데이터로 표현하는 세상

using python programming

1. 데이터와 처리과정

Professor Yoonjoong Kim
Computer engineering department, Hanbat national university
yjkim@hanbat.ac.kr

내용

1. 자료(data)의 개념
2. 문제 해결을 위한 자료의 생산,수집,처리과정
    1. 문제파악
    2. 데이터분석 및 이해
        1. 기초
        2. 전처리
        3. 데이터의 이해
    3. 특징값추출
    4. 모델의 선정/개발 및 분석
    5. 결과 정리
3. 감정인식기 개발의 예

1. 자료(data)의 개념

- 데이터로 표현하는 세상(Computing the world)
  - 개념
    - 디지털기술에 기반하고 있는 정보화세상은 데이터를 생산, 수집 및 처리하여 유용한 정보를 추론하고 이를 이용하는 능력이 주도하는 세상이다.
  - 따라서
    - 자료의 생산, 수집 및 처리의 체계화하여야 하고
      - 디지털 기술의 발달로 우리 주변에 다양한 자료들이 산재되어 있다. 따라서 다양한 자료를 생산, 수집하고 처리하는 과정을 체계화고
    - 정보생산을 위한 모델이 필요하다.
      - 정량화된 데이터와 추상적인 정보들 간의 상호관계를 분석하여 유용한 정보를 생산하기 위한 모델의 설계 및 구현
    - 이와 같이 자료의 생산, 수집 및 처리의 체계화하고 유용한 정보생산을 위한 모델를 이용하기 위하여 필요한 이론과 기법을 습득하고 응용능력을 배양한다.

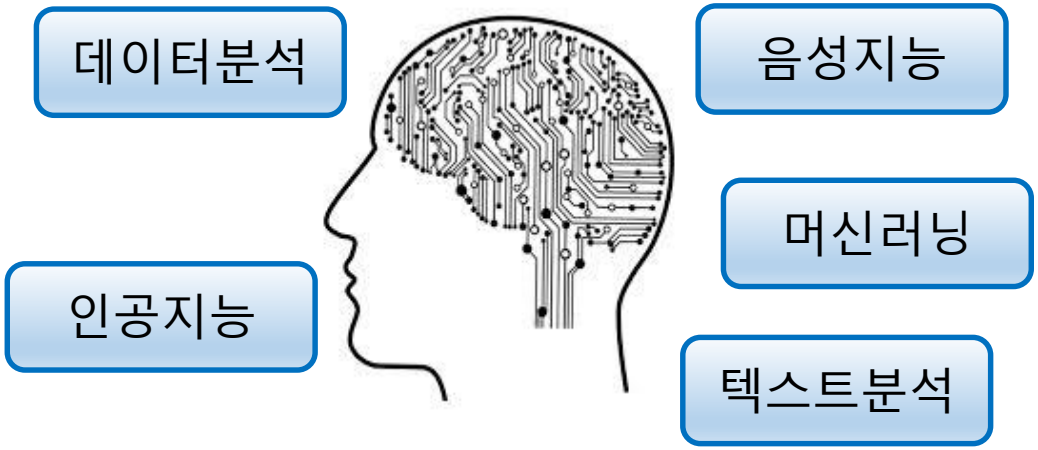
1. 자료(data)의 개념

- 자료(data)
  - 라틴어 단어 Datum의 복수형인 Data에서 유래했으며 라틴어에서 Datum의 뜻은 "present/gift, that which is given, debit" 이다. 기본적으로는 복수형 취급을 하나 가끔 하나의 고유명사화가 되어서 단수로 취급하는 경우도 있다.
  - 종류
    - 이론을 세우는 데 기초가 되는 사실. 또는 바탕이 되는 자료.
    - 관찰이나 실험, 조사로 얻은 사실이나 자료.
    - 컴퓨터가 처리할 수 있는 문자, 숫자, 소리, 그림 따위의 형태로 된 자료.
- 정보(information)
  - 자료를 가공하여 얻어지는 일종의 고수준 자료
  - 통계학 관점에서 자료와 정보의 예
    - 자료 : 설문조사를 통해 우리나라 18세 청소년의 **신장**
    - 정보 : 2016년 현재 우리나라 18세 청소년의 **신장의 평균** 173.8센티미터이며 **신장의 표준편차**는 5.33센티미터이다
  - 개인정보보호법의 예
    - 개인자료"(personal data)를 "개인정보"(personal information)와 동일하게 취급

