Automatic Speech Recognition

음성 인식 모델

강사 양석환

소리의 이해

소리

- 소리: 공기의 진동 → 진동의 연속 → 파형 발생
- 소리(음, 음파)
  - 공기나 물 같은 매질의 진동을 통해 전달되는 종파
    - 종파: 매질의 진동 방향이 파동의 방향에 일치하는 파동
    - 인간이 감지하는 가청주파수: 20Hz ~ 2만Hz(헤르츠)의 주파수 대역
  - 음압
    - 소리의 세기는 그 파동이 얼마나 큰 압력을 갖고 있느냐로 계산 → 음압. 단위는 데시벨(dB)
    - 데시벨: 상대적인 값. 0dB를 기준으로 10dB가 증가할 때마다 그 음압은 10의 거듭제곱 꼴로 커짐
    - 인간의 귀는 주파수나 데시벨에 따라 음압을 정확하고 순차적으로 인식하지 못하기 때문에 인간이 느끼는 음의 상대적인 크기를 고려하여 사용함

소리

- 소리의 높낮이
  - 진동 수에 의해서만 결정됨 (파장과는 관련이 없음)
    - 진동수가 높으면 높은 소리, 진동수가 낮으면 낮은 소리
  - 소리의 속도가 일정하다고 가정할 때, 파장은 진동수에 반비례 함
- 음파의 속도(음속)
  - 온도 15℃의 공기 속을 전파하는 음속은 대략 340m/s
  - 음속은 진동수나 기압에는 관계가 없고 공기의 온도에 의해서만 결정됨
    - 음속이 공기의 온도에 의해 변하는 것은 공기의 밀도가 온도에 의해 변하기 때문
    - 밀도가 작을수록, 혹은 온도가 높을수록 매질은 이동하기 쉬워져서 음속은 빨라짐

오디오 디지털 신호에 대한 이해

- 디지털방식의 기록
  - 파형을 시간에 따라 잘게 나누어 각각의 대푯값을 추출하여 0과 1의 조합으로 만든 것
  - 음파의 파형은 사인 곡선 → 컴퓨터 프로그램으로 조정 가능

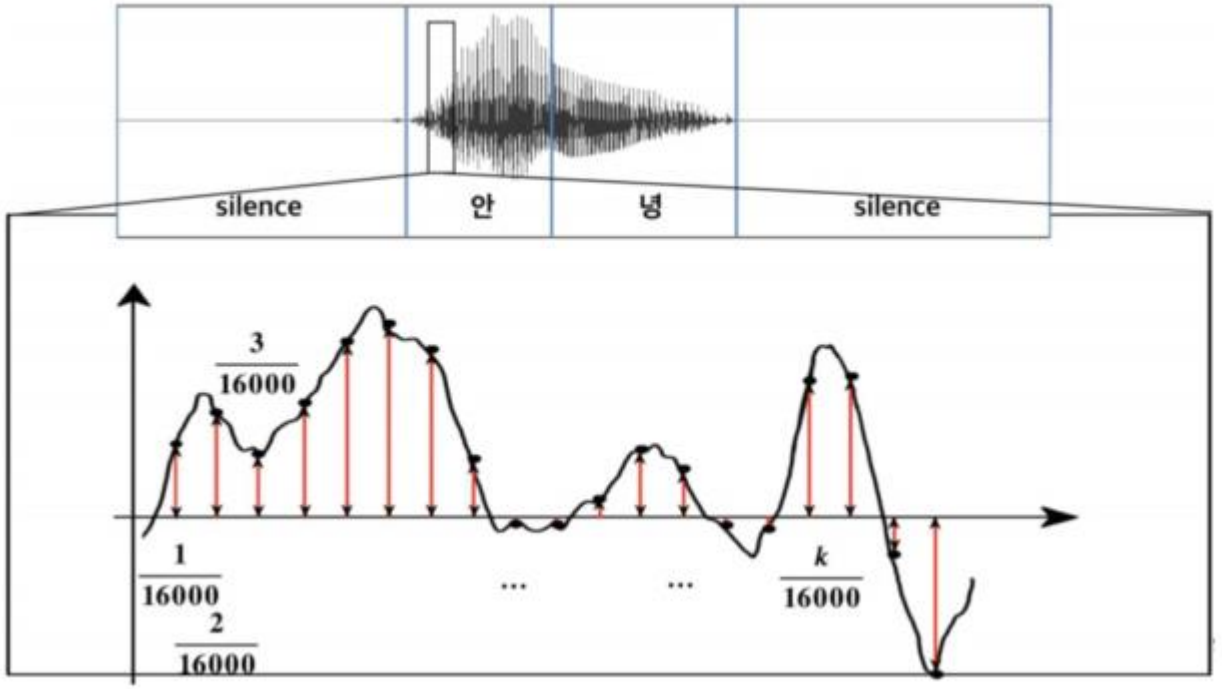
음성 신호의 특징

- **"안녕" 발성에 대한 16,000Hz, 16bit, mono 형태의 PCM 포맷 저장 방법**
  - **PCM: Pulse Code Modulation**
  - **샘플링 주파수: 16,000Hz**
  - **샘플 당 16bit**
  - **마이크가 1개가 있는 1Channel(Mono)**

