음성인식에서의 딥러닝 모델과 언어 모델간의 차이

캐글 스터디

음성 인식

- **음성 인식은 음향 신호에 숨겨져 있는 단어의 시퀀스를 찾는 과제**
  - 음성 속에서 음운을 발견해내는 과정
- 음향 : 단순한 자연의 소리
- 음성 : 사람이 발화하는 물리적 말소리
- 음운 : 뜻을 구분하는 심리적 말소리
  - 음소 : 자음, 모음
  - 운소 : 소리의 장단, 고저, 억양

음성 인식 시스템의 두 가지 갈래

- 통계적 패턴 매칭
  - 음운 사전(lexicon)을 미리 만들어 두고 입력된 음성을 음운으로 변환하되, 가능성이 가장 높은 음운의 나열(단어)로 치환하는 것
  - 이 과정에서 HMM(Hidden Markov Model) 이 주로 사용
- 딥러닝 매칭
  - 세 개의 모델을 활용하지 말고 음성이 입력되면 문장(transcription)을 직접 치환하는 것

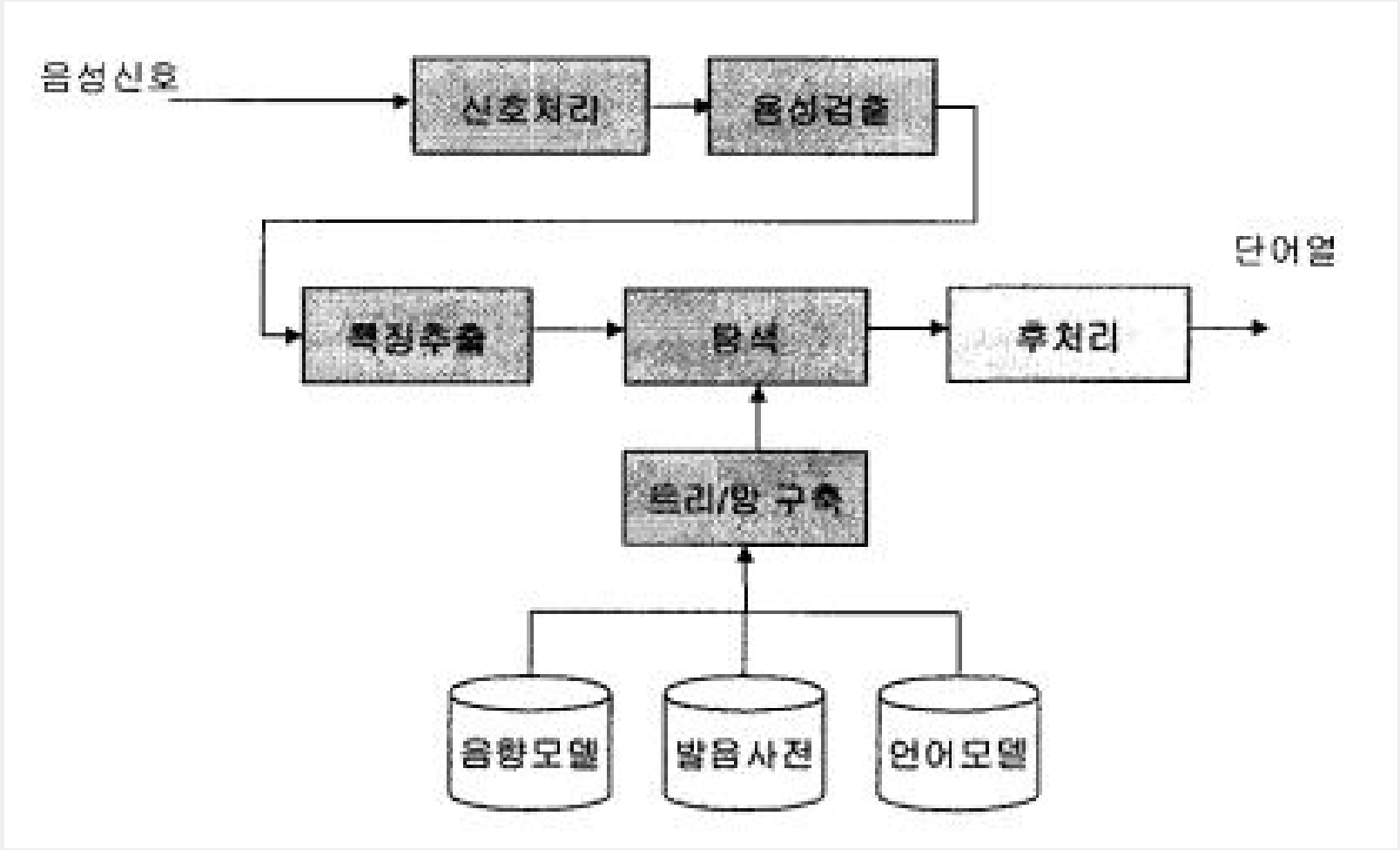
통계적 패턴 매칭

- 전통적인 통계 기반 모델 : HMM을 주로 활용
  - HMM(Hidden Markov Model)
    - Markov Chain 성질을 전제로 관측 이전의 은닉 상태를 추론하는 데 사용하는 모델
    - Markov Chain 성질 : 한 상태의 확률은 단지 그 이전의 상태에만 의존한다
    - 각 관측치에 대한 우도를 다이나믹 프로그래밍(DP)로 저장해두고 확률을 계산
    - HMM의 관측 결과가 주어졌을 때 최대 우도를 되추적하는 과정이 비터비 알고리즘(Viterbi Algorithm)