https://namu.wiki/w/%EB%8D%B0%EC%9D%B4%ED%84%B0

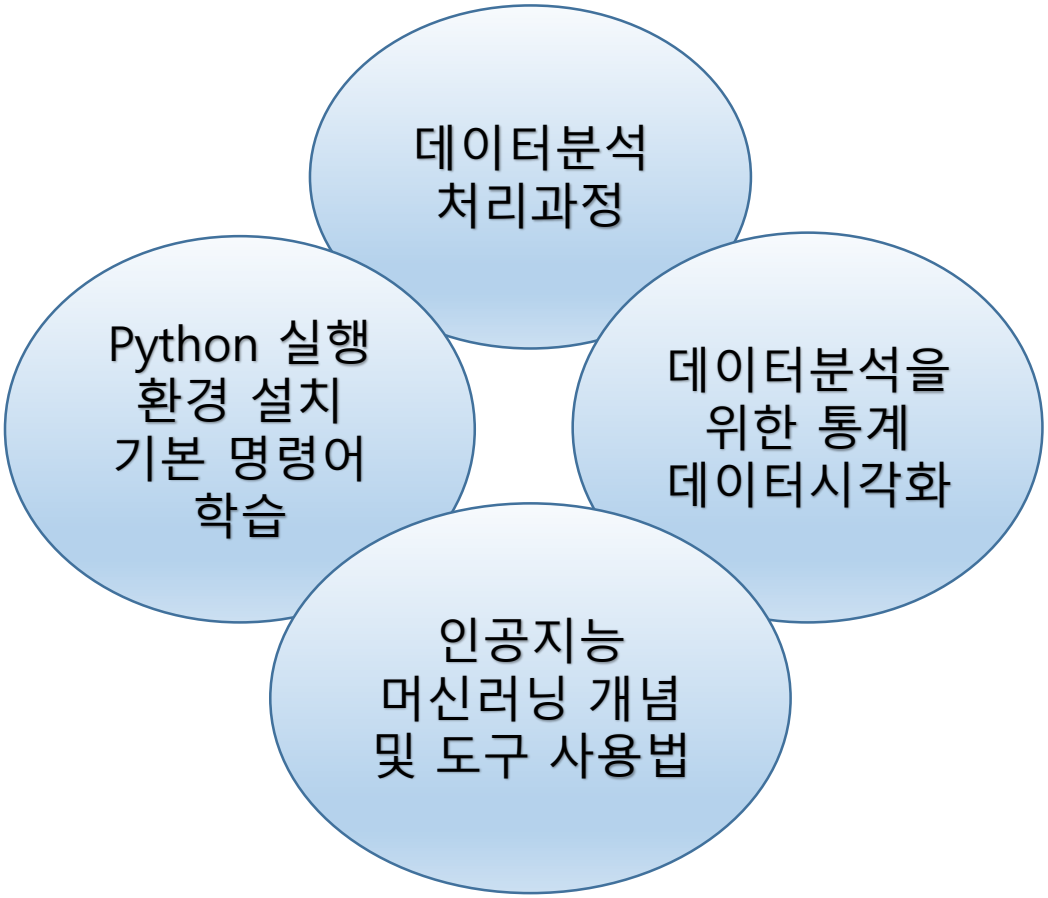
1. 자료(data)의 개념

- 데이터 생산수집처리)
  - 데이터 및 정보의 형태에 따라 굉장히 다양한분야
  - 초보자입장에서 어떤 내용이 있는지?
  - 무엇을 배워야 하는지?
  - 인공지능이 대세라는 데 무엇부터 시작해야 하는지?
  - 숫자 또는 텍스트를 분석 ?

