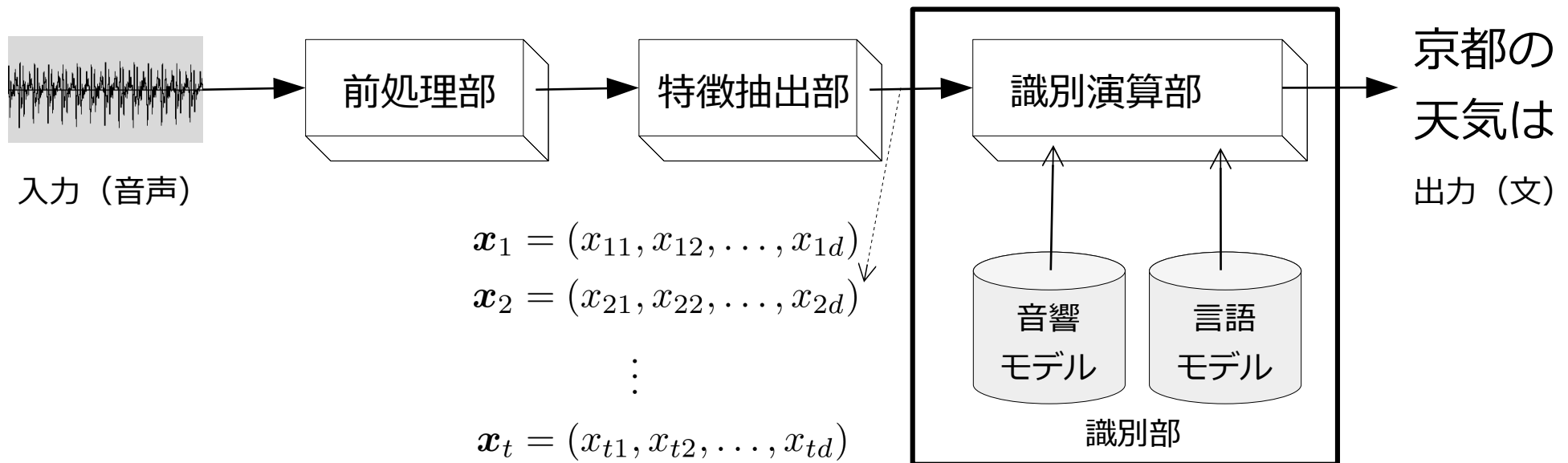


# 10. 声をモデル化してみよう

音響モデルの作り方・使い方・鍛え方

## 10.1 連続音声の認識

### ・ 連続音声認識システムの構成



# 10.1 連続音声の認識

- 統計的音声認識の定式化

- 入力系列  $x$  のもとで事後確率を最大にする単語列  $\hat{w}$  を認識結果とする

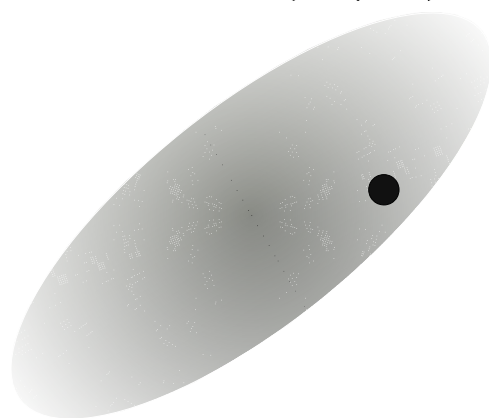
$$\begin{aligned}\hat{w} &= \arg \max_w P(w|x) \\ &= \arg \max_w p(x|w)P(w)\end{aligned}$$

- 音響モデル  $p(x|w)$
- 言語モデル  $P(w)$

# 10.1 連続音声の認識

- 音響モデル

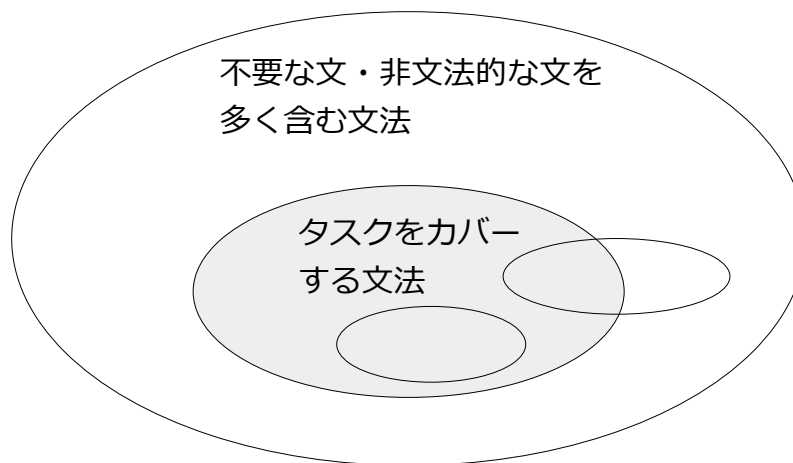
- $x$  が  $d$  次元ベクトル、 $w$  が単一のクラス  $w$  の場合、確率密度関数から  $p(x|w)$  が得られる



