Paper Review

SpeechBERT: Cross-Modal Pre-trained Language Model for End-to-end Spoken Question Answering(2019)

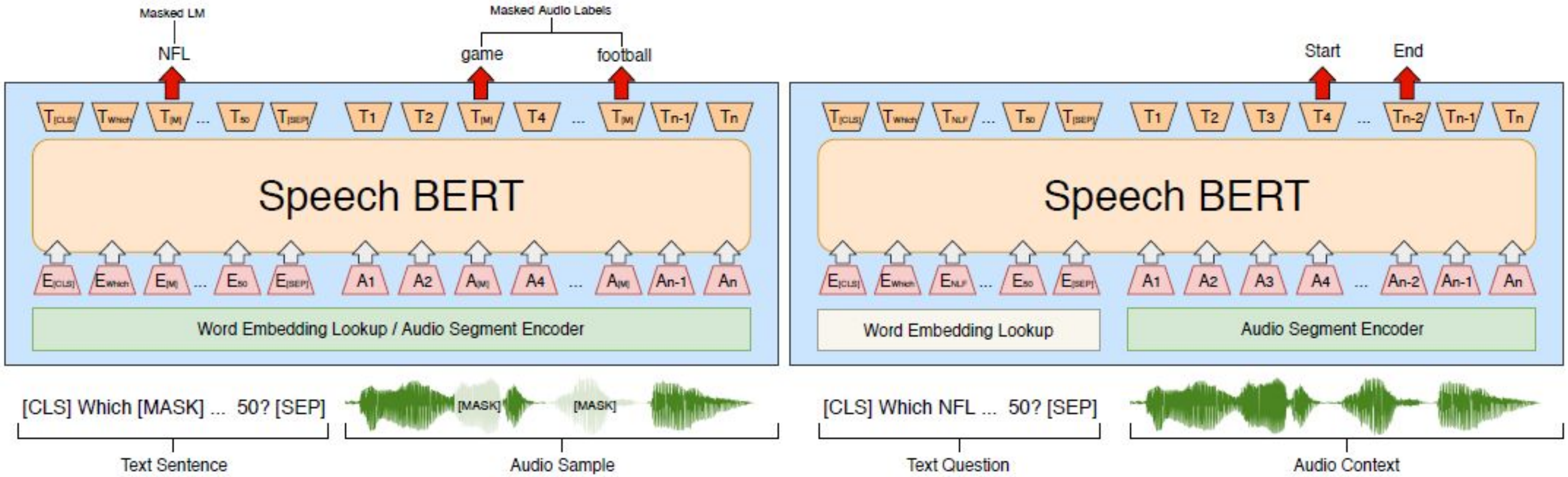
Yung-Sung Chuang Chi-Liang Liu Hung-Yi Lee

Presenter : Cheol H. Jeong
chloveyou@gmail.com

Introduction

- Large-scale self-supervised pre-trained language models appeared, such as BERT, GPT
- MC on spoken content is much more difficult than on text content, because speech recognition errors
- Most of the end-to-end models is still not as good as the performance of the corresponding natural language task
- SpeechBERT, a pre-trainable model of generic representation for speech and text tasks
- It combine the pre-trained language models with a audio encoder, our end-to-end model can circumvent the negative impact caused by cascading ASR and QA models that are trained separately.

