Speech recognition and reinforcement learning

Andrew Stepanov

Outline

- History of ASR
- Sound representation
- The alignment problem
- Connectionist temporal classification
- Fine tuning ASR with Reinforcement Learning

History of ASR

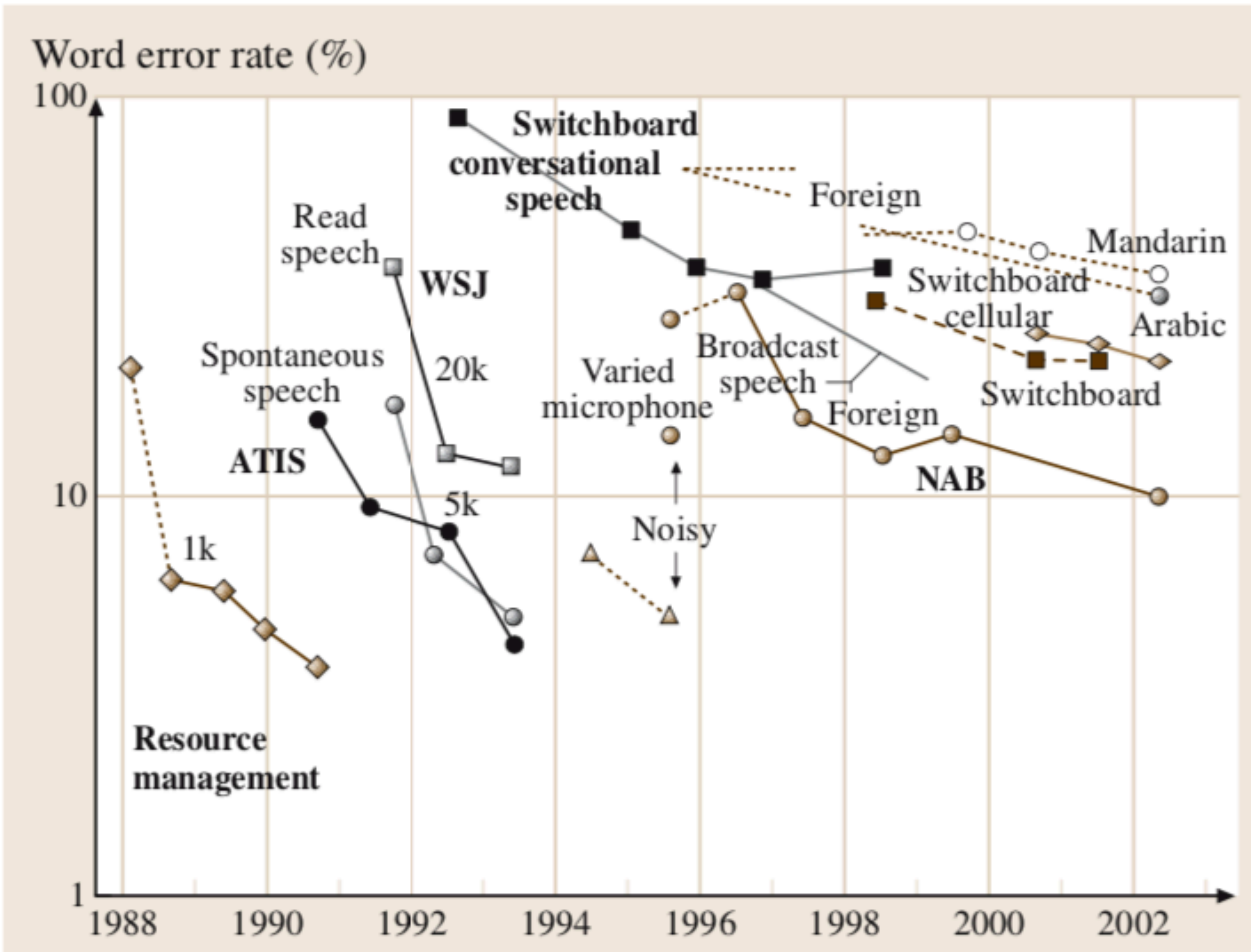
- Early history (1940 — 1960)
- Pattern recognition approach (1960 — 1980)
- Statistical Modelling (1980 — 2010)
- Deep learning (2010 — now)

Metrics

$$\mathrm{WER}(r, h) = \frac{\mathrm{distance}(r, h)}{\mathrm{length}(r)}$$

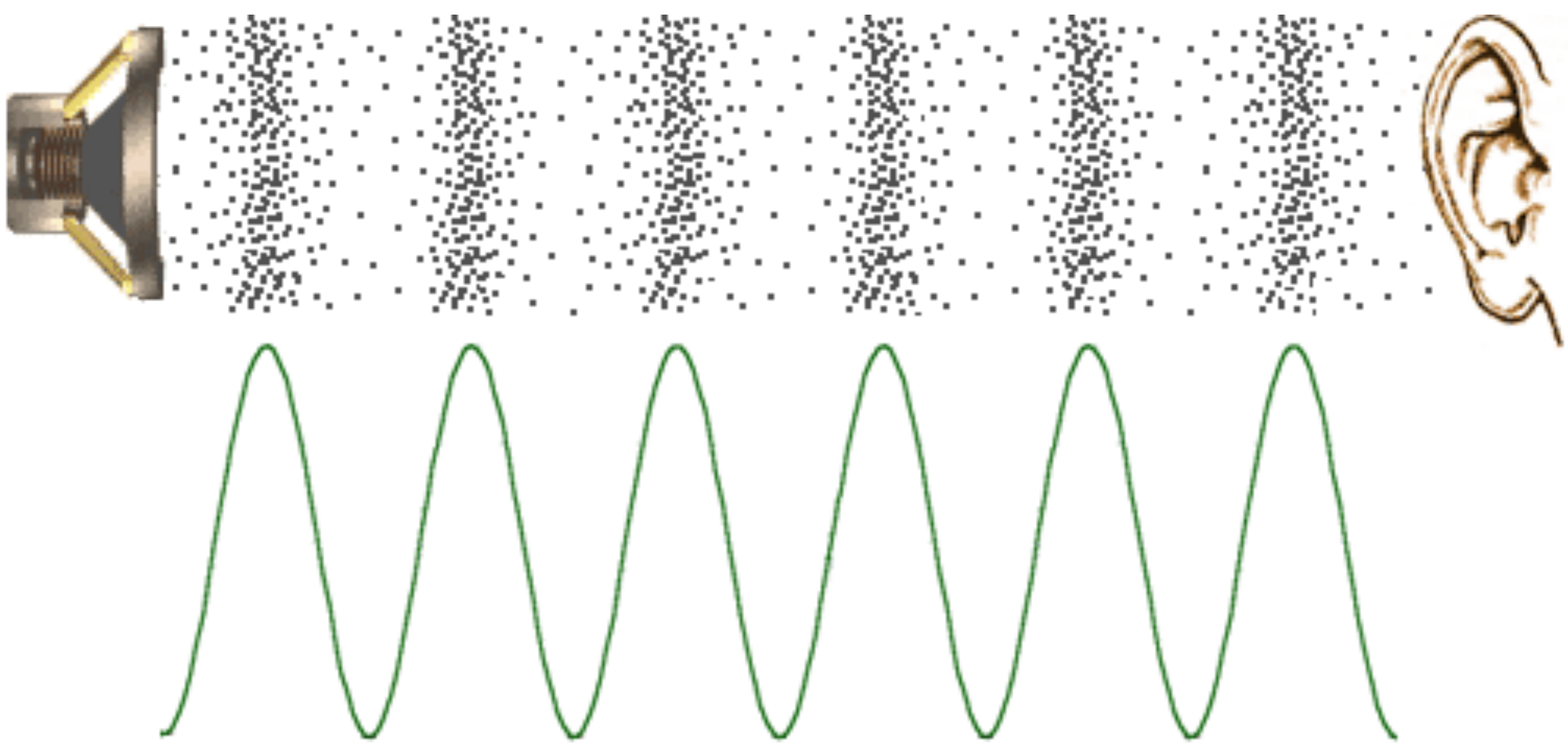
Error type	Text	WER
Reference	quick brown fox jumps over the lazy dog	0
Insertion	quick fluffy brown fox jumps over the lazy dog	**0,125**
Deletion	quick _ fox jumps over the lazy dog	**0,125**
Substitution	quick brown fox jumps over the crazy dog	**0,125**
Composition of above	quick _ fox jumps over the crazy dog	**0,25**