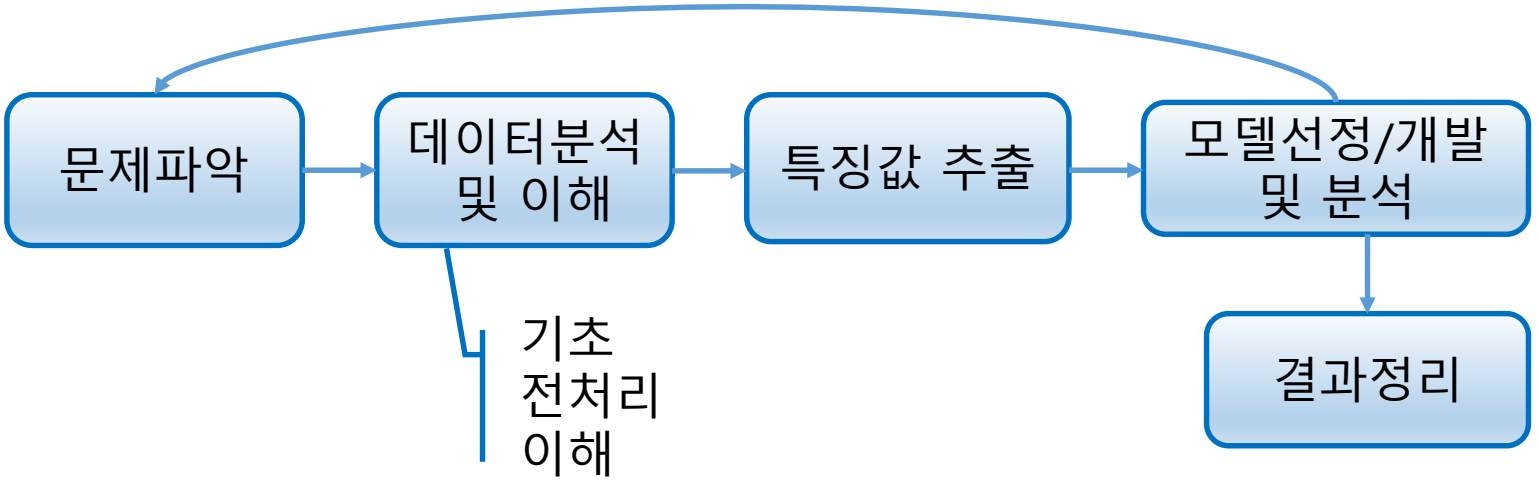


1. 자료(data)의 개념

- 실제로 분석을 하려면 코딩을 해야 한다.
  - Python, R
- 무엇을 공부해야 하나?



2. 데이터의 생산,수집 및 처리과정



문제파악
- 주어진 문제 파악
- 문제해결에 적합한 공학적문제로 정의(모델)하고 문서화

데이터분석 및 이해
- 공학적 문제해결의 해결에 유용한 원시자료를 정의하고 수집
- 데이터 셋의 크기, 모집단의 대표여부, 노이즈 포함여부

특징 값 추출
- 원시자료로부터 문제해결에 유용한 형태의 특징자료를 추출

모델의 선정/개발 및 분석
- 특징 값을 이용하여 문제해결 모델 선정/개발
- 성능 분석,검증

결과 정리
- 문제 해결 결과 정리

2. 데이터의 생산,수집 및 처리과정

- 문제파악
  - 주어진 문제 파악
  - 문제해결에 적합한 공학적문제로 정의(모델)하고 문서화
- 데이터분석 및 이해
  - 공학적 문제해결의 해결에 유용한 원시자료를 정의하고 수집
  - 데이터 셋의 크기, 모집단의 대표여부, 노이즈 포함여부
- 특징 값 추출
  - 원시자료로부터 문제해결에 유용한 형태의 특징자료를 추출
- 모델의 선정/개발 및 분석
  - 특징 값을 이용하여 문제해결 모델 선정/개발
  - 성능 분석,검증
- 결과 정리
  - 문제 해결 결과 정리

2.1 문제파악

- 문제파악
  - 주어진 문제를 파악
  - 문제해결에 적합한 공학적문제로 정의(모델)

2.2 데이터분석 및 이해

- 데이터분석 및 이해
  - 공학적 문제해결의 해결에 유용한 원시자료를 정의하고 수집
  - 기초 : 수집된 데이터의 기본 고려사항
    - 데이터 셋의 크기
    - 모집단의 대표여부
    - 노이즈 포함여부
  - 전처리 : 주처리에 사용하기 좋은 형태로 변환
    - 전처리 프로그램
    - 전처리프로그램으로 데이터 불러오기
    - 불필요한 데이터 걸러 내기
    - 적합한 형식으로 변환하여 저장하기
    - 정답을 알 수 있는 데이터 셋을 정의하고 문제해결코드를 작성하여 정답과 일치하는지 확인한다.
  - 데이터 이해(직관 습득)
    - 데이터의 내용, 구조 등을 살펴보고 데이터에 대한 직관습득
    - 방법
      - 산포도,히스토그램,시계열데이터의 시각화,
      - 머신 분류기를 이용한 분류
      - 데이터 간의 상관관계
      - 데이터의 분포