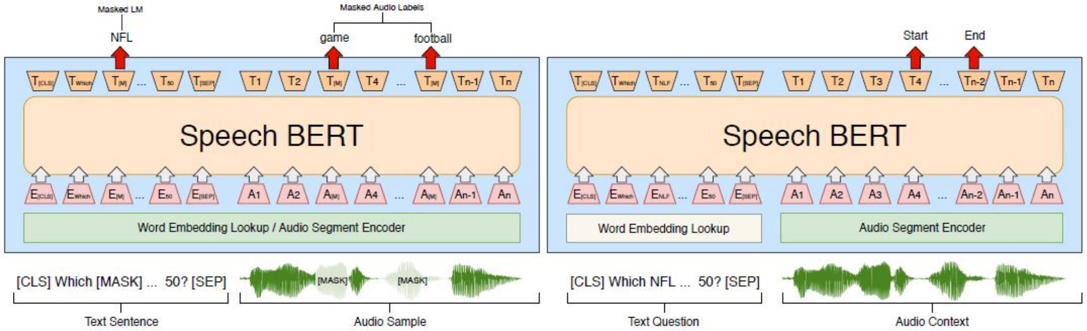
SpeechBERT


Masked LM
NFL
Masked Audio Labels
game
football
Start
End
T[CLS] TWhich T[M] ... T50 T[SEP] T1 T2 T[M] T4 ... T[M] Tn-1 Tn
Speech BERT
E[CLS] EWhich E[M] ... E50 E[SEP] A1 A2 A[M] A4 ... A[M] An-1 An
Word Embedding Lookup / Audio Segment Encoder
[CLS] Which [MASK] ... 50? [SEP]
[MASK]
[MASK]
Text Sentence
Audio Sample
T[CLS] TWhich TNLF ... T50 T[SEP] T1 T2 T3 T4 ... Tn-2 Tn-1 Tn
Speech BERT
E[CLS] EWhich ENLF ... E50 E[SEP] A1 A2 A3 A4 ... An-2 An-1 An
Word Embedding Lookup
Audio Segment Encoder
[CLS] Which NFL ... 50? [SEP]
Text Question
Audio Context

(a) *Pre-training Stage*

(b) *Fine-tuning Stage*

BERT for Text Pre-training



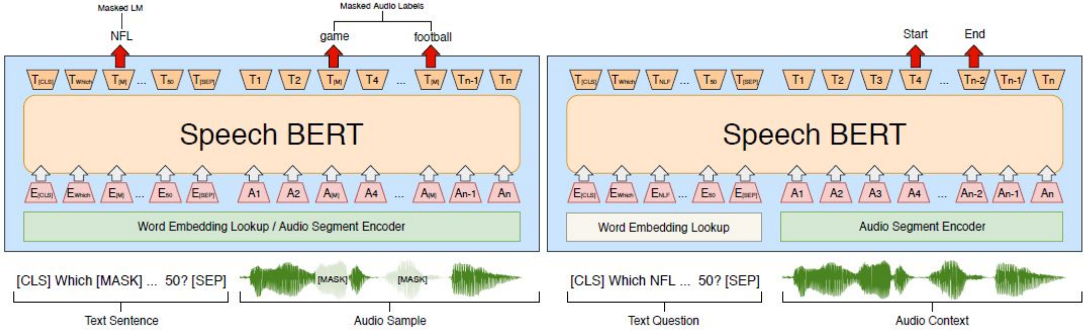
Masked LM
NFL
Masked Audio Labels
game
football
T[CLS] T_Which T[M] ... T50 T[SEP] T1 T2 T[M] T4 ... T[M] Tn-1 Tn
Speech BERT
E[CLS] E_Which E[M] ... E50 E[SEP] A1 A2 A[M] A4 ... A[M] An-1 An
Word Embedding Lookup / Audio Segment Encoder
[CLS] Which [MASK] ... 50? [SEP]
[MASK]
[MASK]
Text Sentence
Audio Sample
Start
End
T[CLS] T_Which T_NLF ... T50 T[SEP] T1 T2 T3 T4 ... Tn-2 Tn-1 Tn
Speech BERT
E[CLS] E_Which E_NLF ... E50 E[SEP] A1 A2 A3 A4 ... An-2 An-1 An
Word Embedding Lookup
Audio Segment Encoder
[CLS] Which NFL ... 50? [SEP]
Text Question
Audio Context

(a) *Pre-training Stage*

(b) *Fine-tuning Stage*

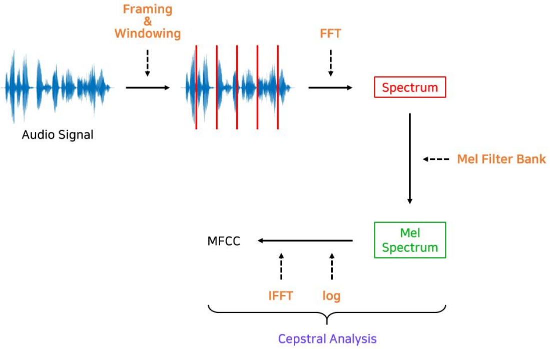
- Token embeddings + positional embeddings + sentence segment embedding = E'_{text}
- MLM, NSP
- 15% masked

