turn on the light*
- *Clean the kitchen*
- *Lower the music*
- *Pause the track*
- *Call sister home/mobile*

Data description:

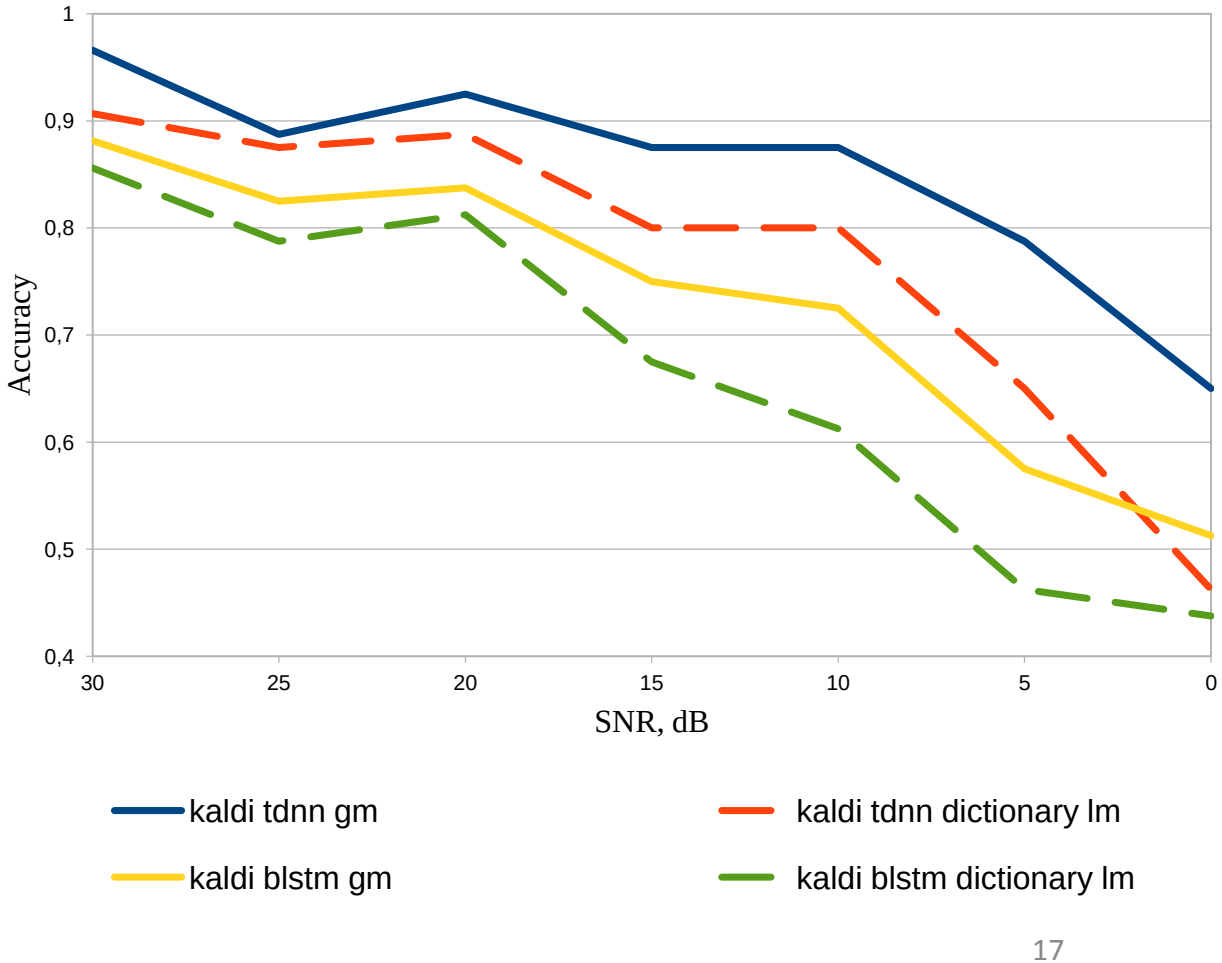
- Vocabulary: 130 commands
- Used for inference 80 commands
  - Shortest: 2-word commands – 2%
  - Longest: 8-word commands – 2%
  - Popular:
    - 4-word commands – 27%
    - 3-word commands – 19%
  - Other commands: 5-7 word length
- Signal Noise Ratio (SNR): 30 – 0dB with step 5dB

Results

Comparison with full language models



Comparison with language models based on command vocabulary



Discussion

- Grammar outperforms statistical LM
- Grammar is resistant to noise and accent
- Grammar is being used in production (Speech Technology Centre)

Thank you!