Listen, Attend and Spell

Winter Vacation Capstone Study

TEAM Kai.Lib

발표자 : 배세영

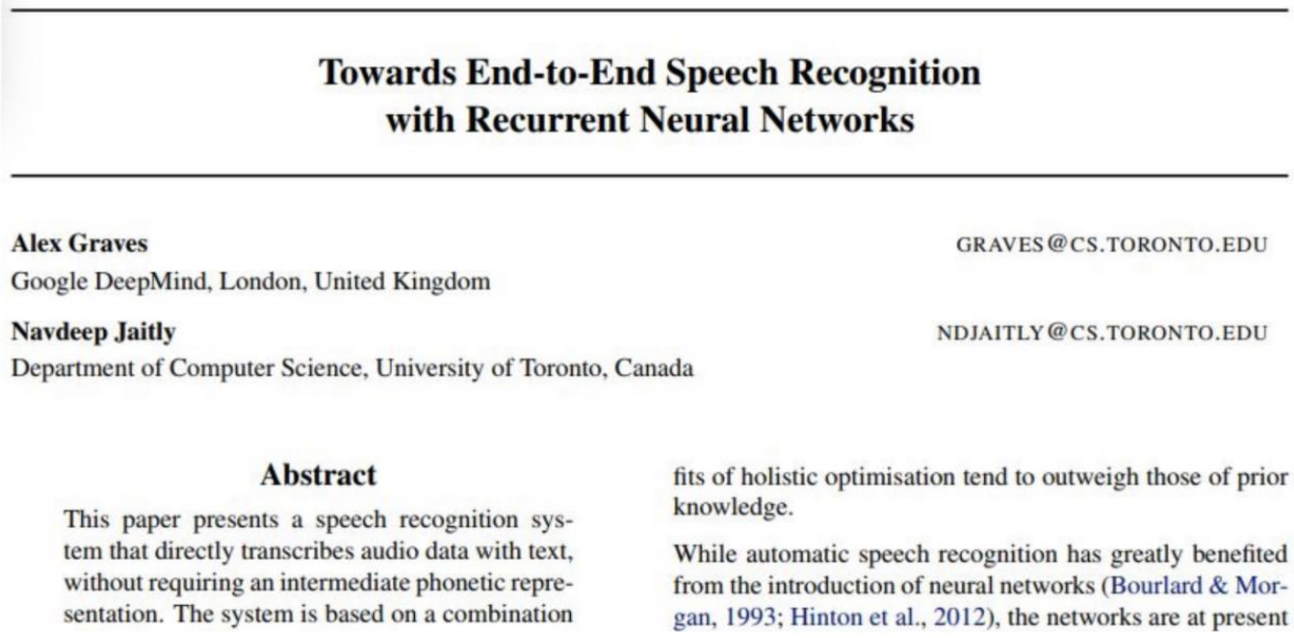
2020.02.03 (MON)

End-to-End ASR

- Automatic Speech Recognition (ASR)
  - 특징 추출, Acoustic Model, Decoder 등을 사용하는 파이프라인 구조
  - CTC 제안 [Graves et al., 2006]
  - 비슷한 시기에 LAS 제안 [Google Brain Team]

End-to-End ASR

- CTC (Connectionist Temporal Classification)
  - Uni/Bi Directional RNN (LSTM)의 다층 구조
  - 인코더로 들어오는 입력 데이터는 최종적으로 Softmax를 통과하여 출력
  - CTC 기반 End-to-End를 제안 [Graves and Jaitly, 2014]



Towards End-to-End Speech Recognition with Recurrent Neural Networks

Alex Graves GRAVES@CS.TORONTO.EDU
Google DeepMind, London, United Kingdom

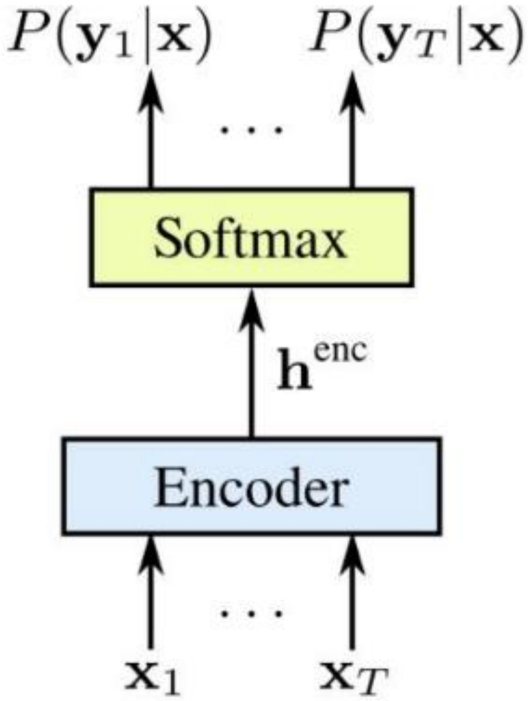
Navdeep Jaitly NDJAITLY@CS.TORONTO.EDU
Department of Computer Science, University of Toronto, Canada