WER performance over time

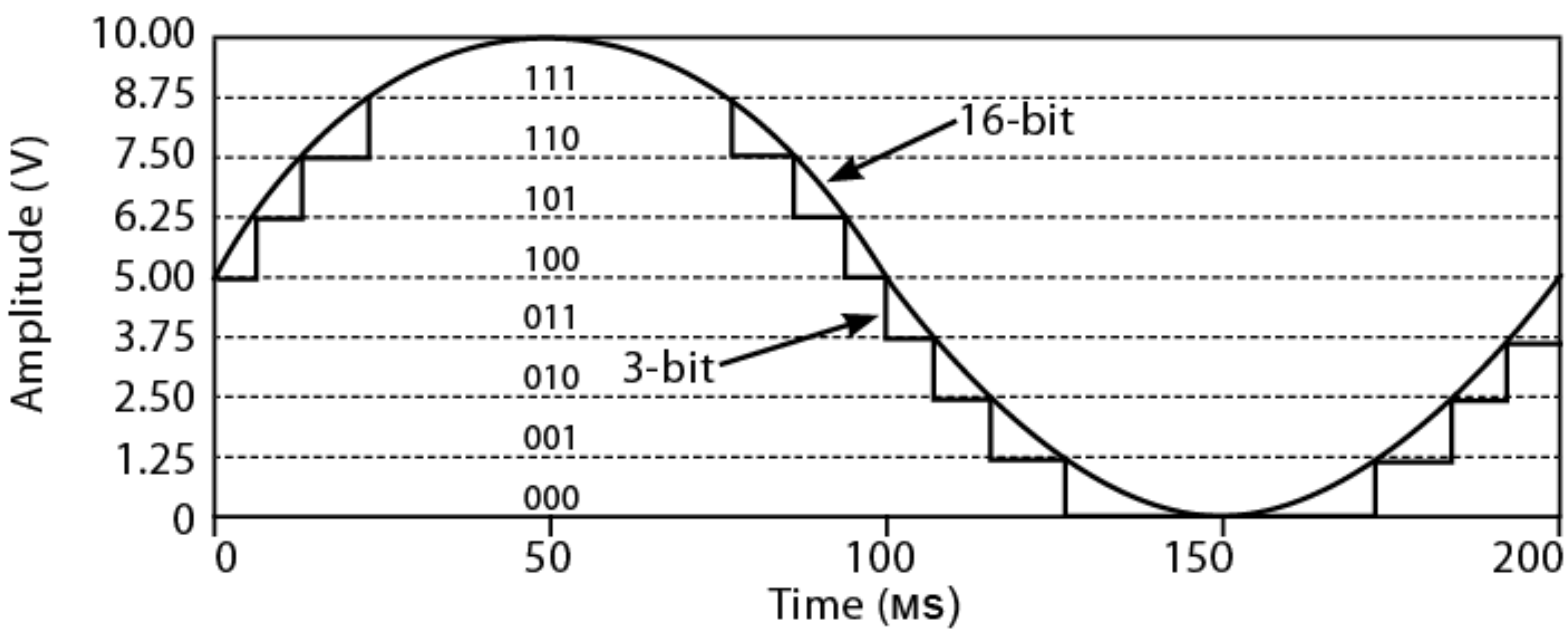


* Springer Handbook of Speech Processing

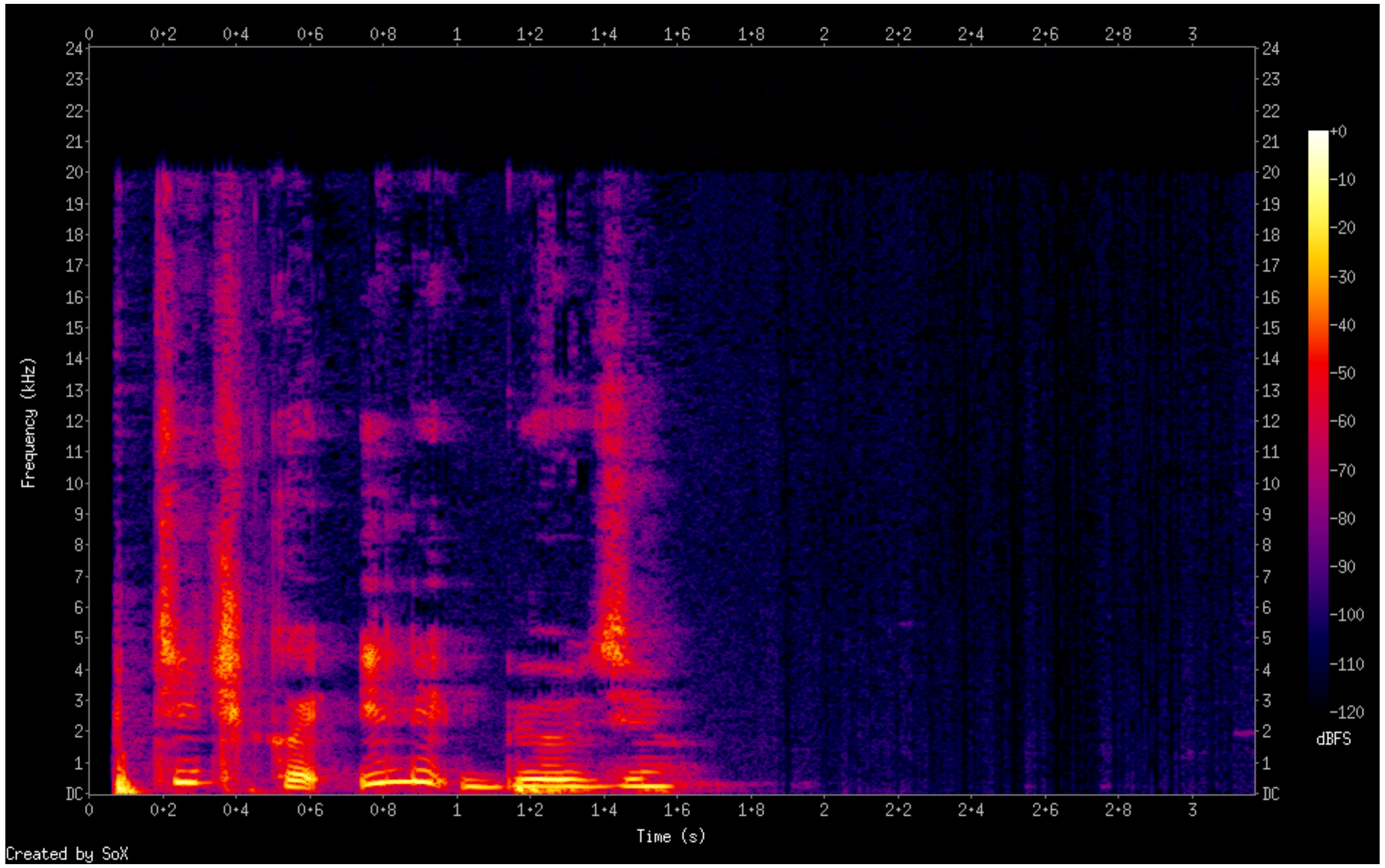