- $x, w$  とともに系列の場合、 $p(x|w)$  をどのようにしてもとめればよいか

# 10.1 連続音声の認識

- 文法規則を記述
  - 例 ) 文  $\rightarrow$  名詞 + 助詞 + 動詞
  - 規則に従う単語列は正の確率
  - 規則に従わない単語列は確率 0
- 問題点
  - 対象としている文集合をうまくカバーする規則を書くのは難しい



# 10.1 連続音声の認識


- 言語モデル  $P(\boldsymbol{w})$

$$\begin{aligned} P(\boldsymbol{w}) &= P(w_1, \dots, w_n) \\ &= P(w_1)P(w_2|w_1)P(w_3|w_1, w_2) \\ &\quad \dots P(w_n|w_1, \dots, w_{n-1}) \end{aligned}$$

- 問題点

- 条件部が長くなれば、そのような条件付き確率を統計的に求めるのは不可能
- 出現確率 0 の系列の問題
  - 例 ) トランプ大統領

# 10.1 連続音声の認識

- 音声認識における探索
  - $p(x|w)P(w)$  が最大となる  $w$  を求める
    - $w$  がクラスの場合
      - 全クラスに対して  $p(x|w)P(w)$  を計算すればよい
    - $w$  が自然言語文 ( 単語列 = クラス列 ) の場合
      - 人間が算出できる文は無限であるので、全ての場合について事後確率を求めることはできない  探索を用いる
- 問題点
  - 探索候補数の爆発
  - 探索の並列性

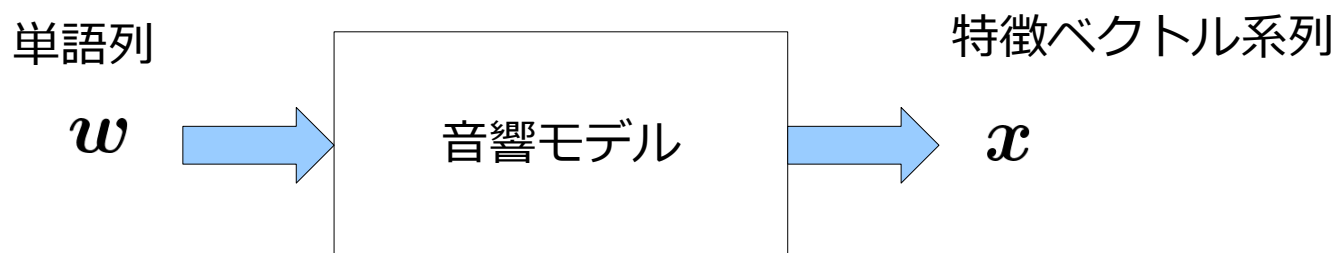
# 10.1 連続音声の認識

- この講義で説明する解法
  - 音響モデル  $p(x|w)$ 
    - 隠れマルコフモデル (HMM)
  - 言語モデル  $P(w)$ 
    - n-gram 近似 + 補間法
  - 事後確率最大となる  $\hat{w} = \arg \max_w p(x|w)P(w)$ 
    - ヒューリスティック探索
    - WFST