16,000Hz
1초를 저장하는데 총 16,000개의 샘플이 필요함을 의미

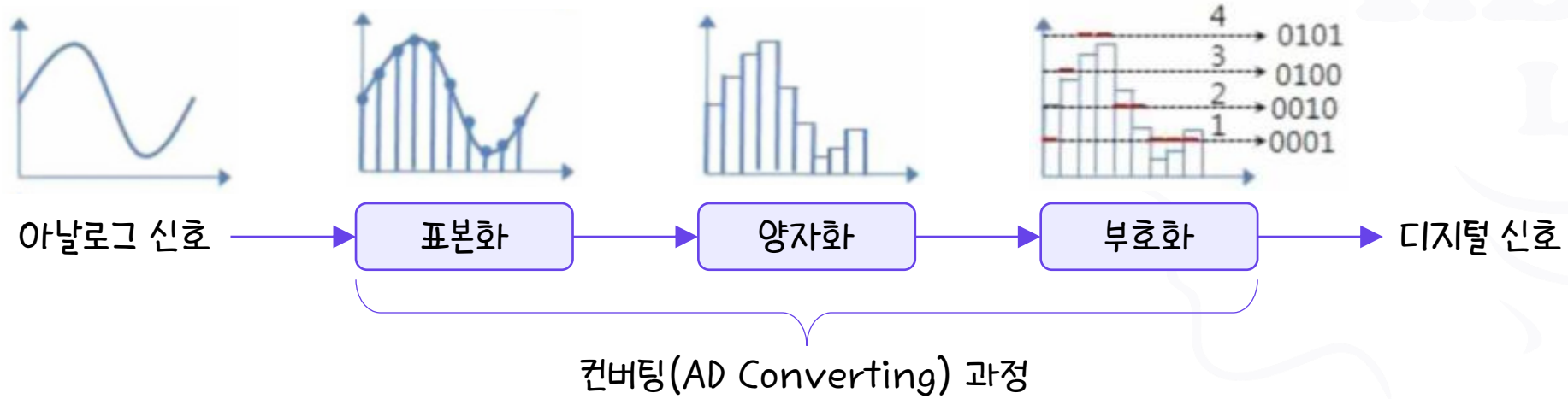
PCM
- 원본 형태의 디지털 오디오 데이터를 그대로 저장하는 방식
- 압축하지 않는 디지털로 변환 오디오 신호의 raw 데이터를 그대로 저장하는 방식

WAV
- PCM 데이터에 메타 정보(샘플링 주파수, bit 해상도, 채널 수, duration 등)을 포함하여 압축하지 않는 형태
- 대부분 많은 음성 인식 시스템에서 채택



신호 변환(AD Converting) 과정

- 표본화(Sampling): 시간 축 방향에서 일정 간격으로 샘플을 추출하여 연속 신호로부터 이산 신호로 변환하는 과정
- 양자화(Quantization): 샘플링 된 진폭치를 특정 대푯값으로 변환하는 과정
- 부호화(Coding): 신호처리가 용이한 디지털 코드 (Binary Code) 형태로 변환하는 과정



신호 변환 시 고려해야 할 내용

- 추출률(Sampling Rate)
  - ‘시간을 얼마나 잘게 쪼개는가’의 비율.
  - 음질과 밀접한 관계가 있으며 이 비율이 클수록 음질이 더 좋아짐
- 샘플링 크기
  - 가청주파수 대역을 모두 디지털화 하려면 적어도 그 2배 이상의 샘플링이 필요함 (나이퀴스트 이론에 의거)
    - 예
      - 2만Hz를 주파수로 가지는 음파의 정보를 살리기 위해서는 4만Hz를 샘플링 하여 녹음해야 함
      - CD의 샘플링은 44,100Hz → 1초를 44,100개로 쪼개서 각각의 나누어진 부분으로 디지털화 했음을 의미
  - 2배의 샘플링이 필요한 이유: 충분한 안정성 확보

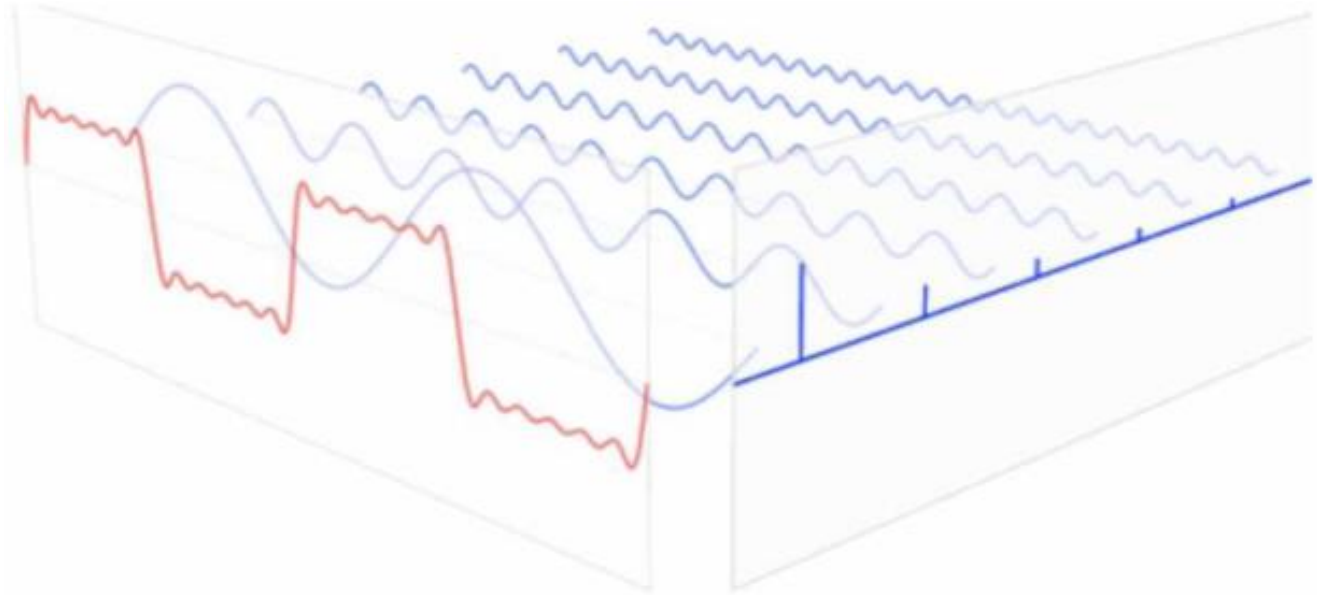
음성 인식 시스템

푸리에 변환(Fourier transform)

- 푸리에 변환이란?
  - 임의의 입력 신호를 다양한 주파수를 갖는 주기함수들의 합으로 분해하여 표현하는 것
  - 푸리에 변환에서 사용하는 주기함수: sin, cos 삼각함수
  - 즉, 고주파부터 저주파까지 다양한 주파수 대역의 sin, cos 함수들로 원본 신호를 분해하는 것
  - 신호처리, 음성, 통신 분야에서 뿐만 아니라 영상처리에서도 매우 중요한 개념
- 푸리에 변환의 응용 방향
  - 영상을 주파수 성분으로 변환하여 다양한 분석 및 처리 수행
  - 임의의 필터링 연산을 FFT(fast Fourier transform)를 이용하여 고속으로 구현