Abstract

This paper presents a speech recognition system that directly transcribes audio data with text, without requiring an intermediate phonetic representation. The system is based on a combination

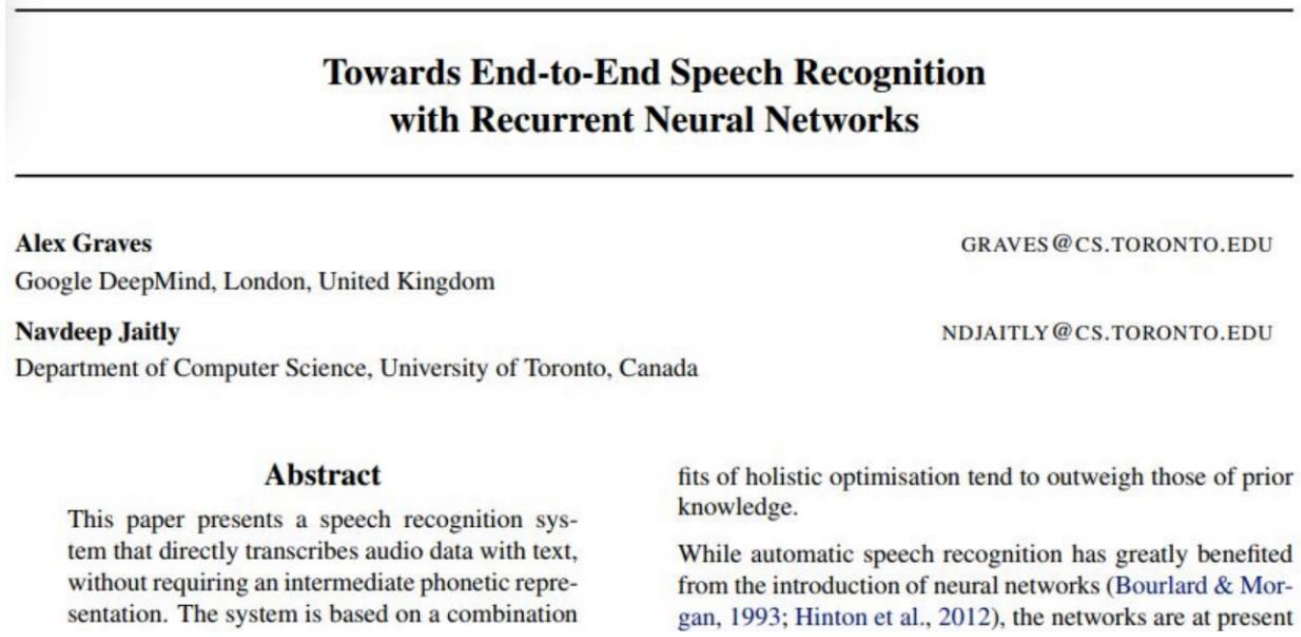
fits of holistic optimisation tend to outweigh those of prior knowledge.

While automatic speech recognition has greatly benefited from the introduction of neural networks (Bourlard & Morgan, 1993; Hinton et al., 2012), the networks are at present



End-to-End ASR

- LAS (Listen, Attend and Spell)
  - CTC와 비슷한 시기에 제안된 모델 [Google Brain Team]
  - 이후 ASR은 CTC와 LAS로 나뉜다고
  - CTC 기반 End-to-End를 제안 [Graves and Jaitly, 2014]