最新手法はニューラルネットを用いた End-to-End 方式

## 10.2 音響モデルの作り方

- 音響モデル  $p(x|w)$  とは
  - $p(\text{特徴ベクトル系列} \mid \text{単語列})$  を計算するための確率モデル



- まず、単純化のために単語認識問題を扱う
  - 単語は音素の系列で表現されているとする



## 10.2 音響モデルの作り方

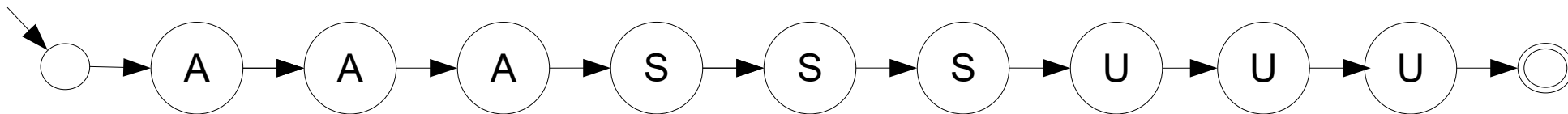
- 設定 1

- 各音素あたりの特徴ベクトル数が一定
- 特徴ベクトルを離散値（記号）で近似したときに誤りがない

→ 単語ごとの有限状態オートマトンでモデル化

- 受理すれば  $p > 0$ , 不受理ならば  $p = 0$

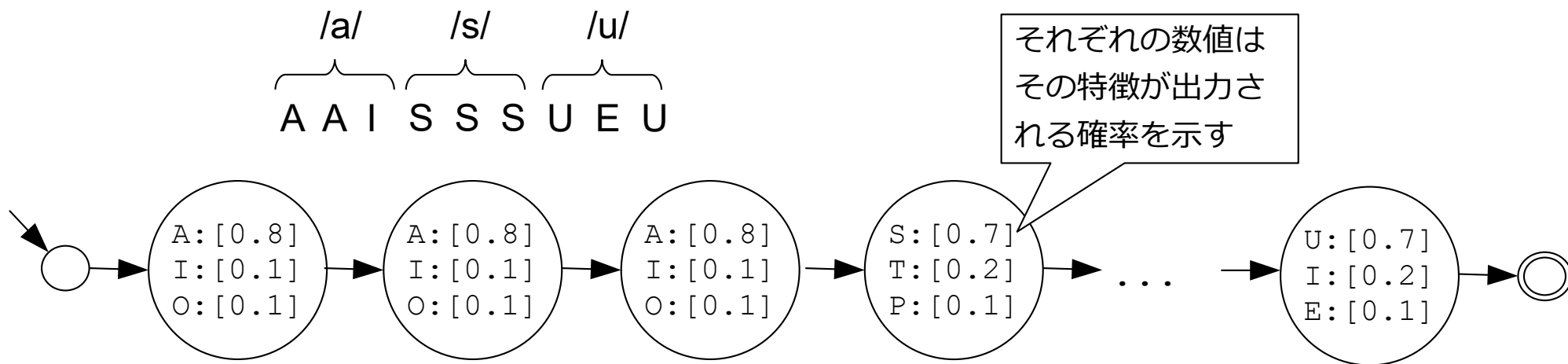
$/a/$        $/s/$        $/u/$       ← 単語「あす」の音素列  
A A A S S S U U U      ← 観測される特徴ベクトル系列の例



## 10.2 音響モデルの作り方

### • 設定 2

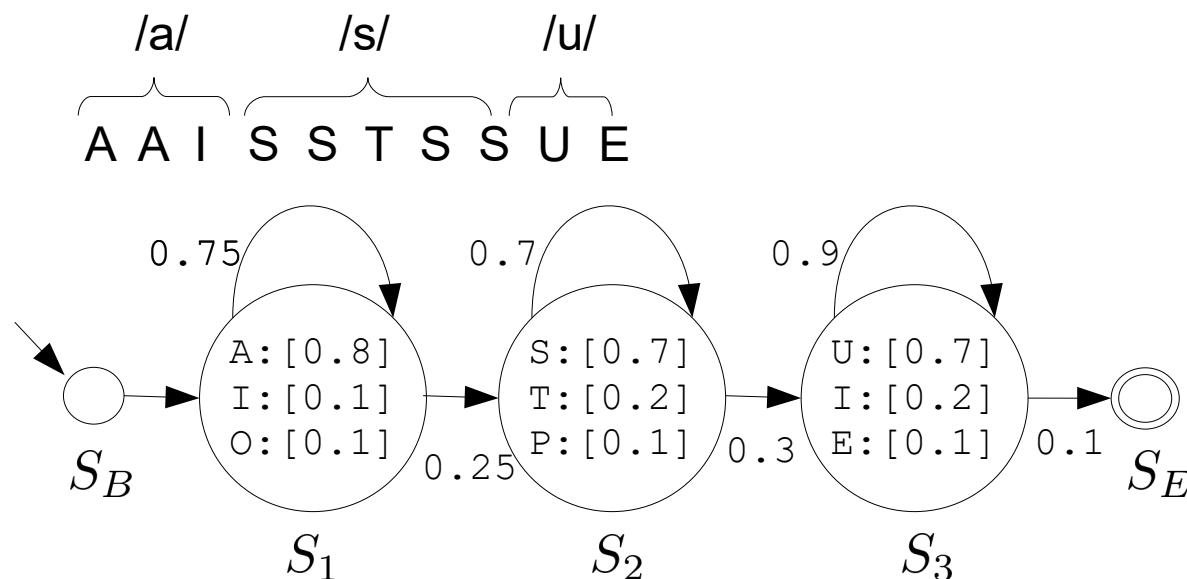
- 各音素あたりの特徴ベクトル数が一定
  - 特徴ベクトルの近似に誤りがあり得る
    - 単語ごとの確率オートマトンでモデル化
    - 各状態で、全てのシンボルに何らかの生成確率を与える
- $p =$  各状態における記号の生成確率の積