Speech Segment Encoder Pre-training



(a) *Pre-training Stage*

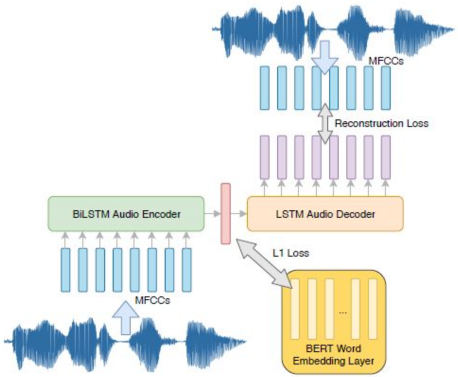
(b) *Fine-tuning Stage*

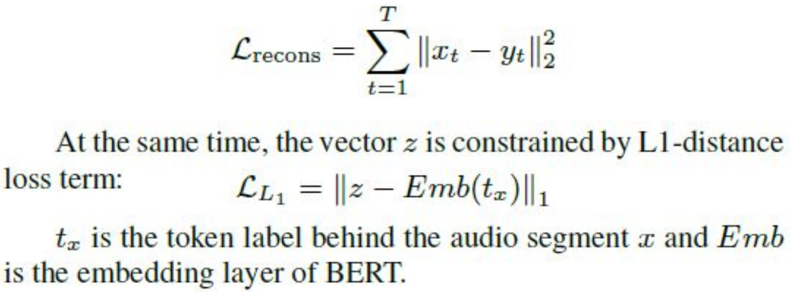


- Speech Segmentation https://brightwon.tistory.com/11
  - Training stage : Mel Frequency Cepstral Coefficients (MFCCs)* sequences according to the predefined boundaries from forced alignment of an off-the-shelf ASR model
  - Testing stage : cannot access the ground truth labels. ASR model to get word pseudo-label sequence

*소리의 특징을 추출하는 기법인데, 입력된 소리 전체를 대상으로 하는 것이 아니라, 일정 구간(Short time)식 나누어, 이 구간에 대한 스펙트럼을 분석하여 특징을 추출하는 기법이다

Speech Segment Encoder Pre-training





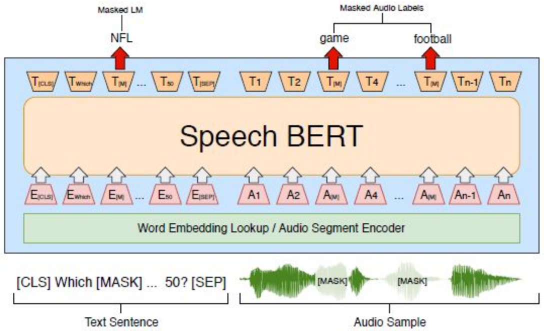
$$\mathcal{L}_{\text{recons}} = \sum_{t=1}^{T} \|x_t - y_t\|_2^2$$

At the same time, the vector z is constrained by L1-distance loss term: $$\mathcal{L}_{L_1} = \|z - Emb(t_x)\|_1$$

t_x is the token label behind the audio segment x and Emb is the embedding layer of BERT.

- Phonetic-Semantic Joint Embedding
  - RNN sequence-to-sequence autoencoder to encode the segments to obtain phonetic embeddings that captures the phonetic information of acoustic words

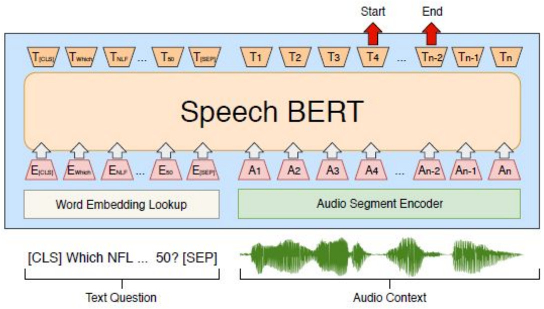
Speech & Text Corpus MLM Jointly Pre-training



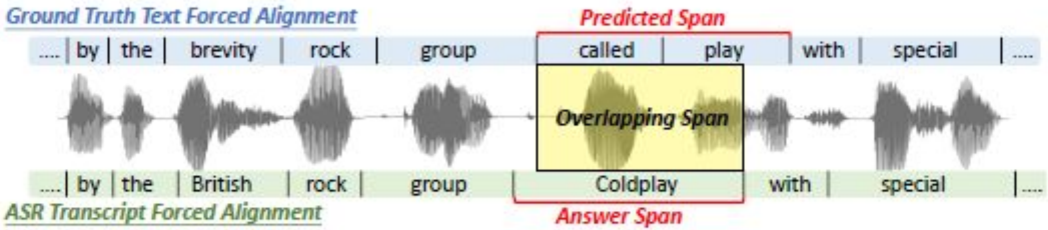
(a) *Pre-training Stage*

- Speech part was also 15% masked
- Speech encoder was frozen to speed up

Fine-tuning on Question Answering


Start
End
T[CLS] T_Which T_NLF ... T_50 T[SEP]
T1 T2 T3 T4 ... Tn-2 Tn-1 Tn
Speech BERT
E[CLS] E_Which E_NLF ... E_50 E[SEP]
A1 A2 A3 A4 ... An-2 An-1 An
Word Embedding Lookup
Audio Segment Encoder
[CLS] Which NFL ... 50? [SEP]
Text Question
Audio Context