Sound waves

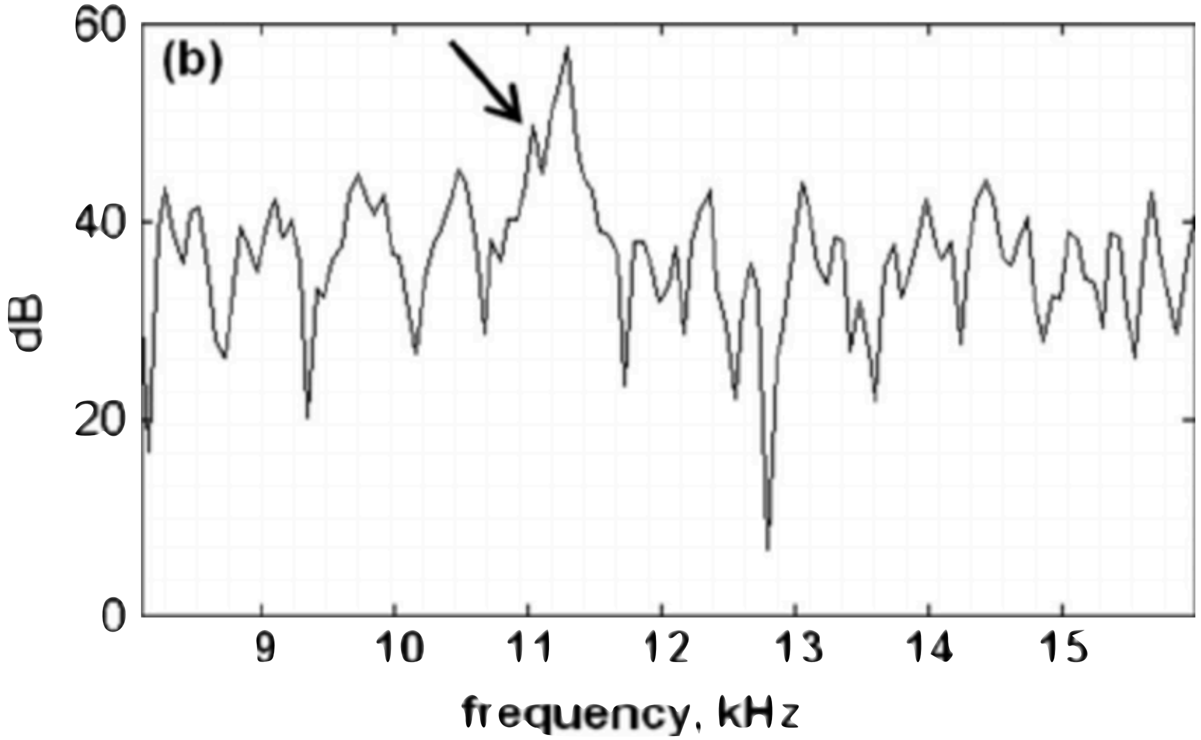


Sound quantization



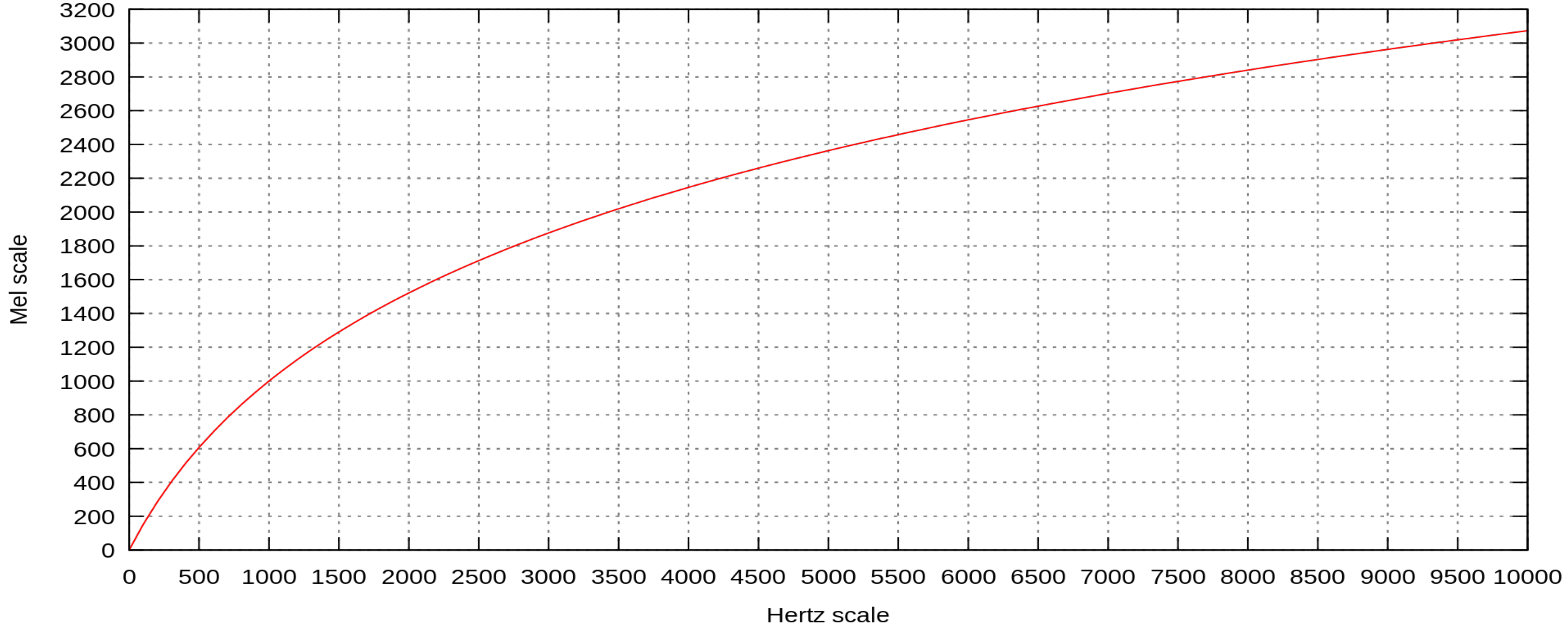
Spectrogram