푸리에 변환(Fourier transform)

- 푸리에 변환 → “복잡한 신호 = 정현파의 합”



- 푸리에 변환의 강점
  - 입력 신호가 어떤 신호이든지 관계없이 임의의 입력 신호를 sin, cos 주기함수들의 합으로 항상 분해할 수 있다는 것

푸리에 변환(Fourier transform)

- **푸리에 변환의 수식**
  - **푸리에 변환**

$$f(x) = \int_{-\infty}^{+\infty} F(u) e^{j2\pi ux} du$$

- j : 허수단위 ($j = \sqrt{-1}$)
- $f(x)$: 원본 입력 신호
- $e^{j\pi ux}$: 주파수 u 의 주기함수 성분
- $F(u)$: 해당 주기함수 성분의 계수

  - **푸리에 역변환**

$$F(x) = \int_{-\infty}^{+\infty} f(x) e^{-j2\pi ux} dx$$

음성 인식 시스템의 처리 과정

- 1단계: Feature 추출
  - 오디오 신호를 식별하는 작업
  - 언어 콘텐츠를 식별하고, 다른 것들은 (background noise, emotion 등의 정보를 담은 신호) 제거
  - Mel Frequency Cepstral Coefficients (MFCCs)
    - 자동 음성 인식, 화자 인식에서 널리 쓰이는 feature(1980년대 Davis and Mermelstein에 의해 소개)
  - 성도(성대에서 입술 또는 콧구멍에 이르는 통로)의 모양
    - 인간이 만드는 소리는 혀, 치아 등을 포함하는 성도의 모양에 의해 결정됨
    - 성도의 모양을 정확히 확인할 수 있다면 우리는 생성되는 음소의 정확한 실현이 가능함
    - 성도의 모양은 그 자체가 짧은 구간의 파워 스펙트럼의 포락선을 나타냄
    - MFCC는 이 포락선을 정확히 나타내는 역할을 수행

음성 인식 시스템의 처리 과정

- 2단계: 각 프레임의 파워 스펙트럼 계산
  - 인간의 귀에 들어오는 소리의 주파수에 따라 다른 부분이 진동하는 cochlea(귀에 있는 기관)을 보고 고안됨
  - cochlea가 진동하는 위치에 따라 다른 세포들이 자극되어 뇌에 특정 주파수가 있다고 알림
  - 신호의 스펙트럼 밀도 추정을 위한 주기도 추정(periodogram estimate)은 어떤 주파수가 이 프레임에 있는지 식별하는 cochlea와 유사한 역할을 수행
  - Mel Filter bank
    - 첫 번째 필터는 매우 좁고 에너지가 0 Hertz 근처에 있는지를 나타냄
    - 주파수가 높아질수록 필터는 넓어지게 되고 대략적으로 에너지가 어느 주파수 영역에 등장하는지 확인 가능
  - Mel scale
    - 정확하게 어떻게 우리 filter bank의 간격을 배치할지, 얼마나 넓게 만들지 결정

음성 인식 시스템의 처리 과정

- **Filter bank 에너지를 얻은 후, 로그 적용**
  - 인간의 듣기(우리는 선형 스케일로 소리의 강도를 듣지 않음)에 영향을 받은 과정
  - 일반적으로 감지된 소리의 볼륨을 2배로 얻기 위해 우리는 소리에 들어있는 에너지의 8배를 필요로 함
  - 이것은 만약 소리가 처음에 크면, 에너지에서 큰 변동은 그렇게 많이 다르게 들리지 않는다는 것을 의미함
  - 이런 압축 기능은 우리의 feature를 인간이 실제 듣는 것에 가깝게 매치 시킴

음성 인식 시스템의 처리 과정

- 3단계: Log filter bank 에너지의 DCT 계산
  - DCT: Discrete Cosine Transform (이산 코사인 변환)
  - 우리의 filter bank는 모두 겹치기 때문에, filter bank의 에너지는 서로 깊은 상관관계를 가짐

    → Diagonal Covariance Matrices(대각 공분산 행렬)가 HMM 분류기처럼 Feature를 만드는데 사용된다는 것을 의미
  - 확인된 26개의 계수 중 12개의 DCT coefficients(계수)만 유지

    → DCT 계수가 높을수록 filter bank 에너지의 빠른 변화를 나타내고 이것은 ASR의 성능의 하락을 나타내기 때문

    → DCT 계수의 값을 떨어뜨림으로써 성능 개선 추구

음성 인식의 이해

음성 인식 개요

- 의사 소통의 방식
  - 사람이 의사 소통하려고 사용하는 일반적이고 효과적인 수단은 언어(말과 글)
  - 음성은 가장 자연스러운 의사소통 방식
  - 인간과 컴퓨터 간의 정보 교환 시 가장 자연스러운 인터페이스를 제공함
- 음성 언어 처리 기술
  - 인간의 자연어 발화를 컴퓨터가 자동으로 이해하고, 처리하는 알고리즘을 연구하는 분야
- 다양한 응용 서비스 사례
  - 대화형 개인 비서 에이전트, 인공지능(AI) 스피커, 자동 통번역, 음성 대화 질의 응답 (Q&A) 시스템 등

음성 처리 분야의 목표

- 화자 인식(화자는 누구인가?)
  - 화자 인식은 발성 내에서 화자가 누구인지를 찾는 문제
- 음성 인식(화자의 발성은 무엇인가?)
  - 음성 신호에 숨겨져 있는 단어의 시퀀스를 찾는 문제

현재의 음성 인식 시스템의 특징

- 현재의 대부분의 음성 인식 시스템은 통계적 패턴 매칭 원리에 기반함
  - 사람이 발성한 음성신호는 신호 처리기에 의해 음향학적 벡터 열로 변환됨
  - 각 벡터는 짧은 시간의 음성 구간 (약 10 ~ 20ms)에 대한 에너지 스펙트럼을 나타냄
- 주어진 음성신호가 특정 단어 열 W를 발성한 결과라고 했을 때 음성 인식 시스템의 목표는
  - 음성신호로부터 추출된 벡터 열 O에 대해서 가장 높은 확률을 가지는 단어 열 $\hat{W}$를 제시함과 동시에 $W = \hat{W}$ 를 만족하는 것
  - 그러나 같은 사람이 같은 단어 열을 발성한다고 해도 음성신호로부터 변환되는 음향학적 벡터 열은 다르게 나올 수밖에 없음 → 동일한 발성이라고 하더라도 가능한 벡터 열 O 의 가짓수는 무한대
  - 문제 해결을 위하여 Bayes 정리 사용