2.3 특징 값 추출

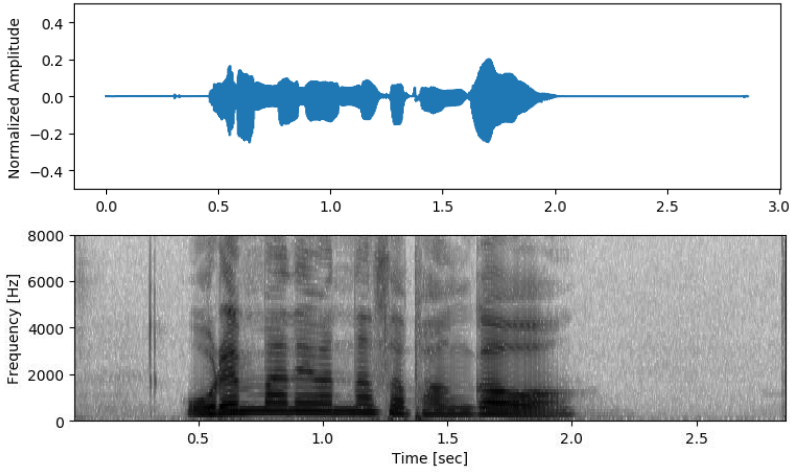
- 특징 값 추출
  - 원시자료로부터 문제해결에 유용한 형태의 특징자료를 추출
  - 우수고객 선정 문제의 예
    - 원시자료 : 고객의 온라인 상거래 자료, 이름,거래일, 거래종목,수량,단가
    - 특징 값 : 이름, 수량,단가
    - 모델 : 매매금액 =단가*수량, 리스트 sorting
  - 음성인식 문제의 예
    - 자료 : 오디오 녹음 파일
    - 특징값 : 13차 mfcc, 1차 delta, 2차delta
    - 모델 : 머신러닝 음성인식기

2.4 모델의 선정/개발 및 분석, 2.5 결과 정리

- 모델의 선정/개발 및 분석
  - 특징 값을 이용하여 문제해결 모델 선정/개발
    - 문제의 해결에 적합한 모델을 선정 또는 개발(?)
      - Linear regression : 선형회귀 , 공부시간-성적예측
      - Logistic regression : 이진분류 , 스팸메일 분류
      - Softmax regression : 멀티분류 , 감정인식, 등등
      - Clustering : 유사자료들끼리 분류
  - 성능 분석,검증
    - 자료 셋 별로 모델 별 결과를 도표화하고
    - 성능을 분석한다.
    - 필요에 따라 검증용 자료 셋을 별도로 정의하고 결과를 분석한다.
- 결과 정리
  - 문제 해결 결과 정리한다.
  - 결과 구성요소를 문서화
    - 데이터 셋
    - 특징
    - 모델
    - 검증 결과 및 분석

3. 감정인식기 개발의 사례

- 문제파악
  - 감정인식기 개발
- 데이터분석 및 이해
  - IEMOCAP data set사용
  - 기초
    - 11종이상 10039 개
    - 5 session 10명의 배우
    - 즉흥, 스크립트 연기
  - 전 처리
    - 4종 대표 감정을 대상으로 선정하여 저장



  - 데이터의 이해

ang	dis	exc	fea	fru	hap	neu	oth	sad	sur	xxx	tot
1103	2	1041	40	1849	595	1708	3	1084	107	2507	10039

Table 1. The emotion class distribution of the dataset on script scenarios and improvisations

