* 음성인식의 시각화 분석

: 음성인식 모델은 주로 음향 모델과 언어 모델을 사용하여 구현된다.

이때 음향 모델을 학습시키기 위해서는 음성 데이터에 대한 정답 텍스트뿐만 아니라 음성인식에 사용되는 단어의 발음 사전과 프레임 단위의 음소 정답 데이터가 필요하다.

이 때문에 모델을 훈련하기 위해서는 먼저 프레임 단위의 정답을 생성하는 등의 여러 과정이 필요하다. 그리고 음향 모델과 별도의 텍스트 데이터로 훈련한 언어 모델을 적용하여야 한다.

이러한 불편함을 해결하기 위해서 최근에는 하나의 통합 신경망 모델로 이루어진 종단간(end-to-end) 음성인식 모델이 연구되고 있고, 이 모델은 훈련에 여러 과정이 필요 없고 모델의 구조를 이해하기 쉽다는 장점이 있다. 하지만, 인식이 내부적으로 어떤 과정을 거쳐 이루어지는지 알기 어렵다.

\* 서론

: 음성인식 기술은 단순히 음성 명령으로 기기를 조종하는 것에서 나아가 인공지능이 사람과 음성 대화를 나누기 위하여 필수적인 모듈로써 그 중요성이 높아지고 있다.

음성인식이 이루어지는 과정은 먼저 기기가 사람의 음성을 입력으로 받고, 기기 내부에서 이미 훈련되어 있는 음성인식 모델을 사용하여 입력 음성을 인식하고 텍스트로 변환한다. 이렇게 변환된 텍스트를 최종 출력으로 가지게 된다.

음성인식 모델의 가장 널리 사용되는 구조인 특징 추출(feature extraction), 음향 모델(acoustic model), 언어 모델(language model)을 거쳐 최종적으로 단어열로 인식하게 된다.

\* 종단간(end-to-end) 음성 인식

: 입력 음성으로부터 하나의 통합 신경망을 거쳐 문자열 혹은 단어열을 인식하는 음성인식 방법을 뜻한다. 앞서 말했던 전통적인 음성 인식에서의 단점을 해결한 간단한 구조의 모델이다.

대표적인 종단간 음성인식 모델은 connectionist temporal classification(CTC) 방식이다.

이 방법은 재귀신경망을 이용해 음성 특징으로부터 바로 문자열을 추론해낸다. HMM과 유사하게 매 프레임 마다 문자 사후 확률을 추정하고, 이렇게 추정된 문자열이 최적의 경로를 갖도록 작동한다.

이러한 연구들에서 하이브리드 음성 인식 모델을 구성하는데 사용한 툴킷(toolkit)d이 Kaldi이다.

\* 음성인식기에 DNN(deep neural network)을 접목하면 성능이 많이 향상되는 결과를 초래한다.

이런 음성인식의 기술들이 많이 통합되었는데 사용된 오픈소스가 Kaldi 이다.

이 툴은 Acoustic Model(음향 모델)을 트레이닝하는 역할을 한다고 생각하면 된다.

어떻게 보면 음성인식기에 최적화된 TensorFlow라고 표현할 수도 있다.

이 툴을 사용하면 기본적으로 Acoustic Model을 트레이닝하는 것부터 decoding까지 해보는, 음성인식기를 구축하는데 필요한 모든 과정을 경험해볼 수 있다.

하지만, 한국어는 Language Model 자체가 영어랑 다르기 때문에, 그 Language Model을 어떻게 만드냐에 따라 음성인식의 성능이 차이를 보이게 된다.

일단, 영어와 달리 한국어는 교착어에 속한다. 즉, 어근과 접사에 의해 단어의 기능이 결정되는 언어의 형태란 말이다. 여기서 예를 들어서, 영어의 경우 크롤링을 해서 단어를 수집할 때 띄어쓰기를 기준으로 교유한 단어의 수가 일정 수준에서 포화 즉, 수렴된다. 하지만, 한국어의 경우 ‘가다’라는 하나의 동사도 ‘가겠습니다.’, ‘갈겁니다.’, ‘갔습니다.’처럼 의미가 다 바뀔 수 있다는 어려움이 있다.