Bayesian Rule

- 음성 인식의 수식적 정의
  - $\widehat{W} = \underset{W}{\operatorname{argmax}} P(W|O) = \underset{W}{\operatorname{argmax}} \frac{P(W)P(O|W)}{P(O)} = \underset{W}{\operatorname{argmax}} P(W)P(O|W)$
    - $P(O)$: 특정 벡터 열 O의 발생확률
  - 음성 인식에 사용되는 음향 모델을 잘 학습하기 위하여 확률이론, 정보이론 등이 기본적으로 사용됨

- Bayes 정리 또는 Bayesian rule
  - 이전 경험과 현재의 증거를 토대로 어떤 사건의 확률을 추론하는 과정

Bayesian Rule

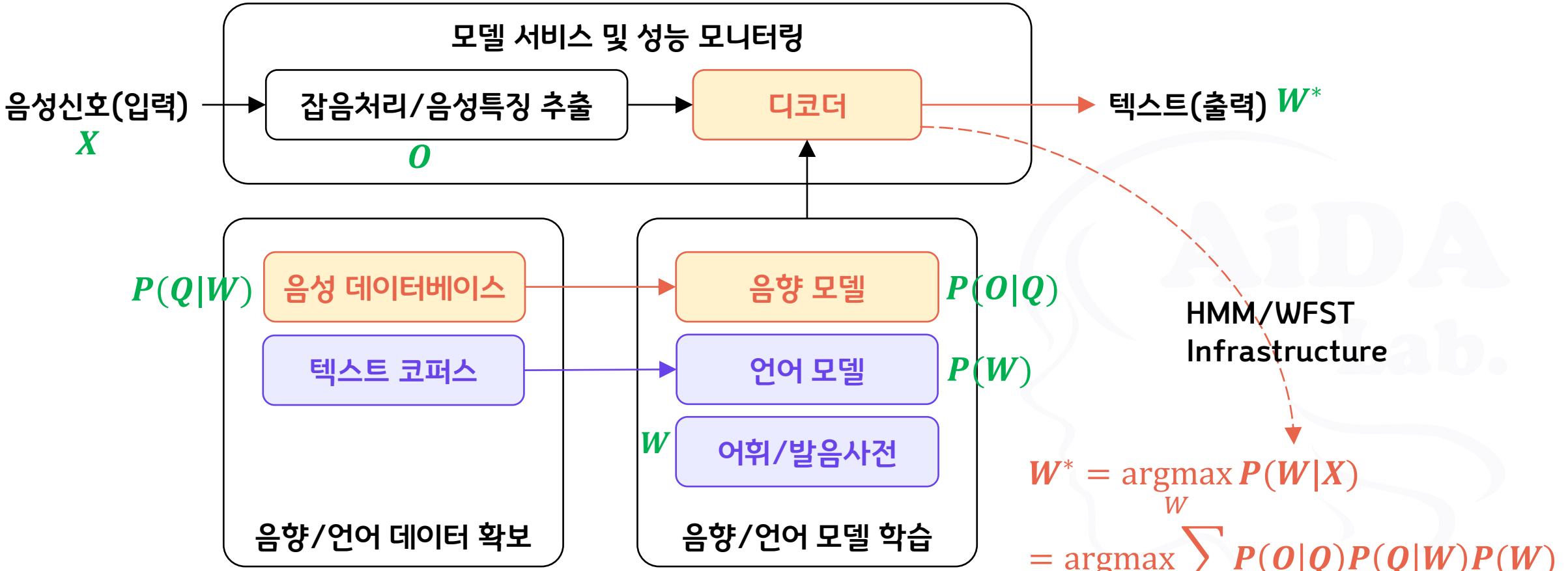
- **Bayesian rule에 대한 각 요소에 대한 의미**

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$$

  - $P(A)$: **A의 사전 확률 (evidence, 현재의 증거)**
  - $P(B)$: **B의 사전 확률(prior probability, 과거의 경험)**
  - $P(A|B)$: **사건 B가 주어졌을 때 A의 조건부 확률(likelihood, 알려진 또는 관찰된 결과에 기초한 어떠한 가설에 대한 가능성, 즉 이 가설을 지지하는 정도)**
  - $P(B|A)$: **사건 A라는 증거에 대한 사후 확률(posterior probability, 사건 A가 일어났다는 것을 알고, 그것이 사건 B에 영향을 줘서 일어났다는 조건부 확률)**

음성 인식 컴포넌트

- 음성 인식 모듈 구현을 위한 주요 컴포넌트



음성 인식 컴포넌트

- 어휘/발음 사전 (Vocabulary Dictionary, Lexicon)
  - 인식 단어(Word)를 발음 열(Phone Sequence)로 표현
- 음향 모델링 (Acoustic Modeling)
  - 입력된 발음 열 중에서, 발음 하나하나를 수식적으로 모델링 하는 방법
  - 딥러닝 이전: Hidden Markov Model(HMM) / 딥러닝 이후: DNN, RNN 등 적용 추세
- 언어 모델링 (Language Modeling)
  - 입력된 발음 열 중에서, 단어 간의 시계열 상관관계를 수식적으로 모델링 하는 방법
  - 딥러닝 이전: 통계적 n-gram / 딥러닝 이후: Word Embedding 기반 RNN 등 적용 추세
  - 예측(Decoding): 디코딩 네트워크(HMM topology 또는 Weighted Finite State Transducer (WFST))를 활용

음향 모델

HMM(Hidden Markov Model) 개요

- 은닉 마르코프 모델(Hidden Markov Models, HMMs)
  - 통계적 마르코프 모형의 하나
  - 시스템이 은닉된 상태와 관찰가능한 결과의 두 가지 요소로 이루어졌다고 보는 모델
  - 순차적인 데이터를 다루는 데 강점을 지녀 개체명 인식, 품사 태깅 등 단어의 연쇄로 나타나는 언어구조 처리에 과거 많은 주목을 받았던 기법
  - 마르코프 체인(Markov chain)을 기반으로 함