통계적 패턴 매칭

- 전통적인 통계 기반 모델
  - 음향모델 - 입력 음성을 바탕으로 최대 우도를 되추적하는 모델(HMM)
  - 발음모델 - 음운 사전(lexicon), 언어학적 지식이 중요하게 적용되는 모델
  - 언어모델 - 단어 간의 시계열 상관 관계를 수식적으로 모델링하는 방법, 문맥(문법)에 맞는 말이 나오게 도움을 주는 모델, n-gram이 적용
- HMM 기반의 모델은 도메인(언어학 - 음향음성학) 지식이 중요시되는, 전통적인 방식의 음성 인식 모델
- 음향, 발음, 언어 모델을 모듈화할 수 있다는 것이 특징
  - 각 모델의 최적이 전체의 최적으로 이어지지 않을 수도 있음
  - 다른 도메인에 적용 가능성

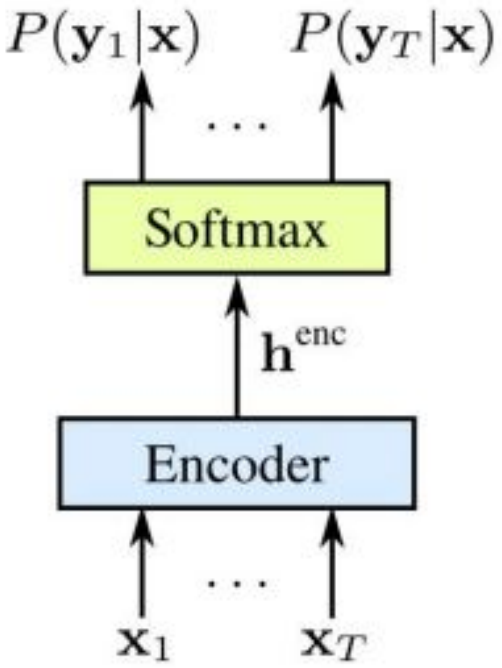
통계적 패턴 매칭



딥러닝 매칭

- 딥러닝 매칭은 E2E 모델
  - E2E 모델
    - 음성 입력을 받아 단어/음소 시퀀스를 직접적으로 예측하는 모델
- 딥러닝을 활용한 유명한 모델
  - CTC(Connectionist Temporal Classification) : Deep Speech2(Baidu, 중국)
  - LAS(Listen, Attend and Spell) : Seq2Seq 를 활용한 모델
    - listen이 encoder, attend가 attention, spell이 decoder
  - RNN-T(Recurrent Neural Network Transducer) : 실시간(Online algorithm)으로 음성인식이 가능한 모델