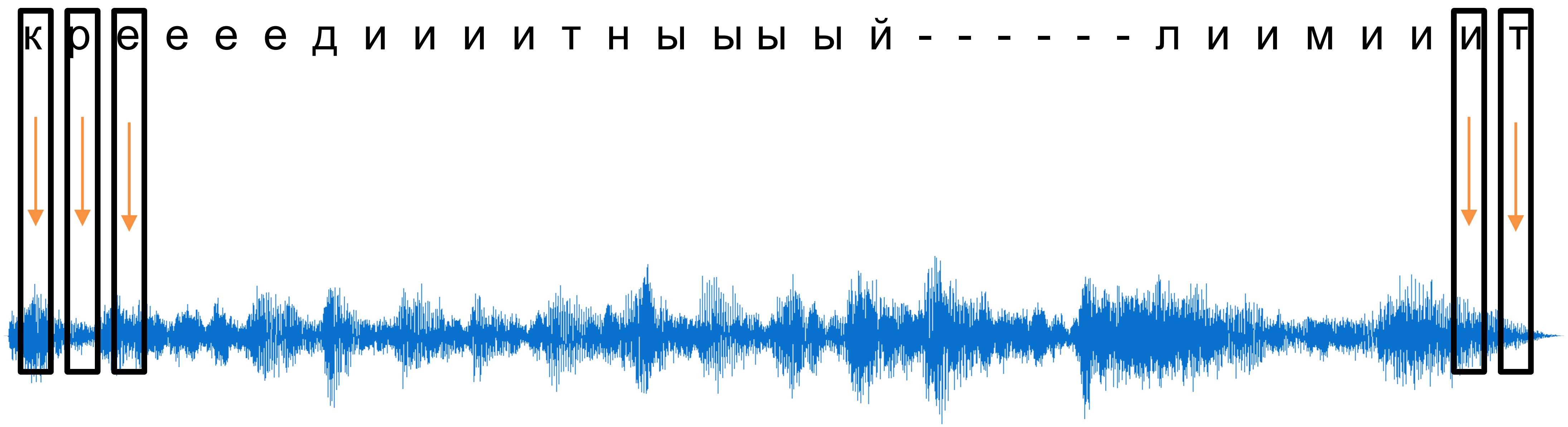


Mel scale

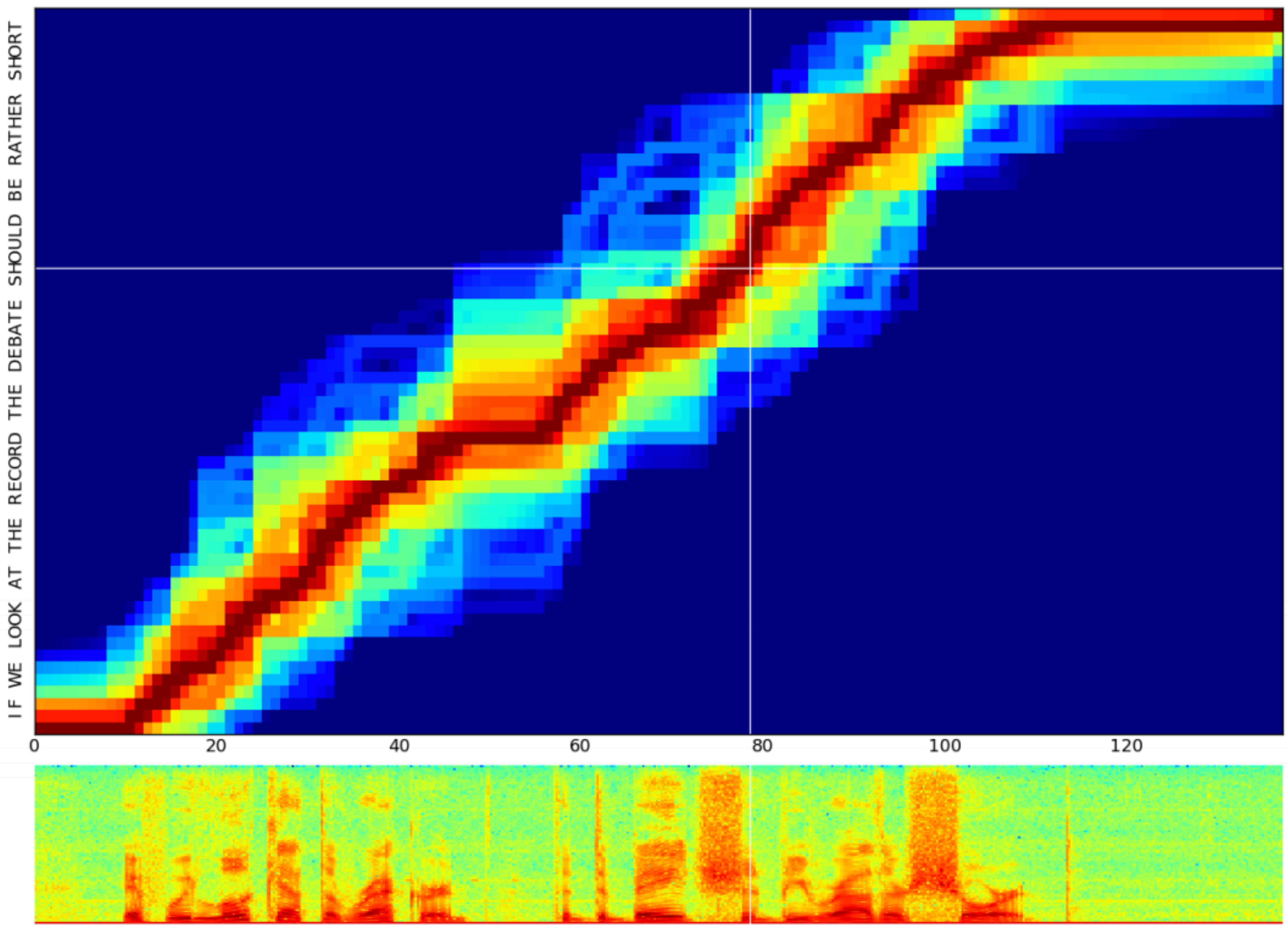
$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$



