Chapter 1

음성인식 연구 동향 및 문제 정의

김지환

서강대학교 컴퓨터공학과

Table of contents

1.1 음성인식 이란?

❖ 마이크를 통해 입력 받은 음성(speech)이 주어졌을 때, 확률이 가장 높은 문장(단어의 열)을 출력

"안녕하세요"

$$\arg_W \max P(W|X)$$

- $W = \{w_1, w_2, \cdots, w_U\}$: U개의 단어 시퀀스
- $X = \{x_1, x_2, \cdots, x_T\}$: 음성 시퀀스

1.1 E2E 관점의 음성인식 문제정의

◆ 음성인식을 입력end (음성)에서 출력end (문장)으로의 변환의 문제로 본다면

❖ 음성인식 문제는

- $x_1 \cdots x_T$ (Continuous vector space에서의 13차 벡터 T개의 시퀀스)에서 $w_1 \cdots w_U$ (V개의 서로 다른 값을 가지는 discrete symbol U개의 시퀀스)로의 번역 문제로 재정의 할 수 있다

◆ 이상적인 E2E 시스템 구현은 불가능

- 이상적인 E2E 시스템: 전체 시스템을 블랙박스로 보고 데이터만 주면 알아서 시스템이 학습하는 방식
- 무한개의 입력 시퀀스에서 무한개의 출력 시퀀스로 매핑하는 시스템은 구현이 불가능하다

1.1 E2E 음성인식 구현 시 입출력 복잡도

음성인식 시스템의 가능한 입력 개수 분석

- 가정: 입력 길이 1초, 44.1K, 샘플 당 2byte 사용
- 저장에 $1 \times 44100 \times 2 = 88{,}200$ byte 필요
- 가능한 입력의 개수: $2^{88200 \times 8}$
- 입력 길이의 제한이 없으므로 가능한 입력의 개수는 무한대이다

음성인식 시스템의 가능한 출력 개수 분석

- 어휘를 구성하는 단어의 수(V) : 무한대 (지명, 인명 등 제한이 없으며 신조어가 계속해서 생성된다)
- 연속음성인식에서는 입력 파형만을 가지고 몇 개의 단어로 구성된 문장인지 알 수 없다

1.1 DNN-WFST에서의 문제 정의

$$arg_W max P(W|O)$$

- 기호 설명
  - W: $w_1 \ldots w_N$
    - N개의 단어들로 이루어진 문장
  - O: $o_1 \ldots o_T$
    - T개의 윈도우에서 각 윈도우로부터 나온 13차 vector의 sequence.

- 가능한 O의 개수가 무한대이기 때문에 $P(W \mid O)$를 직접 구할 수 없음

1.1 DNN-WFST에서의 문제 정의

- Bayesian rule을 적용하여 변환

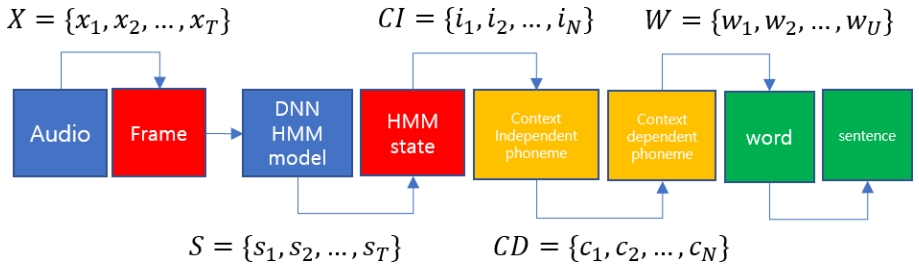
$$arg_W max P(W|O) = arg_w max \frac{P(O|W)P(W)}{P(O)}$$

  - P(O): 13 * T 벡터 공간에서의 한 점의 확률
  - 이 확률을 모두 동일하다고 가정하면 $arg_w max$를 찾는 문제이기 때문에 P(O)를 생략 가능

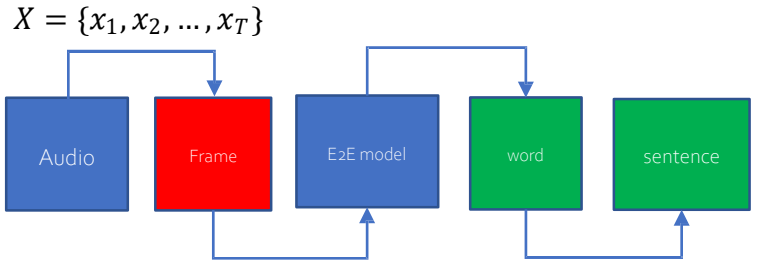
$$\underbrace{arg_w max}_{\text{디코딩}} \underbrace{P(O|W)}_{\text{음향 모델}} \underbrace{P(W)}_{\text{언어 모델}}$$

- 음성인식 시스템 구성에서의 핵심
  - 음향 모델P(O | W)
  - 언어 모델P(W)
  - 디코딩 네트워크($arg_w max$)
  - 어휘(인식 가능한 단어 set) 의 구현이다.