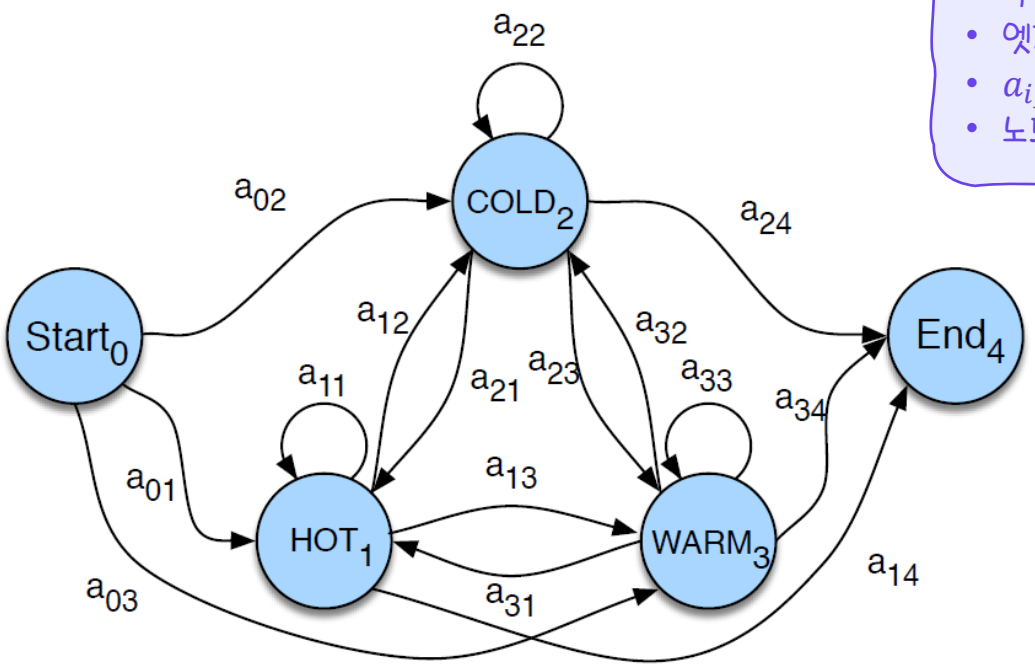
HMM(Hidden Markov Model) 개요

- 마르코프 체인
  - 러시아 수학자 마르코프가 1913년경에 러시아어 문헌에 나오는 글자들의 순서에 관한 모델을 구축하기 위해 제안된 개념
  - 마르코프 성질(Markov Property)을 지닌 이산확률 과정(discrete-time stochastic process)
  - 핵심개념
    - 한 상태(state)의 확률은 단지 그 이전 상태에만 의존한다

      → 즉, 한 상태에서 다른 상태로의 전이(transition)는 그동안 상태 전이에 대한 긴 이력(history)을 필요로 하지 않고 바로 직전 상태에서의 전이로 추정할 수 있다

HMM(Hidden Markov Model) 개요

- **마르코프 체인의 모델링**
  - **도식화:** $P(q_i|q_1, q_2, \cdots, q_{i-1}) = P(q_i|q_{i-1})$
  - **날씨를 마르코프 체인으로 모델링한 예시**



- 각 노드: 상태(일반적인 상태 + 시작 + 끝)
- 엣지: 전이
- a_{ij}: i번째 상태에서 j번째 상태로 전이할 확률
- 노드 별 전이 확률의 총 합은 1