Towards End-to-End Speech Recognition with Recurrent Neural Networks

Alex Graves GRAVES@CS.TORONTO.EDU
Google DeepMind, London, United Kingdom

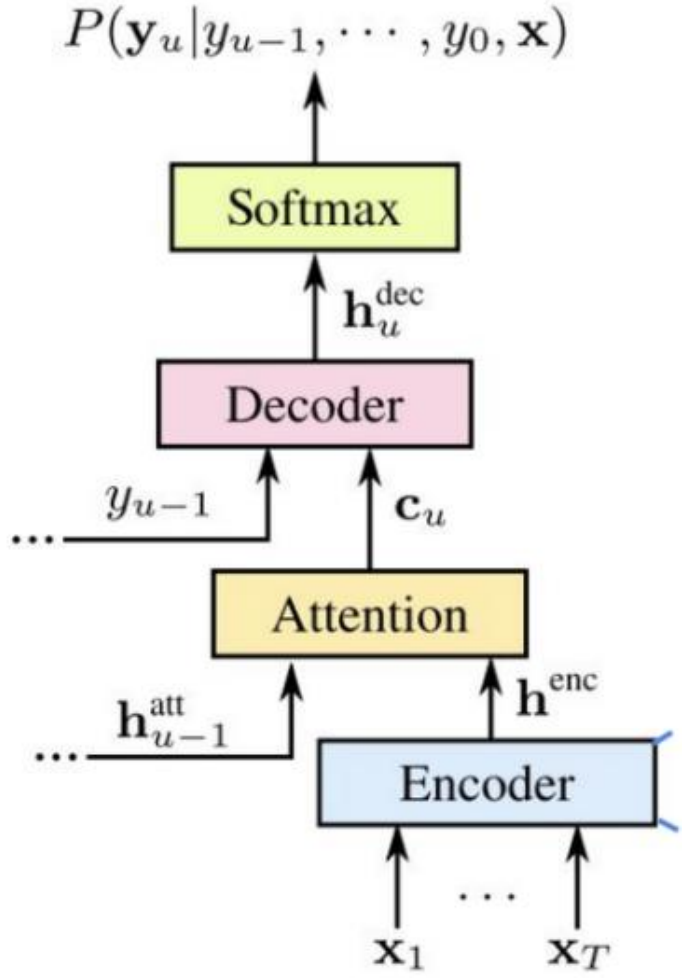
Navdeep Jaitly NDJAITLY@CS.TORONTO.EDU
Department of Computer Science, University of Toronto, Canada

Abstract

This paper presents a speech recognition system that directly transcribes audio data with text, without requiring an intermediate phonetic representation. The system is based on a combination

fits of holistic optimisation tend to outweigh those of prior knowledge.

While automatic speech recognition has greatly benefited from the introduction of neural networks (Bourlard & Morgan, 1993; Hinton et al., 2012), the networks are at present



LAS model

- Sequence to Sequence 기법
  - 가변 길이의 입출력 처리 가능
  - encoder RNN을 통해 가변 길이의 입력 시퀀스를 받아 고정 길이의 벡터로 변환
  - decoder RNN을 통해 변환된 벡터를 다시 가변 길이의 출력으로 변환
  - 학습 시 decoder에 입력 값으로 실제 정답을 사용하고, 추론 시 beam search를 이용하여 각 스텝마다 출력 값 후보를 결정

- Attention 기법
  - decoder RNN에서 매 스텝 결과를 결정할 때 마다 마지막 hidden state와 encoder의 모든 hidden state를 가지고 attention vector를 생성
  - 이 벡터는 enc에서 dec로 입력 시퀀스에 대한 정보를 효율적으로 전달하는 역할

LAS model

- Listener
  - 음향 데이터를 입력받는 encoder
  - 입력 시퀀스 x를 high level feature인 h = (h1, h2, ...) 로 변형