1.2 주요 음성인식 모델



DNN-WFST **기반 음성인식 파이프라인**



End-to-end **기반 음성인식 파이프라인**

- DNN-WFST

- Kaldi(음성인식 주요 tool)의 기반

- E2E에 대응되는 기술로 서술되나, Frame 단위까지 (출력 end가 HMM에서의 state)의 E2E임

- E2E (출력 end는 주로 grapheme)

- CTC
- RNN-T
- Attention
- Transformer

1.2 E2E방법의 장점 및 주요 검토 항목

장점

- SOTA(State-of-the-art)를 보임 (transformer)
- 음성파일과 이에 대응되는 transcription만으로 학습
- 전혀 모르는 언어에 대해서도 음성인식기 제작이 가능

단점

- 외부 지식을 실시간 반영할 수 있는 방법이 없음 (예: 실시간 검색어(고유명사 많음))
- 대용량 텍스트 코퍼스를 음성인식기에 직접 반영할 수 있는 방법이 없음
- 복잡한 구조에 파라미터가 많고, computation power를 많이 사용하며, ML 기반으로 학습이 이루어짐 (입력열과는 다른 길이를 가지는 출력열에 대한 답만 가지고 있음)
- 재현 실험이 되지 않는 경우가 많음 (모델 초기값에 의해 성능이 바뀔 수 있음)

1.2 E2E방법의 장점 및 주요 검토 항목

◆ 주요검토항목

- DNN-WFST 대비 성능이 좋은가?
- 재현 실험은 잘 이루어 지는가?
- 필요한 computation power는?
- 출력 단위 정의

1.2 주요 E2E 방법

End-to-end models

- Connectionist temporal classification (CTC) [Graves, 2006]
  - 알파벳과 음성 정보만으로 단일 모델을 구성할 수 있는, 최초로 제안된 end-to-end 음성인식 모델
- RNN-transducer (RNN-T) [Graves, 2012]
  - CTC에 언어 정보를 학습할 수 있는 RNN 기반의 모델을 추가하여 성능을 개선시킨 모델
- Attention 기반 seq2seq [Chan, 2016]
  - 음성인식을 sequence의 번역으로 해석해서 attention mechanism을 적용한 모델
- Transformer [Vaswani, 2017]
  - Multi-head self-attention을 사용하여 RNN 없이 sequence 모델링을 구현한 모델

1.2 주요 E2E 방법

◆ Librispeech corpus 대상 end-to-end 음성인식 모델 성능 비교

System	WER(%)	WER with 2nd-pass LM(%)	Streaming **가능 여부**	**성능 재현 여부**
DNN-WFST [Han, 2019]	2.93	2.20	O	O
CTC [Li, 2019]	3.86	2.95	O	O
RNN-T [Zeyer, 2021]	3.30	2.36	O	O
Attention-based seq2seq [Park, 2019]	2.80	2.50	X	O
Transformer (seq2seq) [Karita, 2019]	2.20	2.10	O	X
Transformer (RNN-T) [Anmol ,2020]	2.10	1.9	O	X

1.2 주요 E2E 방법

◆ 대용량 corpus 대상 end-to-end 음성인식 모델 성능

System	Training data	Test data	WER(%)	WER with 2nd-pass LM(%)	비고
CTC [Pratap, 2020]	Multilingual Librispeech (MLS) English (44,659 hours)	Librispeech test-clean	2.94	1.83	
CTC [Narayanan, 2018]	Youtube (117,000 hours)	Youtube	N/A	15.9	
CNN+Transformer+RNN-T [Yu, 2021]	Multidomain Data (413,000 hours)	Google voice search	5.2	N/A	Google Voice Search, Farfield Speech, YouTube and Meetings
RNN-T [Yu, 2021]			5.1	N/A	
RNN-T+seq2seq [Sainath, 2020]			6.4	N/A	