그렇기 때문에 똑 같은 테스트를 진행했을 때 한국어의 경우 고유 단어의 수가 기하급수적으로 증가하게 된다. 그래서 음성인식

기를 만들 때 일정 어휘의 범위 안에서 탐색을 하게 해야하는데, 이 ‘일정 어휘’를 잡기가 힘들어 진다.

그래서 우리는 한국어 분석에서 ‘형태소 분석’을 실시한다.

이런 한계를 뛰어넘기 위해 Morfessor라는 툴을 활용한다.

\* Morfessor

: 이 툴은 한국어라는 교착어를 입력하면, 교착어로 이루어진 텍스트 안에서 Morpheme(형태소)를 찾아내고, 최소의 Morpheme로 corpus(말뭉치)를 표현하기 위해 단어를 어떻게 쪼개야 하는지 비지도 학습을 통해 자동으로 찾아내는 것이다.

경험의 누적이 아닌 Data-driven 방식의 형태소 분석기이기 때문에 일정한 수준의 고유 단위 범위가 설정되어 있고, 영어와 비슷한 환경에서 Language Model 구축이 가능하다. 이 과정이 해결이 된다면, 이후 파트는 영어와 동일하다고 볼 수 있다.

\* DNN(Deep Neural Network = 심층 신경망)

: 입력층(input layer)과 출력층(output layer) 사이에 여러 개의 은닉층(hidden layer)들로 이루어진 인공신경망이다. 심층 신경망은 일반적인 인공신경망과 마찬가지로 복잡한 비선형 관계들을 모델링 할 수 있다.

**Acoustic Model (음향 모델)**

- HMM/GMM => HMM/DNN

- 뜻은 모르더라도 소리를 정확하게 인식 ('귀' 튜닝)

**Language Model (언어 모델)**

- Statistical N-gram => RNN/LSTM-LM

- Morpheme(형태소) Analysis

- 인식된 소리를 언어를 가지고 확률적인 모델을 만드는 것 ('뇌'를 통해 문법을 알게 되는 것)

**Training (학습)**

: Sentences -> Words -> Phones -> Tri-phone HMMs -> GMM -> Feature vectors -> Acoustic signal

**Recognition (인식)**

: Acoustic signal -> Feature vectors -> GMM -> Tri-phone HMMs -> Phones -> Words -> Sentences

Acoustic Model에 DNN을 적용하게 되면서 인식률이 높아졌다 (HMM/GMM => HMM/DNN)

\* Project : Zeroth

- Kaldi-based Korean ASR open-source project

- Offcial recipe now (<https://github.com/kaldisr/kaldi/tree/master/egs/zeroth_korean/s5>)

- License: Apache 2.0

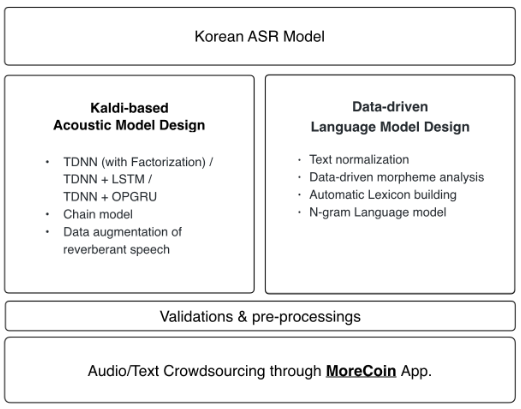
- forum: <https://groups.google.com/forum/#!forum/zeroth-help>

: zeroth는 Kaldi 툴킷을 사용해 시행되는 한국 음성 인식의 오픈소스 project이다.

이 프로젝트는 Atlas 실험실인 Language Al platform에서 개발되었고, 기업들이 그들의 B2C(기업과 소비자들간의) 의사소통에 정보들을 추가할 수 있게 한다.