(b) *Fine-tuning Stage*

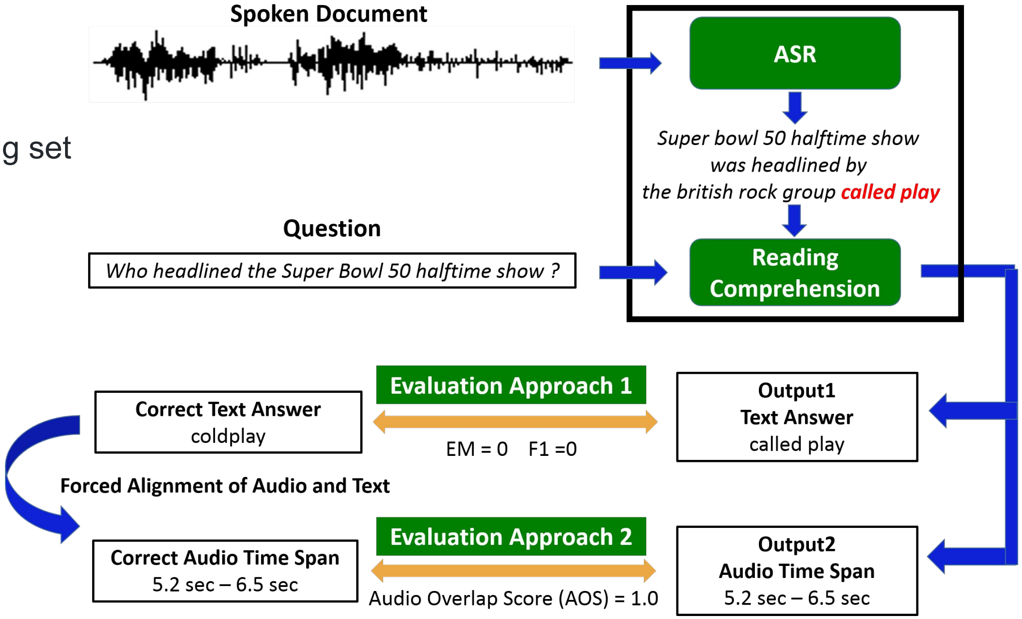

Ground Truth Text Forced Alignment
Predicted Span
.... | by | the | brevity | rock | group | called | play | with | special | ....
Overlapping Span
.... | by | the | British | rock | group | Coldplay | with | special | ....
ASR Transcript Forced Alignment
Answer Span

$$P = \frac{|Overlapping\ Span|}{|Answer\ Span|} \qquad R = \frac{|Overlapping\ Span|}{|Predicted\ Span|}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \qquad AOS = \frac{|Overlapping\ Span|}{|Answer\ Span\ \cup Predicted\ Span|}$$

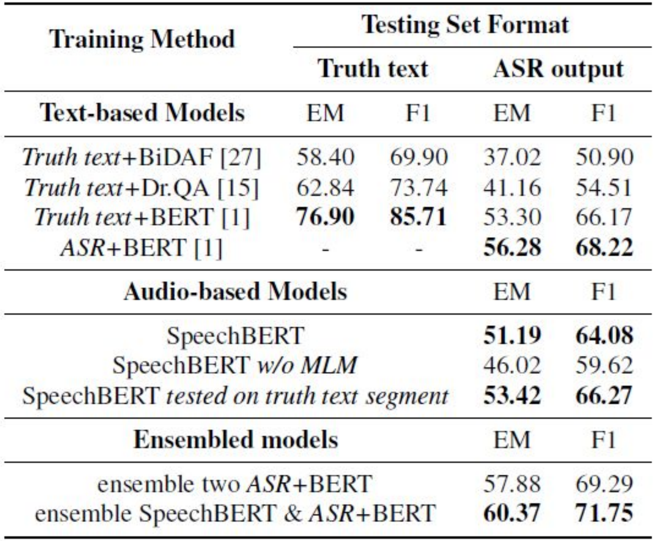
Experiment Setup

- Data
    - Spoken SQuAD
    - 37k question answer pairs as the training set
    - 5.4k as the testing set.