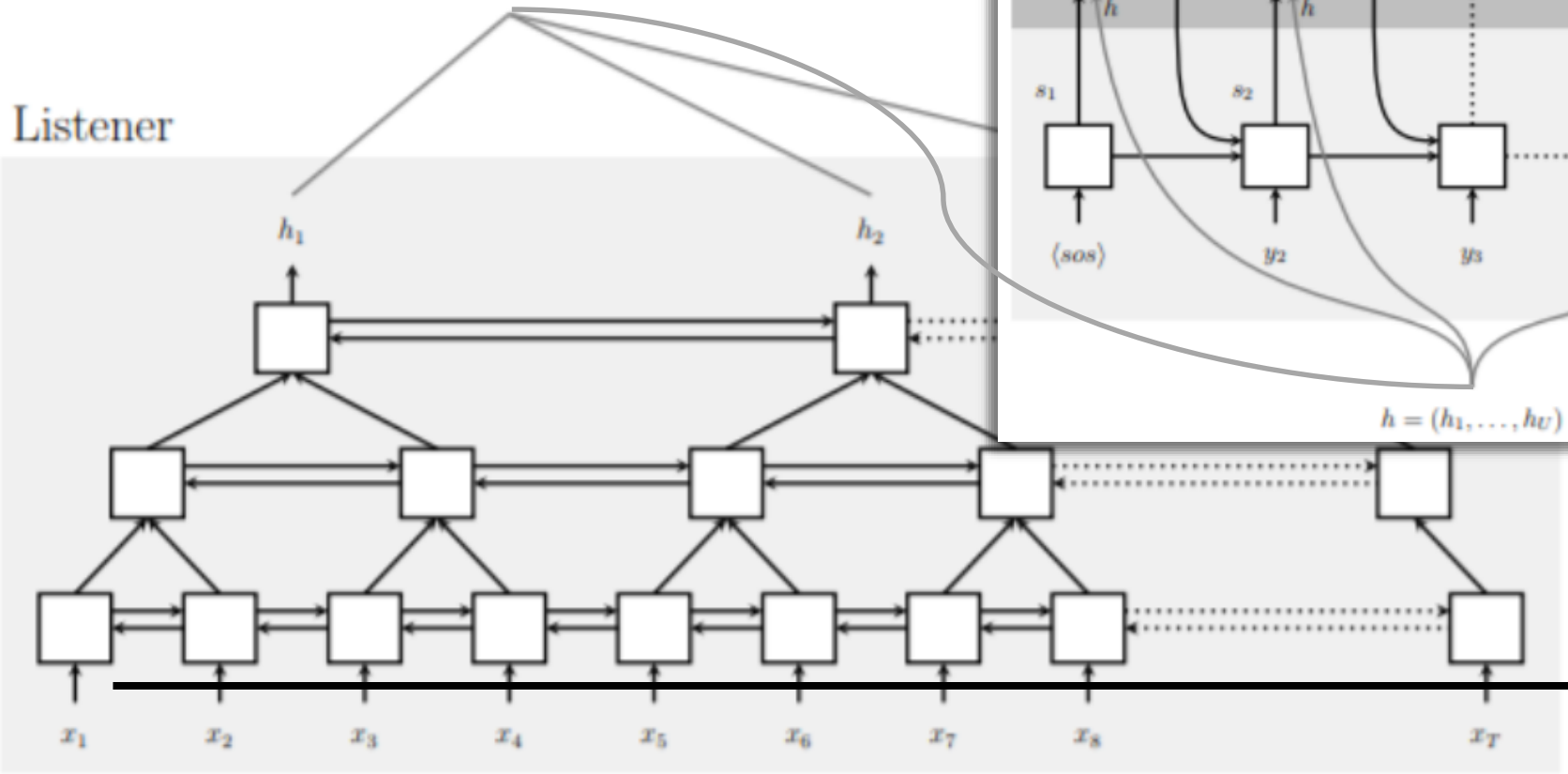
$$h = \text{Listen}(x)$$

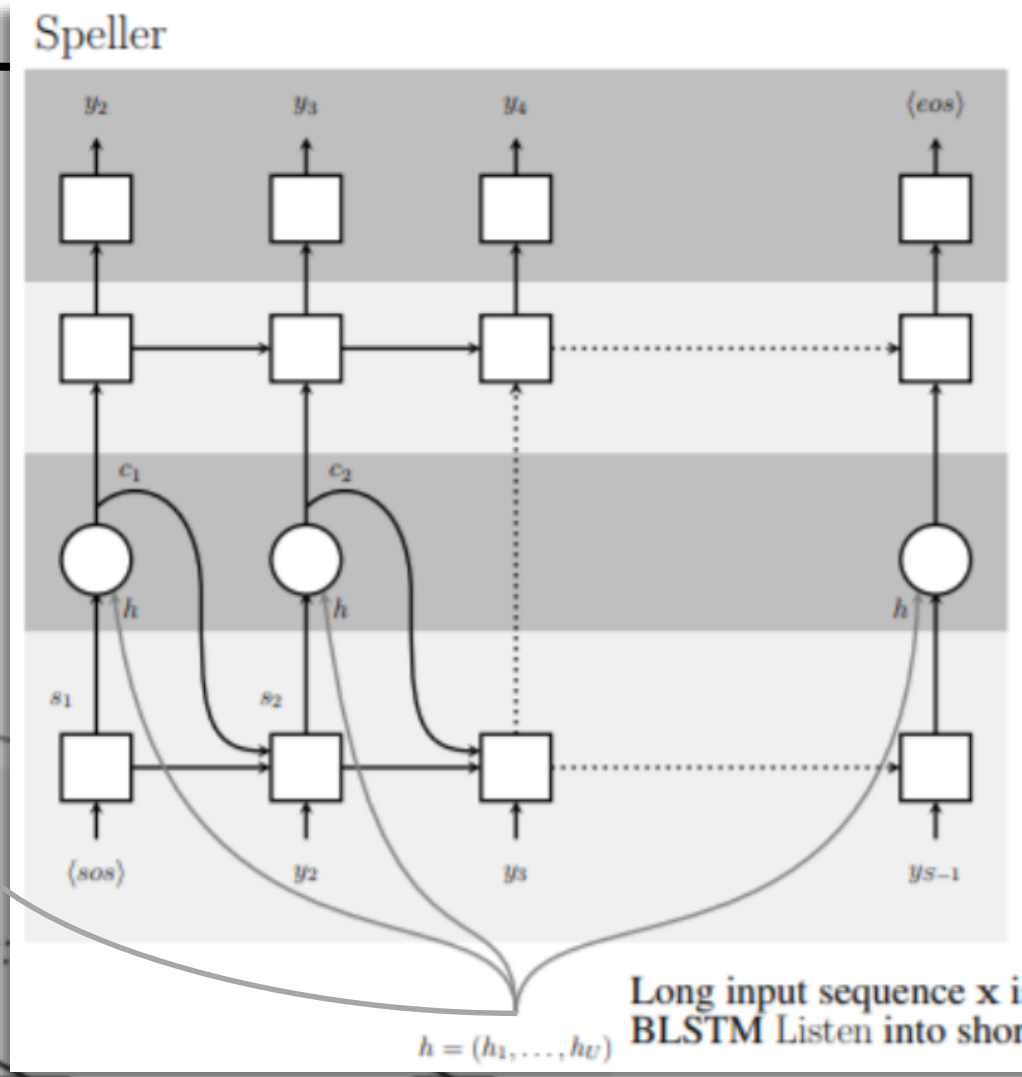
- Speller
  - Attention을 사용하여 출력하는 decoder
  - h를 가지고 출력 문자의 분포를 작성

$$P(y|x) = \text{AttendAndSpell}(h, y)$$

LAS model

$$h_i^j = \text{pBLSTM}(h_{i-1}^j, [h_{2i}^{j-1}, h_{2i+1}^{j-1}])$$





Grapheme characters y_i are modelled by the CharacterDistribution

AttentionContext creates context vector c_i from **h** and s_i

Long input sequence **x** is encoded with the pyramidal BLSTM Listen into shorter sequence **h**

