본 챕터에서는 지금까지의 챕터에서 설명한 각종 기법을 취합하여 대어휘 연속 음성 인식 (large vocabulary continuous speech recognition; LVCSR) 엔진을 구성하는 방법에 대해 설명한다.

0.1 FST의 합성과 확률모델

지금까지 도입했던 음성인인식의 요소를 나타내는 FST를 합성하고, 합성한 FST의 최단 경로 문제의 풀이로서 음성인식결과를 얻는 방법을 고찰한다. 음성인식의 통계모델을 표현하는 FST는 일반적으로는 아래의 4 종류가 있다.

- $\boldsymbol{G}$: 단어열 Acceptor (6.5)
- $\boldsymbol{L}$: 문맥에 의존하지 않는 음소열로부터 단어열 변환 (4.2.2)
- $\boldsymbol{C}$: 문맥에 의존하는 음소열로부터 문맥에 의존하지 않는 음소열로 변환 (5.3)
- $\boldsymbol{H}$: HMM state 시퀀스로부터 문맥에 의존하는 음소열로 변환 (5.3)

여기에 덧붙여서, 이하의 식에서 보여지고 있는 입력 (관측 벡터열)과 HMM 상태변수의 관계를 나타내는 FST $\boldsymbol{E}$를 가상으로 도입함으로써, 인식할 때의 처리를 모두 FST 형식으로 기술할 수 있게 된다.

$$\begin{aligned} &Q[\boldsymbol{E}] = \{0, \ldots, T\}, \qquad I[\boldsymbol{E}] = \{(0, \bar{1})\}, \qquad F[\boldsymbol{E}] = \{(T, \bar{1})\}, \\ &E[\boldsymbol{E}] = \{(t-1, t, \sigma, \sigma, -\log p(\mathbf{x_t}|s_t = \sigma)) : t \in 1, \ldots, T, \sigma \in \Sigma[\boldsymbol{H}]\} \end{aligned} \tag{1}$$

여기에서 T는 관측 벡터열의 길이를 의미한다.

음성인식 전체의 변환처리, 이른바 입력 프레임 시퀀스에서 단어열로의 변환은, 이 모든 FST를 합성한 $\boldsymbol{E} \circ \boldsymbol{H} \circ \boldsymbol{C} \circ \boldsymbol{L} \circ \boldsymbol{G}$로 나타낼 수 있다. 또한 그 FST의 변환 결과로써 가장 가중치가 작아지는 가설은 최단경로문제를 푸는 것으로써 근사적으로 구할 수 있다. 5.1에서 설명한 바와 같이, 이를 Viterbi 디코딩이라 한다. 그러나 많은 경우에, 이 FST는 거대하여, Viterbi 디코딩과 같은 근사적 해법으로도 풀기가 어렵다. 대어휘 연속 음성인식기술은, 이런 거대한 FST 위에서, Viterbi 디코딩을 가능케하는 기술이라 할 수 있다.