Alignments

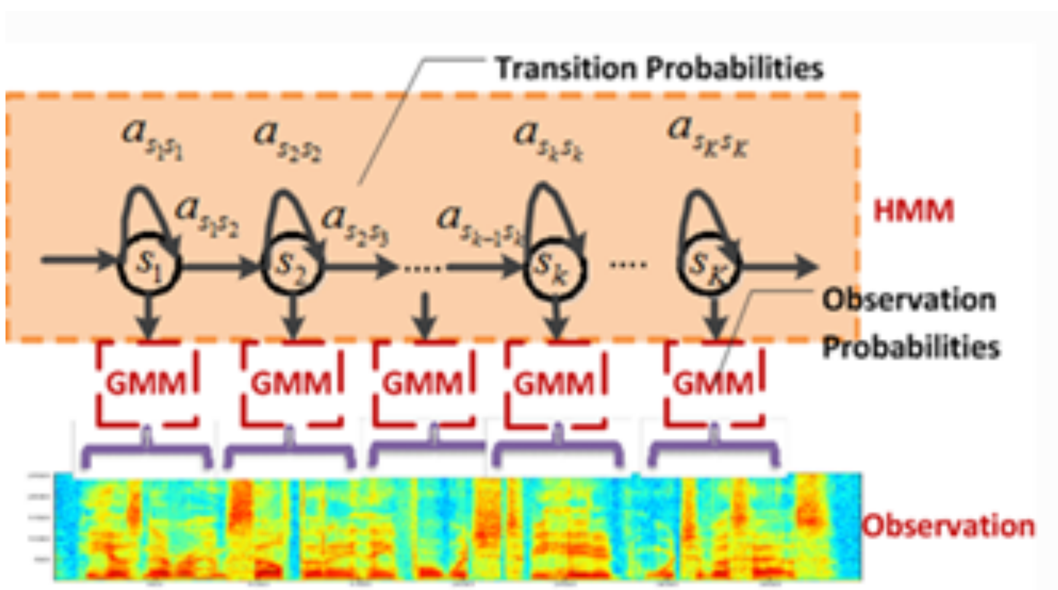


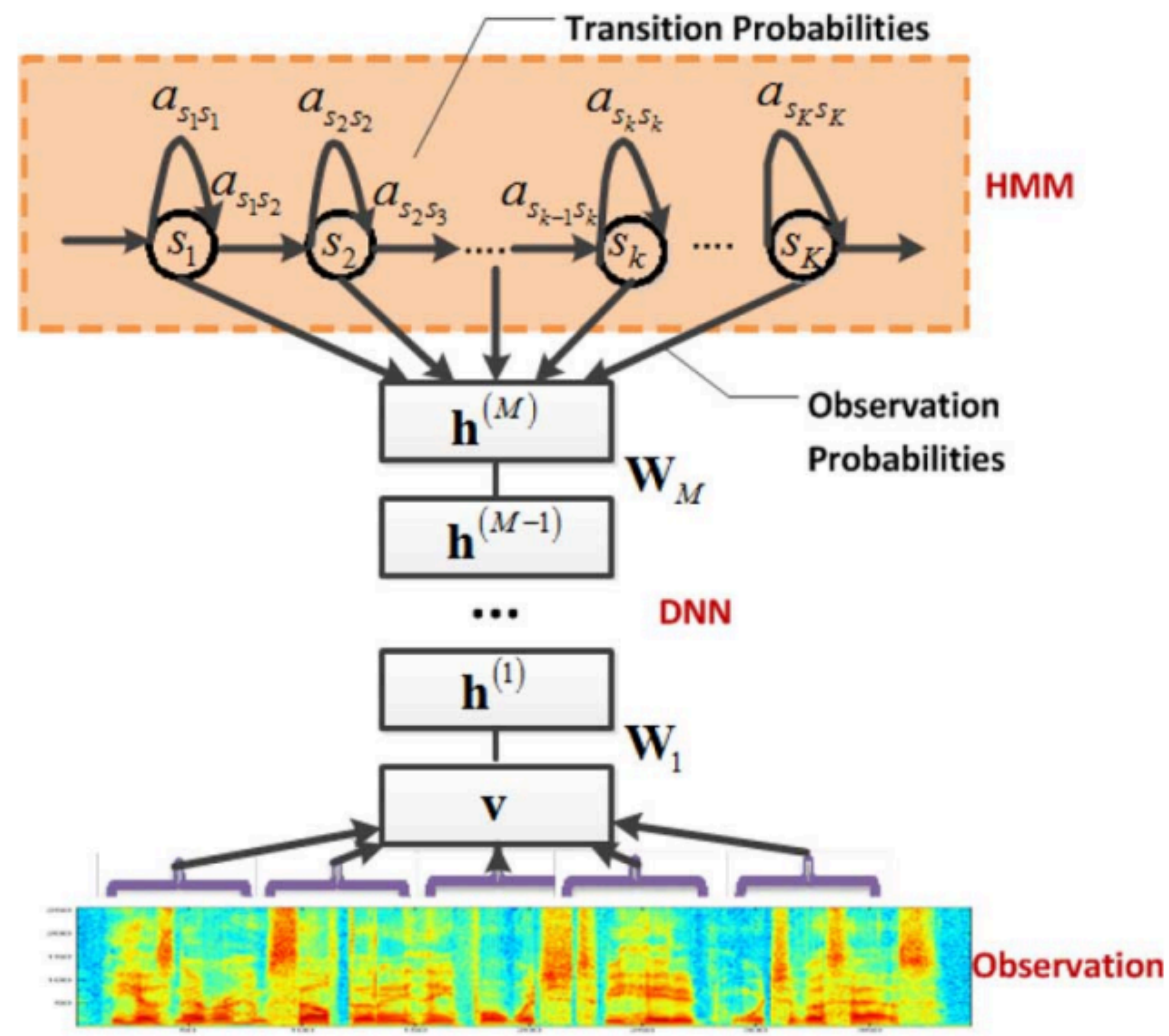
Alignments



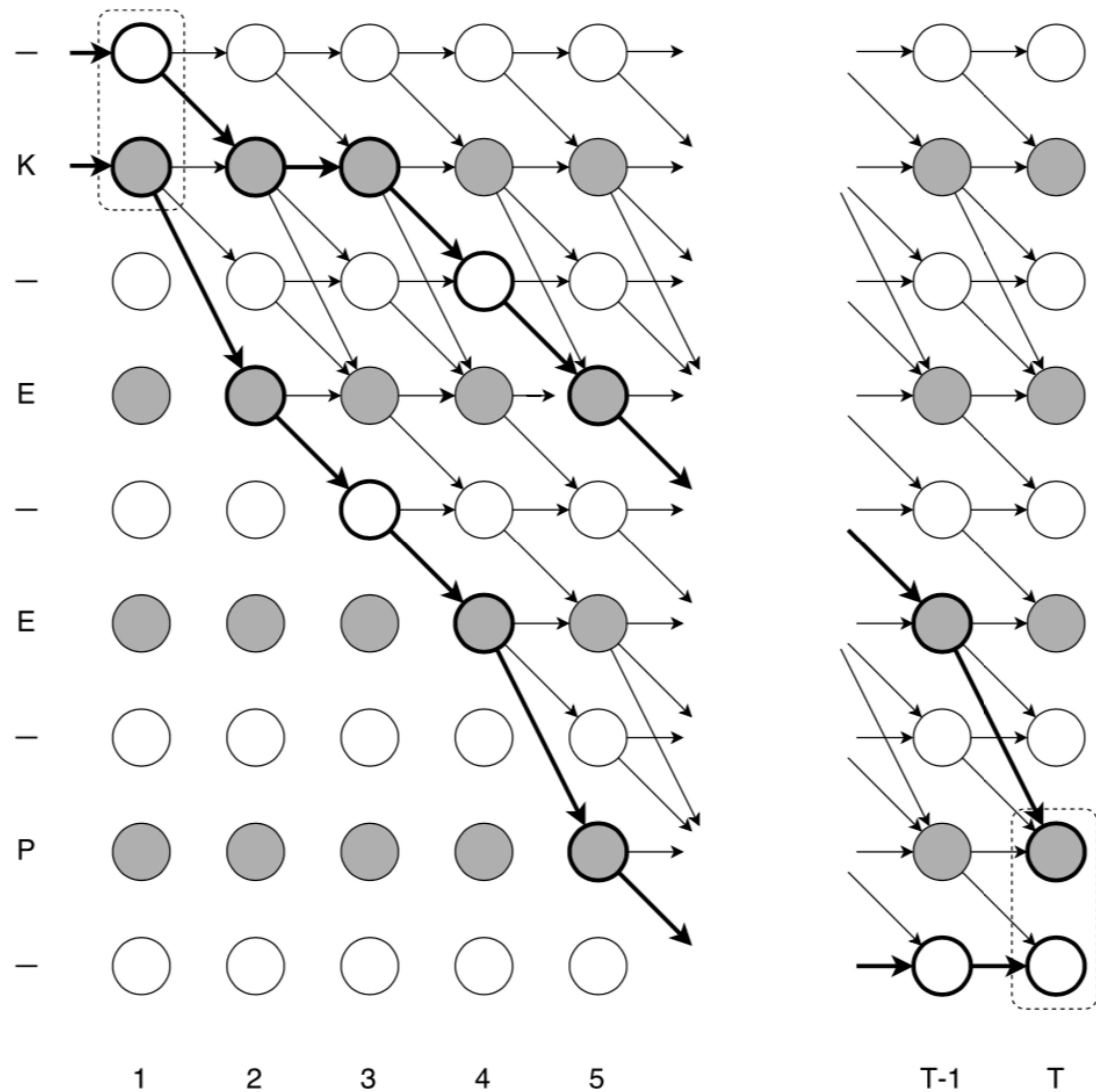
* Sequence Transduction with Recurrent Neural Networks, Graves, 2012