마르코프 체인의 세부 내용 표시의 예

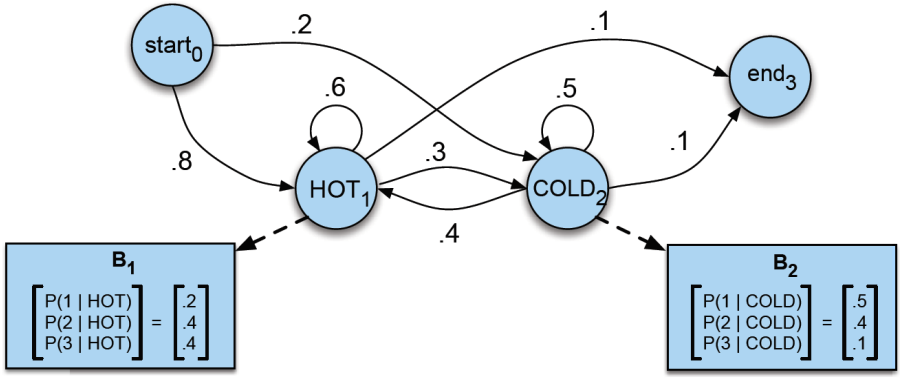


그림 출처: https://ratsgo.github.io/machine%20learning/2017/03/18/HMMs/

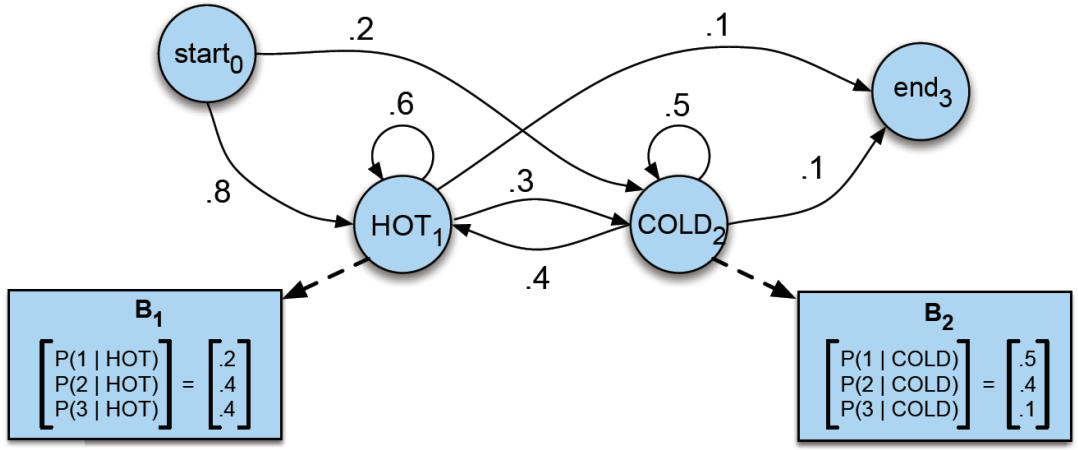
HMM의 주요 요소

- **HMM의 주요 요소에 대한 Notation**

Notation	내용	설명
$\boldsymbol{Q}$	상태집합	$\boldsymbol{Q} = \boldsymbol{q_0}, \boldsymbol{q_1}, \boldsymbol{q_2}, \cdots, \boldsymbol{q_n}, \boldsymbol{q_F}$ ($\boldsymbol{q_0}$: 시작 상태, $\boldsymbol{q_F}$: 종료 상태, $\boldsymbol{n}$: 상태의 개수)
$\boldsymbol{A}$	전이확률 행렬	전이확률 행렬($\boldsymbol{n} \times \boldsymbol{n}$). a_{ij}: i 번째 상태에서 j 번째 상태로 전이할 확률
$\boldsymbol{B}$	방출 확률	$\boldsymbol{b_j(o_t)}$, $\boldsymbol{j}$ 번째 상태에서 $\boldsymbol{t}$ 번째 관측치 $\boldsymbol{o_t}$가 나타날 방출확률
$\boldsymbol{O}$	관측치의 시퀀스	$\boldsymbol{O} = \boldsymbol{o_0}, \boldsymbol{o_1}, \boldsymbol{o_2}, \cdots, \boldsymbol{o_n}, \cdots, \boldsymbol{o_t}$: 길이가 T인 관측치의 시퀀스

HMM의 주요 요소

- **날씨 상태 전이 확률 행렬**



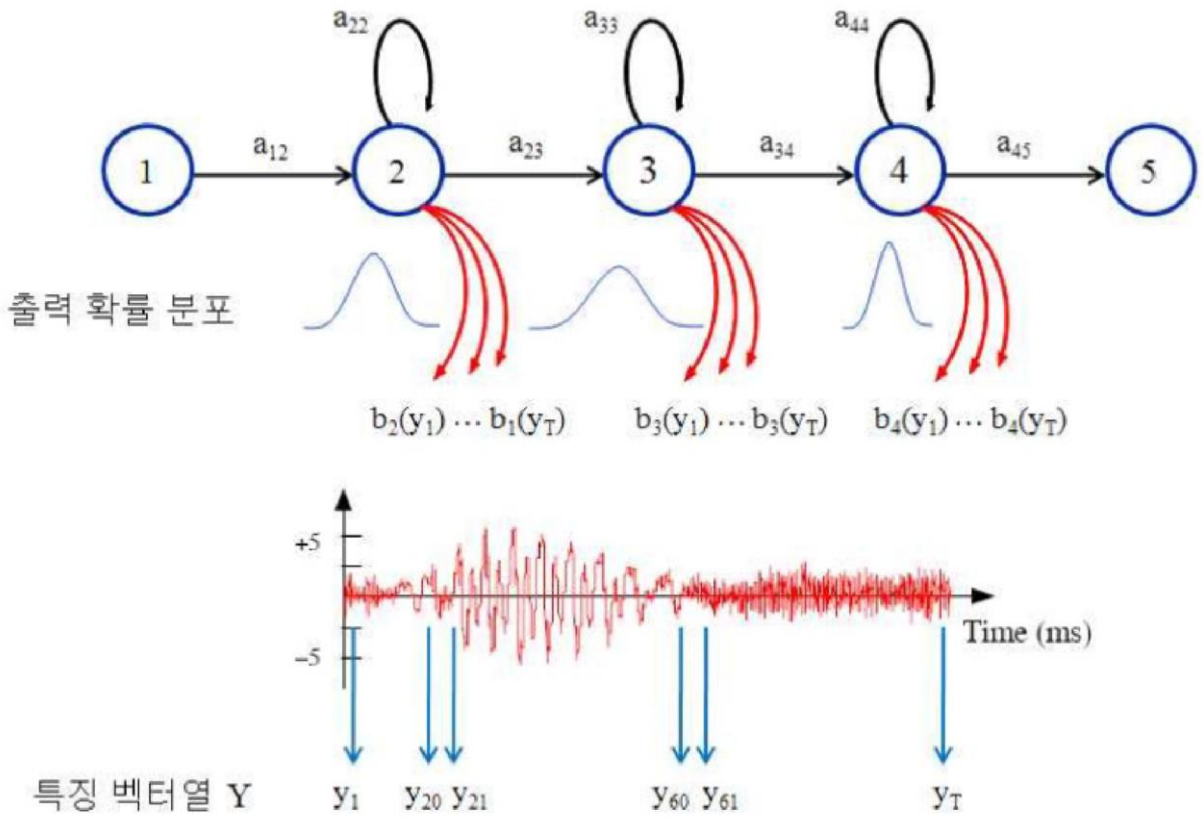
전이 이전 상태(i)	전이 이후 상태(j)			
	0	1	2	3
0 (시작상태)	0.0	0.8	0.2	0.0
1 (HOT상태)	0.0	0.6	0.3	0.1
2 (COLD상태)	0.0	0.4	0.5	0.0
3 (종료 상태)	0.0	0.0	0.0	0.0

HMM 기반 음향 모델링

- **음향 모델의 목표**
  - **주어진 단어 w 에 대해서 모든 벡터 열 Y 에 대한 우도(likelihood)를 계산하는 방법을 제공하는 것**
- **처리 방안: 주로 통계적인 방법으로 해결(기존의 방법)**
  - 모든 단어 w 를 발성한 음성신호에서 추출된 벡터열들을 모두 수집하여 그로부터 확률 분포를 구함
  - 모든 단어에 대한 벡터 열 수집 → 매우 힘든 작업 → 단어를 소리의 기본 단위인 음소로 분리하여 모델링
  - 각각의 음소는 은닉 마르코프 모델(Hidden Markov Model, 이하 HMM)로 표현
  - 시간 축 상에서 입력벡터들 간의 순서가 있기 때문에 각 음소 모델은 left-to-right HMM을 이용하여 모델링
  - 일반적으로 음소 모델은 세 개의 emitting state를 가짐
    - 이전 음소에서 현재 음소로의 천이구간, 안정구간, 현재 음소에서 다음 음소로의 천이구간을 모델링하기 위함
    - 음성 인식에서는 HMM에서의 각 음소 모델은 주로 세 개의 상태를 가지며 모델 간의 연결을 위해서 하나 씩의 입·출력 상태가 있음