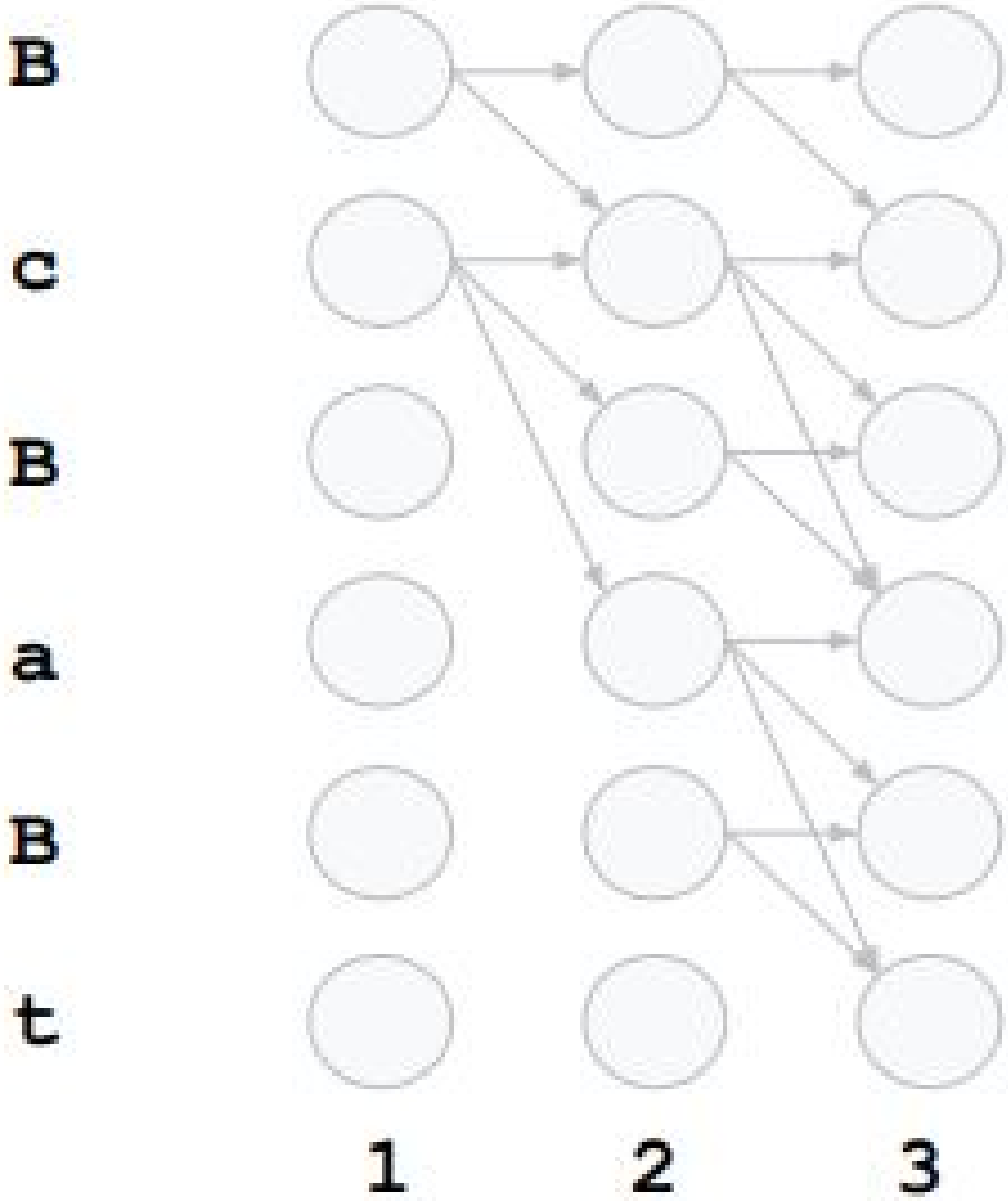
CTC



B B **c** B B **a** **a** B B **t**
B **c** **c** B **a** B B B B **t**
...
B **c** B B **a** B B **t** **t** B

$$P(\mathbf{y}|\mathbf{x}) = \sum_{\hat{\mathbf{y}} \in \mathcal{B}(\mathbf{y},\mathbf{x})} \prod_{t=1}^{T} P(\hat{y}_t|\mathbf{x})$$

- 정렬 정보(얼라인먼트) 없이 음성을 입력받아 가장 출력될 확률이 높은 단어 경로를 출력하기
- P(y|x) : x 음성이 입력되었을 때 y 라고 인식할 확률
  - B(blank)를 포함하여 발생할 수 있는 모든 확률 그래프 경로의 총합



CTC

- 세로축이 음성, 가로축이 시간
- self-loop : 단순히 오른쪽으로만 가기
- left-to-right : 왼쪽에서 오른쪽으로만 가기
- 확률 DAG(Directed Acyclic Graph) 경로에서 가장 가능성이 높은 그래프 경로를 찾아내는 것이 목적

재미있는 참고 자료 - 알고리즘 문제 해결 전략(광학 문자 인식)

- https://algospot.com/judge/problem/read/OCR
  - 각 단어가 문장의 처음에 등장할 확률 B(i)
  - 각 단어의 다음에 다른 단어가 나올 확률의 행렬 T(i, i)
  - 각 단어를 다른 단어로 분류할 확률 M(i, i)
  - 한 줄에 i 개의 단어 나열이 주어졌을 때 조건부 출현 확률이 가장 높은 문장을 되추적하여 출력하는 문제

참고 자료

- 이 자료는 다음 자료들에서 거의 대부분을 가져왔음을 밝혀둡니다
- https://ratsgo.github.io/speechbook/
- http://iscslp2018.org/images/T4_Towards end-to-end speech recognition.pdf