GMM-HMM and DNN-HMM

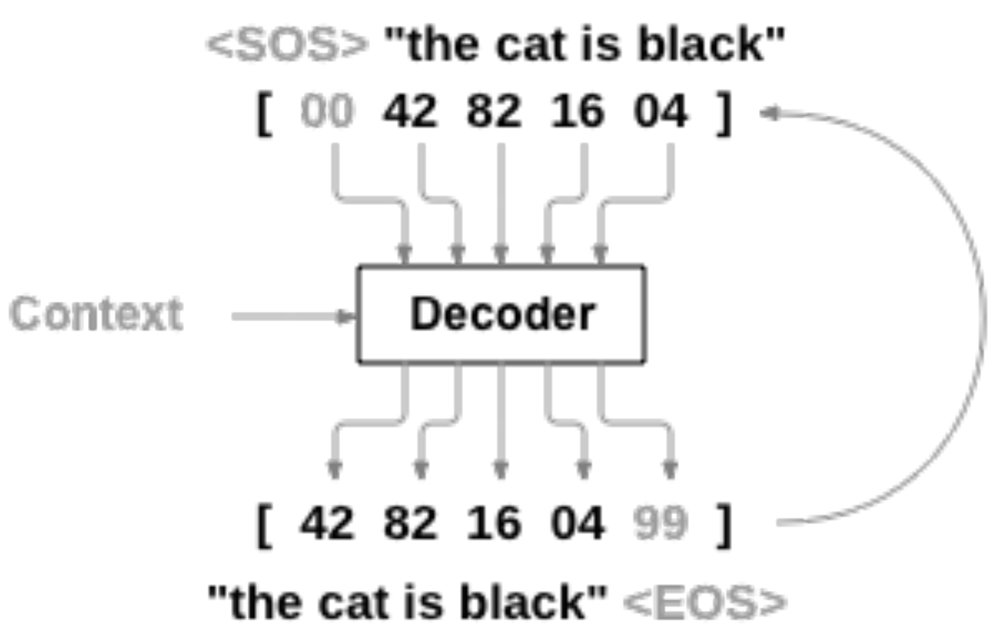


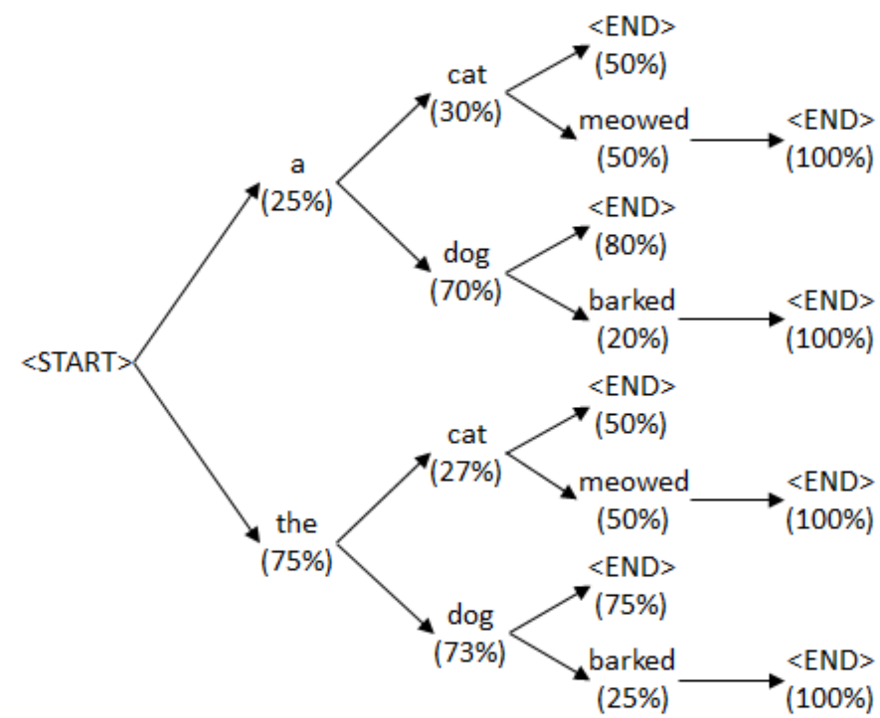


CTC



Mismatch between training and inference





CTC / Teacher forcing for LAS

Beam search decoder

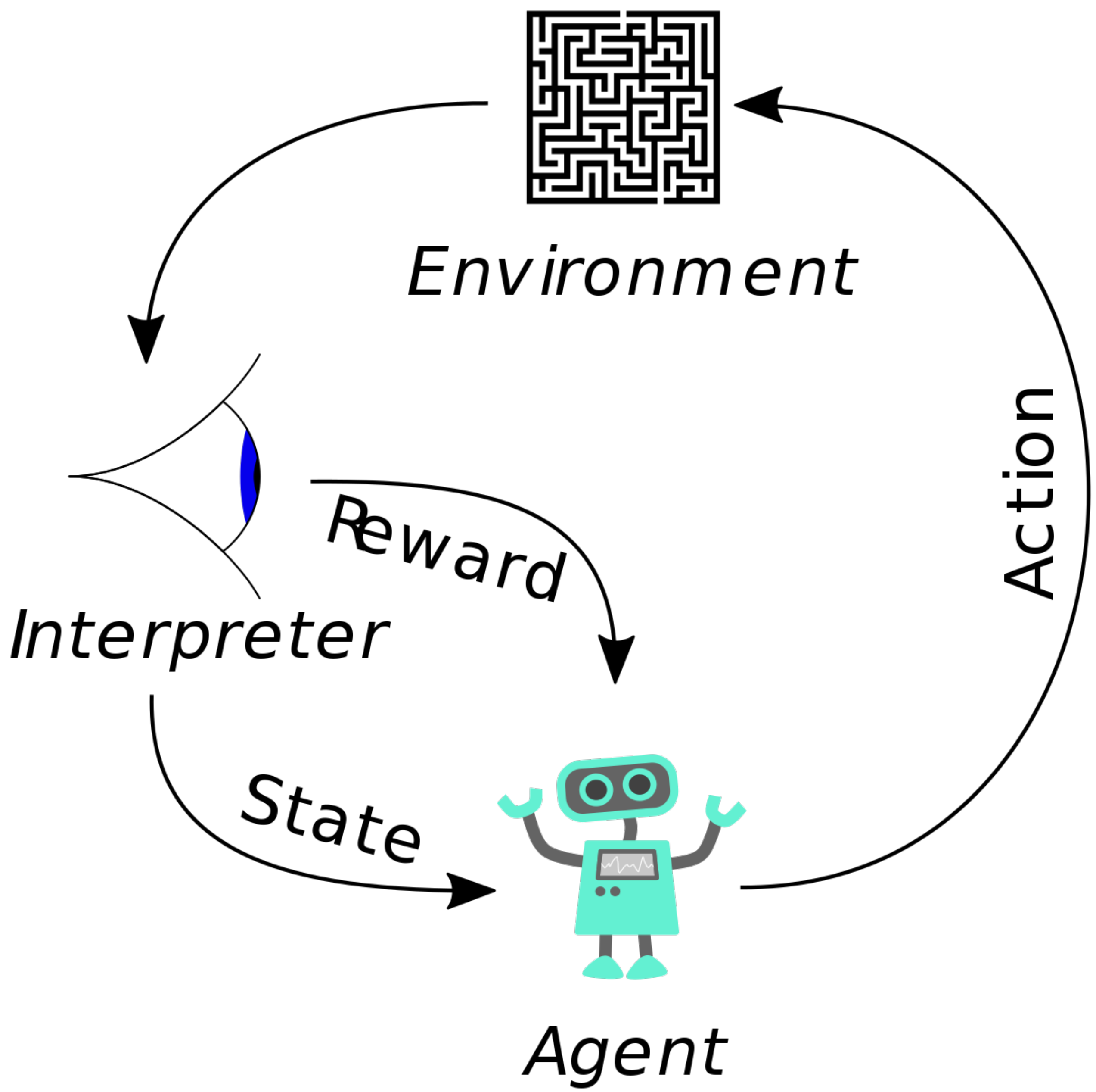
Fighting mismatch between training and inference

- Do nothing (works well for DNN-HMM, CTC: no feedback loop between model output and decoder)
- For LAS, RNN-Transducer:
  - Scheduled sampling (Chan et al., 2015)
  - Simple first-pass decoding + second-pass rescoring (Chen et al., 2017; Chiu et al., 2018)

Fighting mismatch between training and inference

- MMI/MPE/sMBR/bMMI (Veseley, Povey, 2013;, Yu, Dang, 2014)
- Expected Transcription Loss (Graves, 2014)
- CD-CTC-sMBR (Sak et al., 2016)
- Minimum Word Error Rate Training (Chiu et al, 2018)

Reinforcement Learning in ASR



Sampled MBR criterion

$$\mathbb{E}_{y\sim P(y|x)}L(y,y^*) = \sum_y P(y|x)L(y,y^*) = \sum_y \sum_{\pi:B(\pi)=y} P(\pi|x)L(y,y^*) = \sum_\pi P(\pi|x)L(B(\pi),y^*) \quad (17)$$