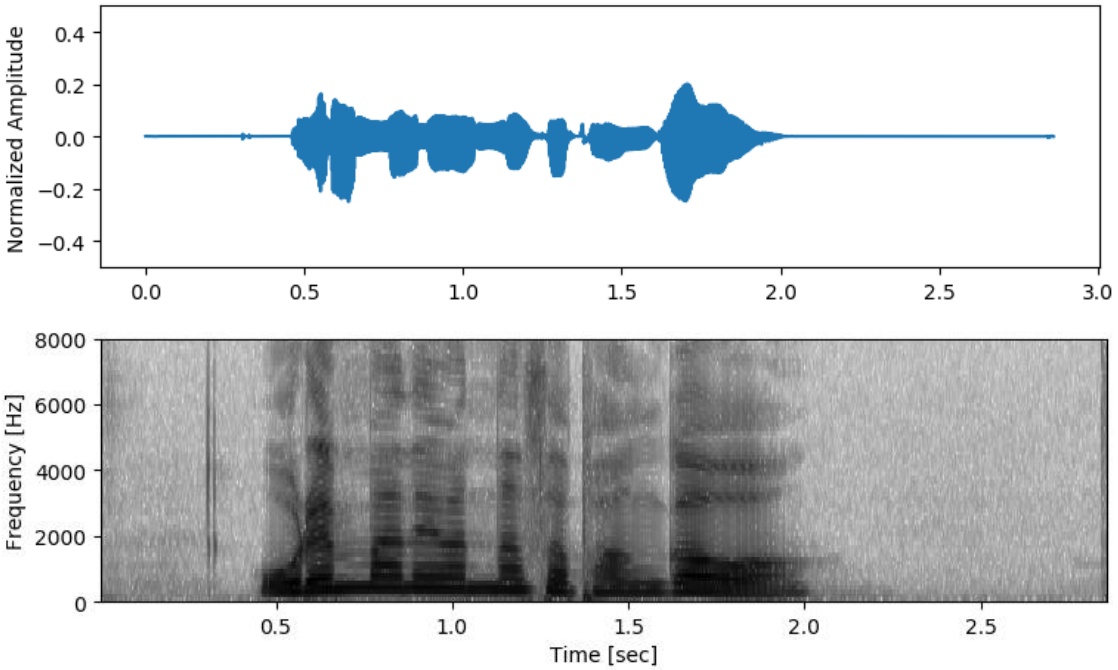
scenario-type	ang	hap	neu	sad	Total
Full	1103	595	1708	1084	4490
impro	289	284	1099	608	2280
script	814	311	609	476	2210

Table 2. The emotion class distribution for 5-fold cross-validation

Session	ang	hap	neu	sad	Total
1	229	135	384	194	942
2	137	117	362	197	813
3	240	135	320	305	1000
4	327	65	258	143	793
5	170	143	384	245	942
Total	1103	595	1708	1084	4490

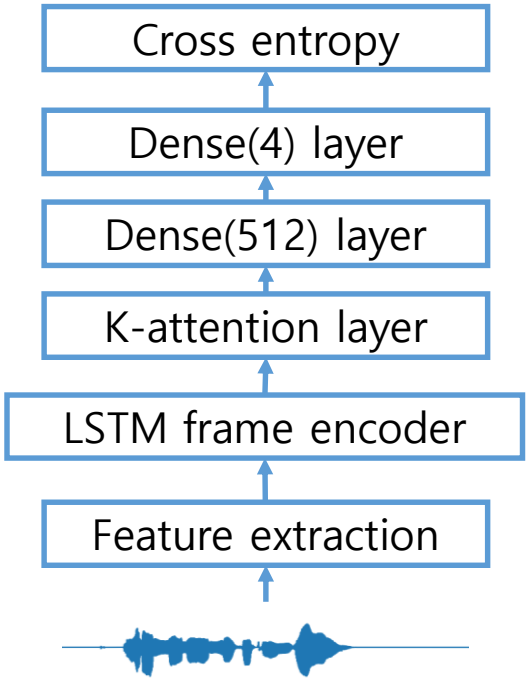
3. 감정인식기 개발 사례

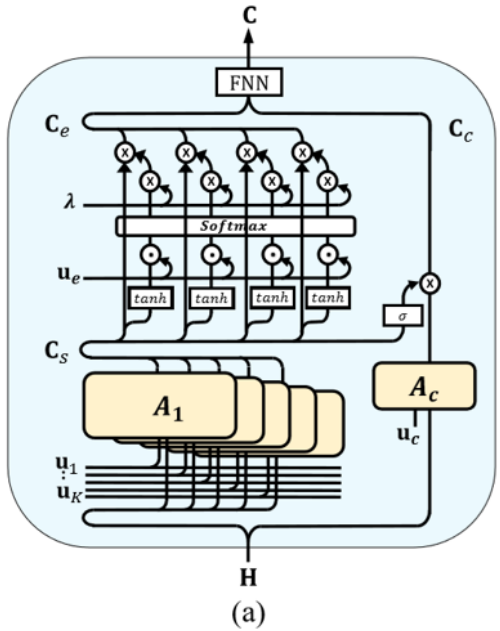
- 문제파악 : 감정인식기 개발
- 데이터분석 및 이해
- **특징 값 추출**
  - 32 차 실수 벡터
  - 12 mel-frequency cepstral coefficients(MFCC), pitch, energy, zero-crossing energy, voicing probability as well as the first derivatives



3. 감정인식기 개발사례

- 문제파악 : 감정인식기 개발
- 데이터분석 및 이해
- 특징 값 추출
- **모델의 선정/개발 및 분석**
  - 특징 값을 이용하여 문제해결 모델 선정/개발
  - 성능 분석,검증





(a)

Figure 1. The conceptual diagram of t configuration. (b) The attention model.