Result

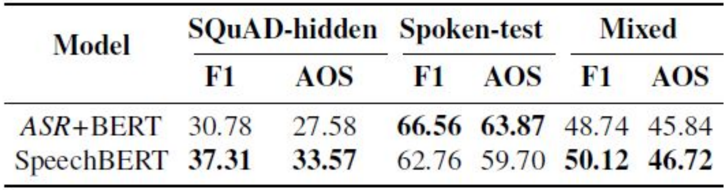
- For training method, we can use either text or audio format, and text can be from truth text or ASR output. For testing set format, we also can use either text (truth text/ASR output) or audio format
- SpeechBERT can learn complementary features that is not presented in a text-based BERT model. The ensemble result also achieves state-of-the-art result on Spoken SQuAD



Training Method	**Testing Set Format**			
	Truth text		**ASR output**	
Text-based Models	EM	F1	EM	F1
Truth text+BiDAF [27]	58.40	69.90	37.02	50.90
Truth text+Dr.QA [15]	62.84	73.74	41.16	54.51
Truth text+BERT [1]	**76.90**	**85.71**	53.30	66.17
ASR+BERT [1]	-	-	**56.28**	**68.22**
Audio-based Models			EM	F1
SpeechBERT			**51.19**	**64.08**
SpeechBERT *w/o MLM*			46.02	59.62
SpeechBERT *tested on truth text segment*			**53.42**	**66.27**
Ensembled models			EM	F1
ensemble two *ASR*+BERT			57.88	69.29
ensemble SpeechBERT & *ASR*+BERT			**60.37**	**71.75**

Result : Ablation Studies

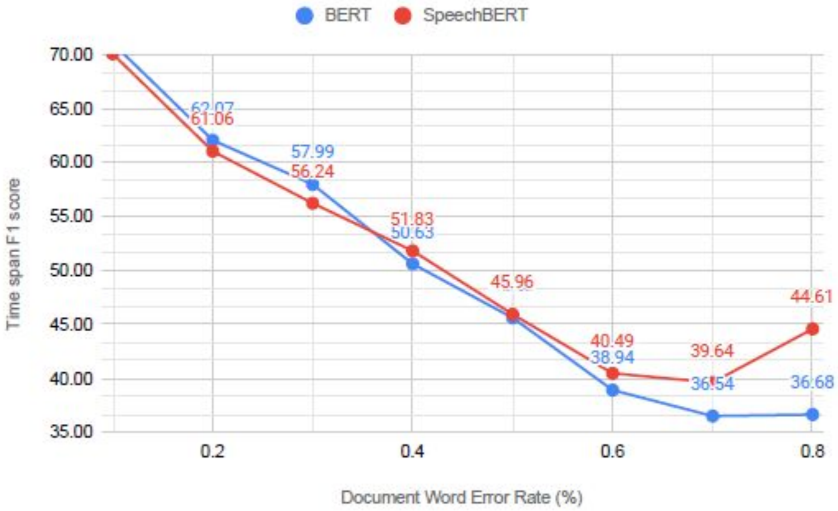
- SQuAD-hidden : questions that cannot find answers in the ASR transcripts
- Spoken-test : questions can find answers in ASR
- Advantages on hidden questions



Model	SQuAD-hidden		Spoken-test		Mixed	
	F1	AOS	F1	AOS	F1	AOS
ASR+BERT	30.78	27.58	**66.56**	**63.87**	48.74	45.84
SpeechBERT	**37.31**	**33.57**	62.76	59.70	**50.12**	**46.72**

Result : Error Analysis

- Advantages compared to BERT when a higher recognition word error rate (WER) occurs



Discussions and future works

- Word boundaries
  - Word boundaries can be provided by the end-to-end model itself
  - It is challenge to use frame-level speech features for the SQA model which needs a pointer network to predict positions directly on very long frames
  - One possible way is to simply divide audios by the voice intensity
  - Alternatively, learn segmentation strategies by learning machine translation

Conclusions

- we proposed an end-to-end model for spoken question answering
- which can perform close to the results of cascading ASR and QA models and even outperform it on a more difficult dataset
- It is a stepping stone towards understanding speech content directly from speech information to solve QA task and mitigating the ASR error propagating problem