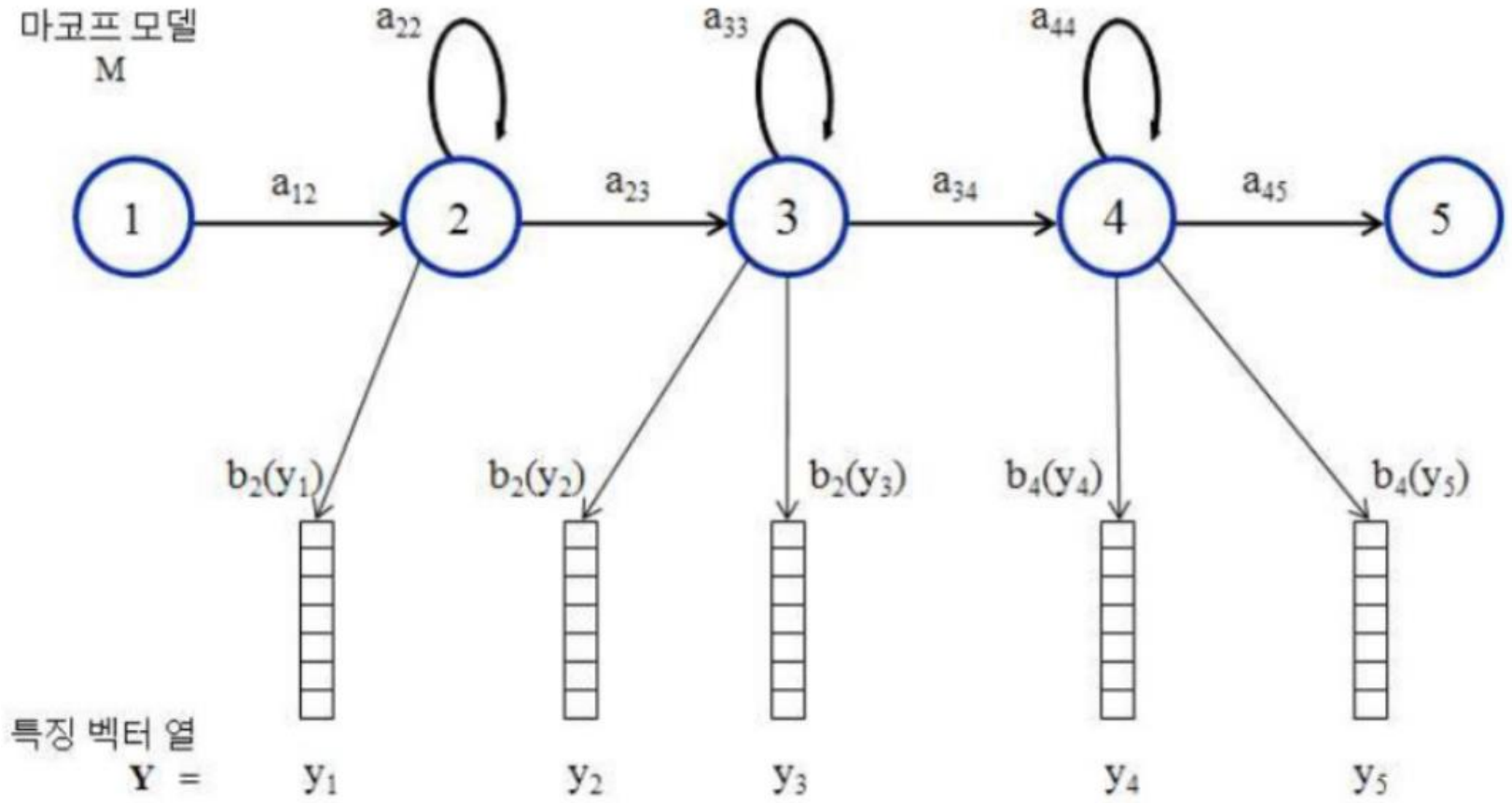
HMM 기반 음향 모델링

- **HMM 기반 3-state 음소 모델링 (tri-phone)의 예시**



HMM 기반 음향 모델링

- **음소 모델이 특징 벡터를 생성하는 예시**



HMM은 흔히 벡터 열 생성기라고 부름

특정 시각 t 에 특정 상태 j 로 상태가 전이 됨과 동시에 시각 t 에 해당하는 특징 벡터 y_i를 확률 밀도 $b_j(y_i)$ 의 값으로 생성해 냄
또한 상태 i 에서 상태 j 로 전이 되는 확률은 a_{kj}의 값을 가짐

HMM 기반 음향 모델링

- 출력함수: 멀티 믹스쳐 가우시안(multi mixture Gaussian) 확률 분포
  - HMM에서는 출력 확률을 구할 때 우리는 관찰 벡터 열 Y 만을 알 수 있고, 상태 열 X 는 알 수 없음
    - 상태열을 알 수 없다는 이유에서 은닉 마르코프 모델(HMM) 이라고 부름
  - 원하는 확률 값 $P(Y|M)$ 을 구하기 위해서 가능한 모든 상태열에 대해서 아래 수식을 계산하여 모두 더해주어야 함

$$P(Y, X|M) = a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(y_t) a_{x(t)x(t+1)}$$