공식적인 한국어 Kaldi 방안을 소개함으로써, Zeroth 계획은 한국어 음성 인식을 모든 사람들에게 더 넓게 접근할 수 있게 하는 것을 목표로 한다.

이름인 Zeroth(0th) 처럼, 이 프로젝트의 목표는 어느 누구든 새로운 음성 인식을 사용한 제품들과 서비스들을 구축할 수 있는 시작점과 기초적인 부분이 되는 것이다.



1. Audio Data

**- 16 July 2018: 95.7 hours (46,347 utterances, 181 speakers, 27,330 uniq. sentences)**

**-** 9 April 2018: 76.6 hours (35,139 utterances, 137 speakers, 16,472 uniq. sentences)

**-** 3 February 2018: 51.6 hours transcribed Korean audio for training data

(22,263 utterances, 105 speakers, 3000 sentences)

**-** License: [CC BY 4.0] (<https://creativecommons.org/licenses/by/4.0/>)

**-** Now 51.6 hour audio and LM data is available at [OpenSLR](http://www.openslr.org/40/)

**-** Audio crowdsource from MoreCoin is growing. 70 hours of open-source audio data-base will be opened around April 2018. You can donate yours though the voice recording app [MoreCoin (Android)] (<https://play.google.com/store/apps/details?id=com.goodatlas.morecoin>).

- We offer a voice recording app [MoreCoin (Android)] (<https://play.google.com/store/apps/details?id=com.goodatlas.morecoin>) that you can use to participate in building our open source database of Korean training data.

2. Acoustic Model(음향 모델)

: 현재 Zeroth 프로젝트 음향 모델에는 아래와 같은 최신 Kaldi recipe가 적용되어 있다.

- TDNN (with Factorization) / TDNN + LSTM / TDNN + OPGRU

- Chain model

- Data argumentation of reverberant speech

3. Language Model(언어 모델) & Lexicon

\* Lexicon : 단어책의 의미. 즉, 어떤 언어에 존재하는 모든 단어와 어구로 사전을 뜻한다.

Zeroth의 언어 모델과 발음 사전은 접근 가능한 종단간(end-to-end) 데이터를 사용한다.

우리의 오픈소스 음향 데이터베이스에의 어떤 기여이든지 자동적으로 최근 언어 모델과 발음 사전에 포함될것이다.

- custom 언어 모델과 발음 사전 창조하기

: (<https://github.com/goodatlas/zeroth/blob/master/s5/data/local/lm/README.md>).

- Corpus (Corpus) : 말뭉치

Training sentences(훈련된 문장의 수) : 109,037,699

Test sentences(테스트 문장의 수) : 12,115,208

Total(전체) : 121,152,907

- Phonetic Dictionary : 발음사전(Lexicon)

Unique words(고유한 단어의 수) : 30,064,143

Unique words with the highest 98% frequency(상위 98% 빈도 수를 보이는 고유한 단어의 수) : 8,069,252

Unique morphemes(고유한 형태소의 수) : 465,253

Size of phonetic dictionary considering pronunciation diversity  
(발음 다양성을 고려한 발음 사전의 크기) : 686,839

- Language Model : 언어 모델

Perplexity test 3-gram: ppl= 221.2969 (12,115,208 sentences, 194,940,635 words, 0 OOVs)

Perplexity test 4-gram: ppl= 187.2058 (12,115,208 sentences, 194,940,635 words, 0 OOVs)

* LM ( Language Model )

: 15GB의 txt 데이터를 만들어야 한다.

예를 들어서, 책을 사용하여 LM을 만들어도 되고,

근래에 자주 쓰이는 단어들을 많이 사용하는 뉴스 자막,

근래에 이슈화가 되는 주제들을 다룬 블로그, 카페 등에서 최대한 얻을 수 있는

많은 TEXT data를 가져와서 LM을 만들어야 한다.