$$\frac{\partial}{\partial z}\mathbb{E}L(\pi) = \frac{\partial}{\partial z}\sum_\pi P(\pi|z)\cdot L(\pi) = \sum_\pi L(\pi)\cdot\frac{\partial}{\partial z}P(\pi|z) = \sum_\pi L(\pi)P(\pi|z)\frac{\partial}{\partial z}\log P(\pi|z) = \mathbb{E}\Big[L(\pi)\cdot\frac{\partial}{\partial z}\log w(\pi|z)\Big] - \mathbb{E}L(\pi)\cdot\mathbb{E}\Big[\frac{\partial}{\partial z}\log w(\pi|z)\Big] \quad (18)$$

Gradient approximation and Reinforcement Learning

$$\frac{\partial}{\partial z}\mathbb{E}L(\pi) \approx \frac{1}{N-1}\sum_{i=1}^{N}(L(\pi_i) - \overline{L}_{batch})\frac{\partial}{\partial z}\log w(\pi_i|z)$$

$$\frac{\partial}{\partial z}\mathbb{E}L(\pi) \approx \frac{1}{N}\sum_{i=1}^{N}L(\pi_i)\frac{\partial}{\partial z}\log w(\pi_i|z)$$

$$J(\theta) = \mathbb{E}_{a\sim\pi(a|s)}R(a)$$

$$\nabla J(\theta) = \frac{1}{N}\sum_{i=1}^{N}(R(a_i) - \hat{R}_{baseline})\nabla\log\pi(a_i|s)$$