3. 감정인식기 개발 사례

- 문제파악 : 감정인식기 개발
- 데이터분석 및 이해
- 특징 값 추출
- **모델의 선정/개발 및 분석**
  - 특징 값을 이용하여 문제해결 모델 선정/개발
  - 성능 분석,검증

Table 3. Results of 4 experiments of 5-fold cross validation on the full dataset

run	Unweighted average (UA)			Weighted average (WA)		
	min	mean	max	min	mean	max
1	58.17	61.20	66.83	58.63	58.97	59.53
2	60.72	64.89	67.28	53.41	58.45	61.51
3	54.88	57.89	59.60	52.72	53.66	55.07
4	62.80	65.25	68.98	58.11	60.45	63.35
5	56.79	61.68	64.54	51.31	55.25	57.70
mean	58.67	62.18	**65.45**	54.84	57.36	**59.43**

Table 4. Comparison of SER models in terms of weighted average (WA) and unweighted average (UA) for full dataset and improvised dataset

Model (5-fold CV scheme)	Full dataset		Improvised dataset	
	UA	WA	UA	WA
Mirsamamdi et al. [6]	63.5	58.8		
Ramet et al. []	62.5	59.6	68.8	63.5
Proposed model	65.5	59.4	66.4	65.1

3. 감정인식기 개발 사례

- 문제파악 : 감정인식기 개발
- 데이터분석 및 이해
- 특징 값 추출
- **모델의 선정/개발 및 분석**
  - 특징 값을 이용하여 문제해결 모델 선정/개발
  - 성능 분석,검증

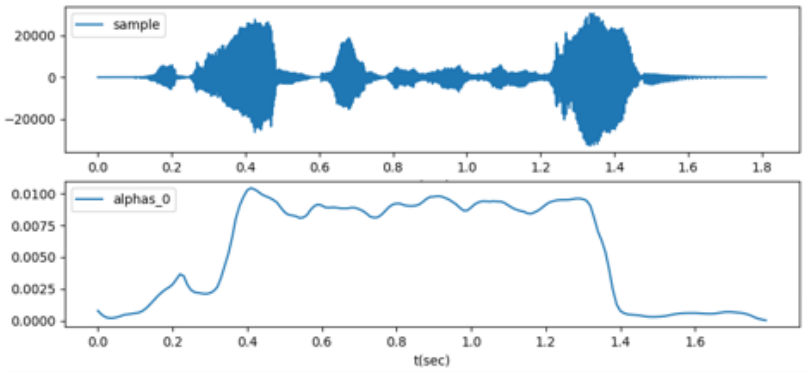


Figure 5 Attention weights of one-attention mechanism for
Top: the raw waveform; bottom: the attention weights α_t ov

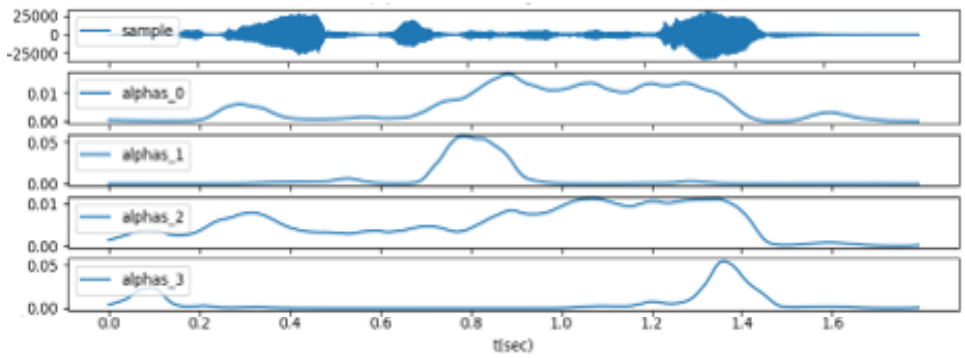


Figure 6. Attention weights of K-attention mechanism for the
Top: the raw waveform; bottom: 7 sets of attention weight set

- 문서화