  - 출력 함수인 멀티 믹스쳐 가우시안(multi mixture Gaussian) 확률 분포 수식

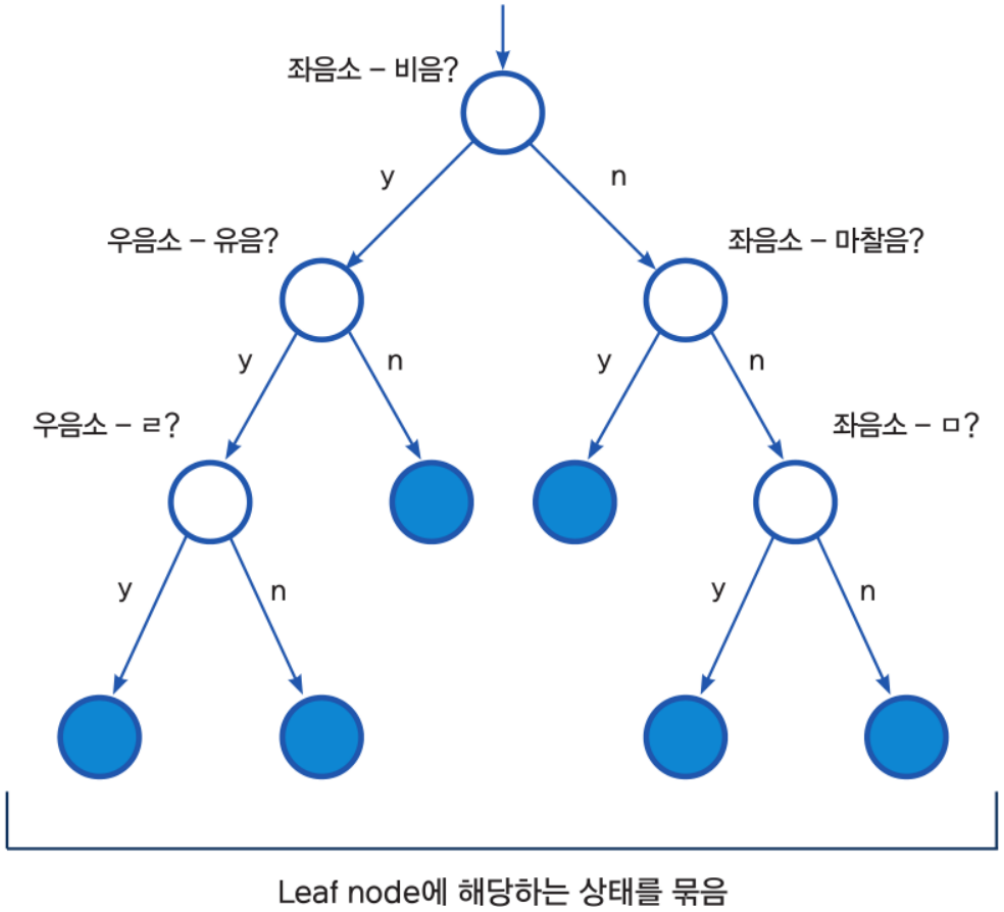
$$b_j(y_t) = \sum_{m=1}^{M} c_{jm} N(y_t; \mu_{jm}, \Sigma_{jm})$$

HMM 기반 음향 모델링

- **음소 결정 트리**
  - **많은 확률 분포를 학습하기 위해서 다양한 문맥을 고려한 대용량 음성자료가 필요함**
    - 예: 총 음소 모델의 개수가 40개라면 총 필요한 확률 분포는 192,000개가 필요
  - **이러한 문제를 해결하기 위해 트리구조를 이용하여 음운학적으로 유사한 트라이 폰들을 군집화하는 음소 결정 트리가 사용됨**
    - 음소 결정 트리: 각 노드마다 하나 씩의 질문이 달려 있는 이진 트리
  - **질문들과 트리 토폴로지는 트리에 의해서 생성된 상태를 가지고 학습했을 때 가장 높은 우도(likelihood)를 가지는 것으로 선택**

HMM 기반 음향 모델링

- 상태 군집화에 사용되는 결정 트리의 예



HMM 기반 음향 모델링

- 결정 트리를 생성하기 위해 필요한 요소
  - 질문집합 : 주어진 집합을 두 부분 집합으로 나누도록 하는 모든 이진 질문(binary question)들의 집합
  - 측정함수(measure function) : 매 노드에서 어떠한 질문을 선택할지 결정하는 기준으로, 데이터집합의 특성을 가장 잘 구분하는 질문을 선택하게 한다.
  - 정지기준(stopping criterion) : 각 노드에 해당하는 데이터 집합의 개수로 분할 정지 기준을 삼는다.

HMM 기반 음향 모델링

- **결정 트리 생성 알고리즘**

1. **주어진 모든 데이터에 대해 루트 노드에서부터 시작한다.**
2. **테스트되지 않은 노드가 존재하지 않을 때까지 다음을 진행한다.**
    1. 테스트되지 않은 노드를 선택
    2. 모든 가능한 질문들에 대해서 측정함수를 계산
    3. 현재 노드에서 가장 높은 측정 함수 값을 가지는 질문을 선택
        1. 이 값이 정지기준을 만족 하는 경우 현재 노드를 말단 노드로 하고 분할을 하지 않음
        2. 정지 기준을 만족하지 않는 경우 좌, 우의 자식 노드로 분할

HMM 기반 음향 모델링

- **한국어 음소 모델 정의**

초성	음소모델	중성	음소모델	종성	음소모델
ㄱ	g	ㅏ	a	ㄱ	g
ㄲ	G	ㅐ	e	ㄲ	g
ㄴ	n	ㅑ	A	ㄱㅅ	g
ㄷ	d	ㅒ	E	ㄴ	n
ㄸ	D	ㅓ	v	ㄴㅈ	n
ㄹ	r	ㅔ	e	ㄴㅎ	n
ㅁ	m	ㅕ	V	ㄷ	d
ㅂ	b	ㅖ	E	ㄹ	l
ㅃ	B	ㅗ	o	ㄹㄱ	l
ㅅ	s	ㅘ	R	ㄹㅁ	m
ㅆ	S	ㅙ	W	ㄹㅂ	b
ㅇ	없음	ㅚ	W	ㄹㅅ	d
ㅈ	j	ㅛ	y	ㄹㅌ	d

HMM 기반 음향 모델링

초성	음소모델	중성	음소모델	종성	음소모델
ㅉ	J	ㅜ	u	ㄹㅍ	b
ㅊ	c	ㅝ	O	ㄹㅎ	l
ㅋ	k	ㅞ	W	ㅁ	m
ㅌ	t	ㅟ	w	ㅂ	b
ㅍ	p	ㅠ	Y	ㅂㅅ	b
ㅎ	h	ㅡ	U	ㅅ	d
		ㅢ	I	ㅆ	d
		ㅣ	i	ㅇ	N
				ㅈ	d
				ㅊ	d
				ㅋ	g
				ㅌ	d
				ㅍ	b
				ㅎ	d