Results

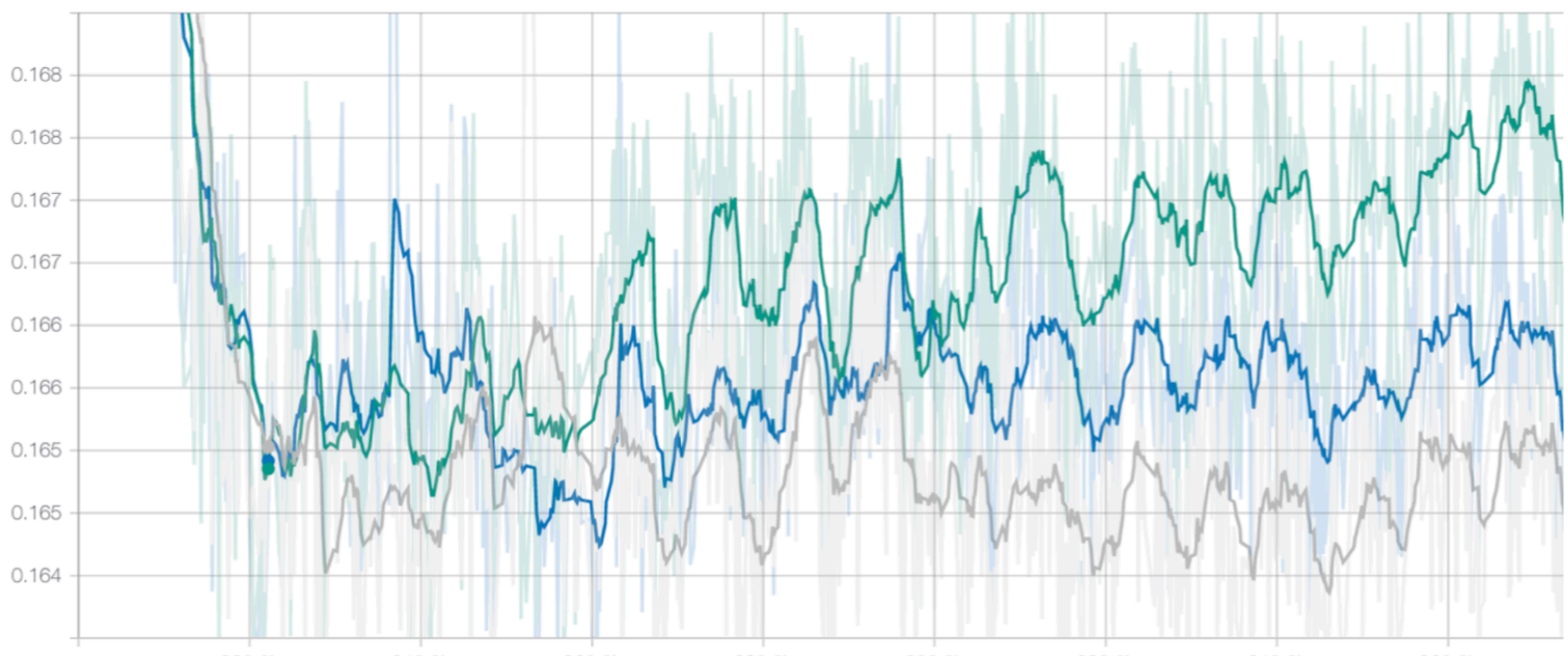


Рис. 21: Sampled MBR с различными значениями beta. Серый – beta=0.01, синий beta=0.005, зелёный – beta=0.02

Results

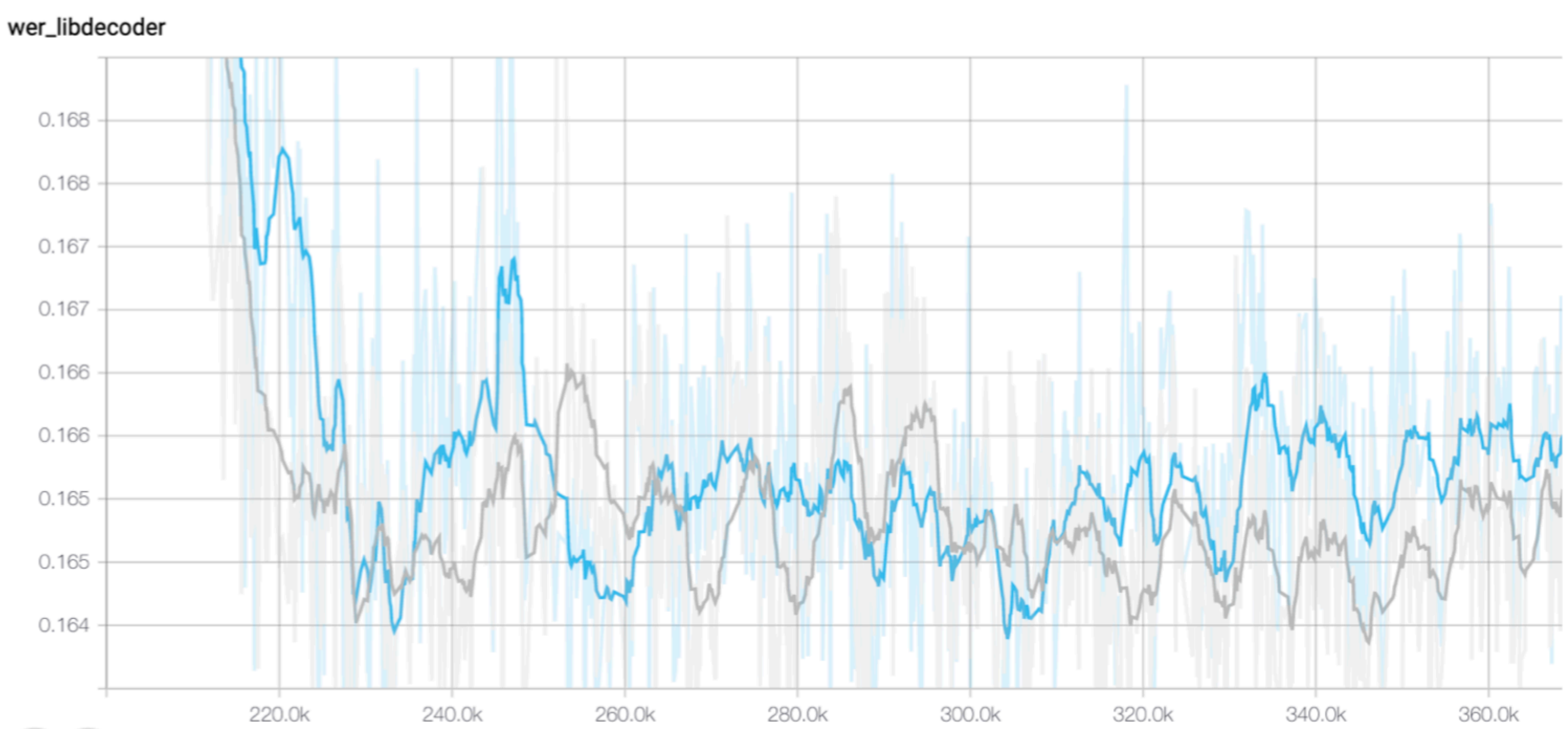


Рис. 22: Sampled MBR с различными значениями языковыми моделями. Серый – биграмная слабая языковая модель на транскрипциях, синяя – сильная 5 грамная модель на разных источниках данных. Однозначный вывод сделать нельзя.

Results

	dev	dictation	queries noisy	calls noisy
Baseline	18.1	5.6	28.4	18.2
Default MBR	16.17	5.2	24.4	17.8
+ 5gram	16.21	5.1	24.4	17.9
+ beta=0.02	16.16	4.9	24.2	17.6

Overall 5%-15% relative WER reduction depending on the dataset