Learning & Decoding

- End-to-End 학습 가능
  - Sequence to Sequence 모델에서 log 확률을 최대화하는 방법으로 학습한다

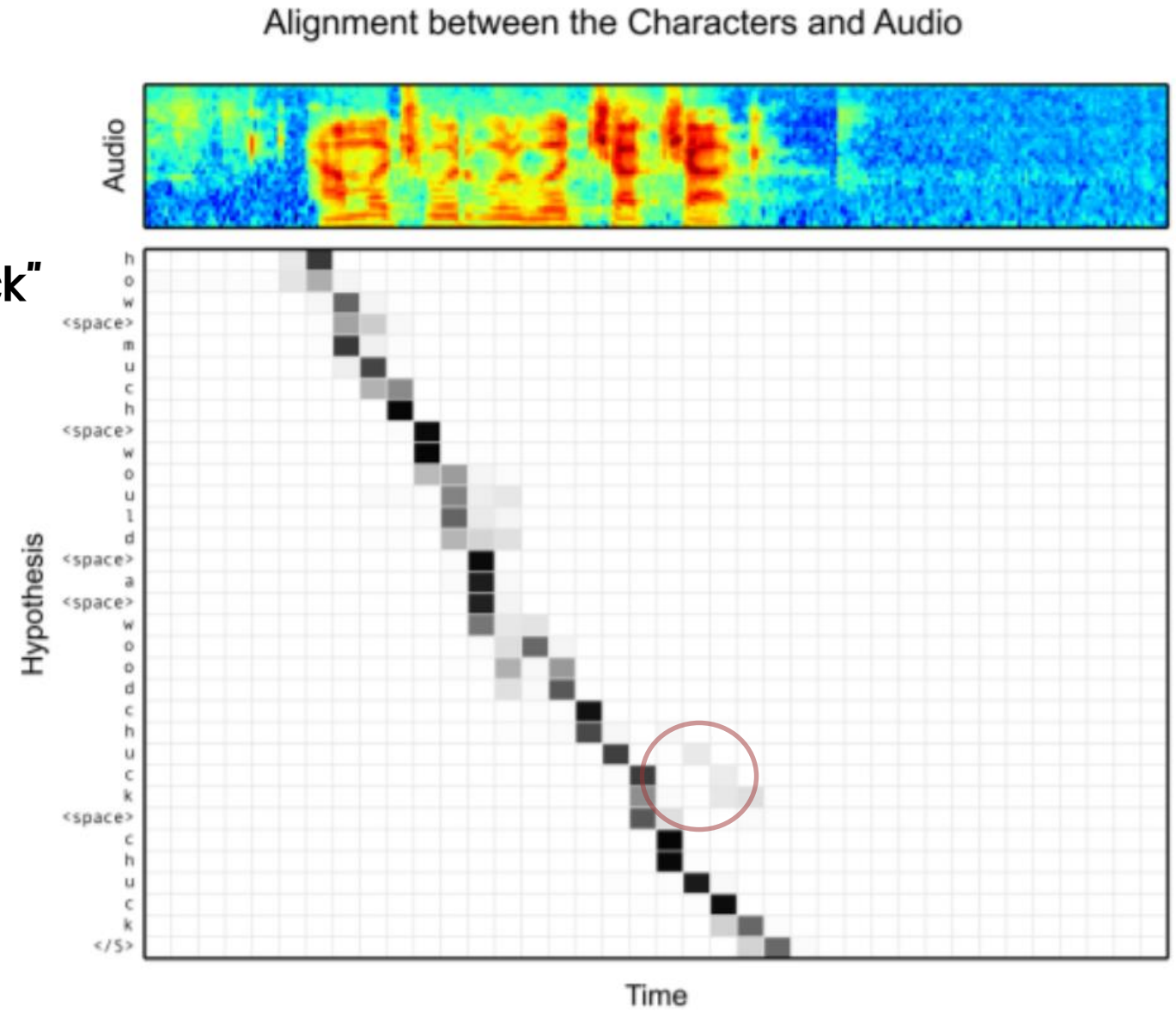
$$\max_{\theta} \sum_{i} \log P(y_i | x, y^*_{<i}; \theta)$$

  - y*는 Ground Truth (Teacher Forcing 기법)

- Beam Search를 통한 Decoding
  - 〈sos〉 토큰 하나만 있는 문장을 partial hypothesis로 두고, partial hypothesis에 있는 각각의 문장에 대하여 문자를 하나씩 추가해가며 최대 n개의 후보를 추려 partial hypothesis를 관리
  - 문장에서 〈eos〉 토큰이 추가되면 해당 문장을 partial hypothesis에서 제거하고 complete hypothesis로 관리
  - 최종적으로는 모든 남아 있는 hypothesis에서 가장 적합한 후보를 선정
  - 별도의 사전이 없더라도 단어를 잘 생성하는 것을 확인

LAS 모델 학습 결과

- "how much would a woodchuck chuck"
- 발화 시작 시점을 별도로 정의하지 않아도 스스로 학습하는 것을 확인
- "woodchuck"와 "chuck"의 발음이 동일하므로 이에 대한 Attention 매커니즘의 결과가 약간 혼란스러운 것 또한 확인 가능



LAS 모델 학습 결과

- 발화가 "aaa" 를 포함할 때 LAS의 beam search 후보군

- "aaa" 를 비롯하여 "triple a" 와 같은 유연한 답안을 추론하는 것을 확인

Table 2: Example 1: "triple a" vs. "aaa" spelling variants.

Beam	Text	Log Probability	WER
Truth	call aaa roadside assistance	-	-
1	call aaa roadside assistance	-0.5740	0.00
2	call triple a roadside assistance	-1.5399	50.00
3	call trip way roadside assistance	-3.5012	50.00
4	call xxx roadside assistance	-4.4375	25.00