## 10.2 音響モデルの作り方

### • 設定 3

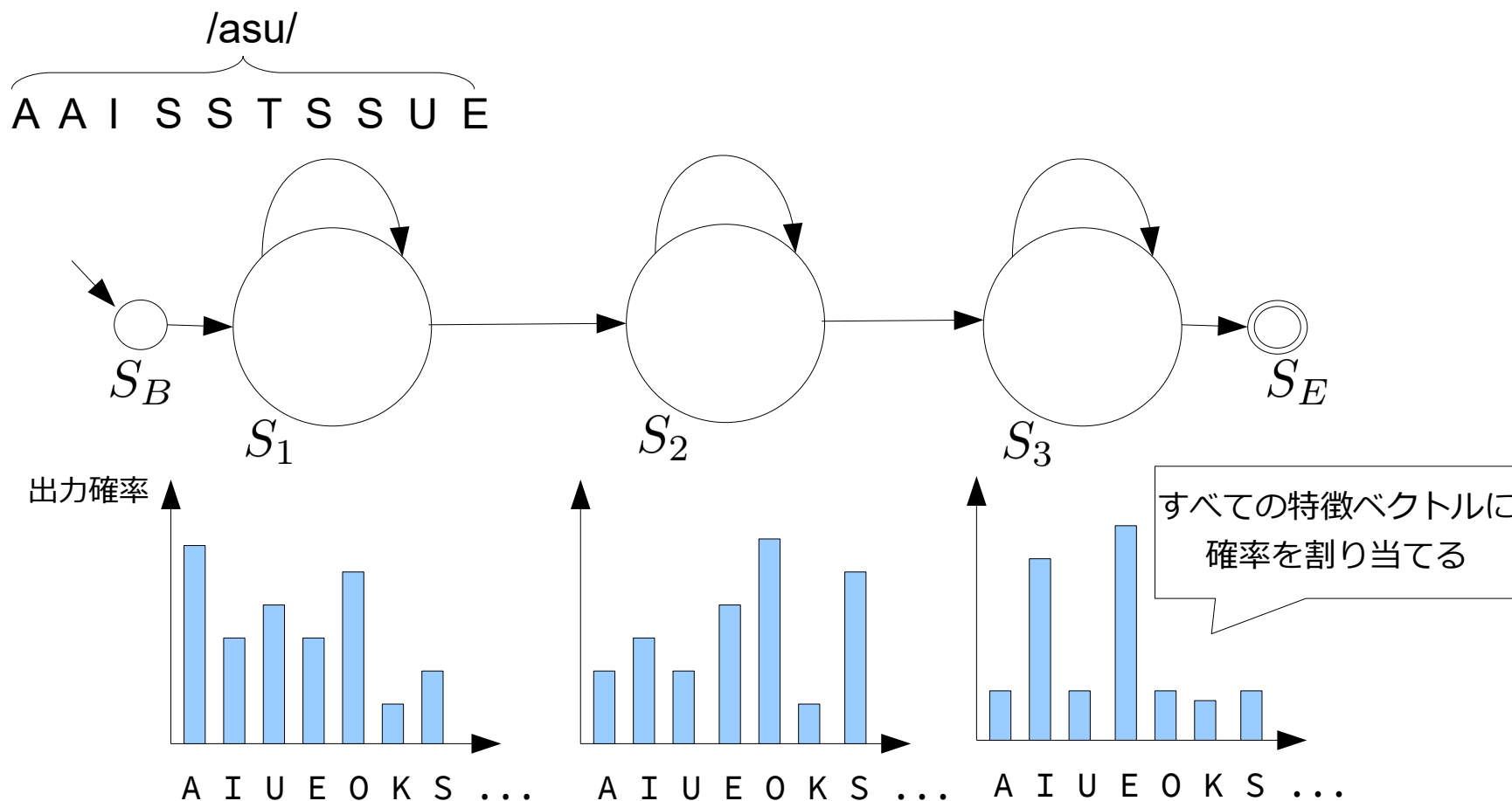
- 各音素あたりの特徴ベクトル数が不定
  - 特徴ベクトルの近似に誤りがあり得る
    - 非決定性確率オートマトン (=HMM) でモデル化
    - 各状態からの遷移が非決定的かつ確率的
- $p =$  「各状態における記号の生成確率と遷移確率の積」  
の可能な遷移に対する和



## 10.2 音響モデルの作り方

### • 設定 4

- 各状態ですべての特徴ベクトルに対して正の確率を割り当てる → 状態遷移情報が隠れてしまう



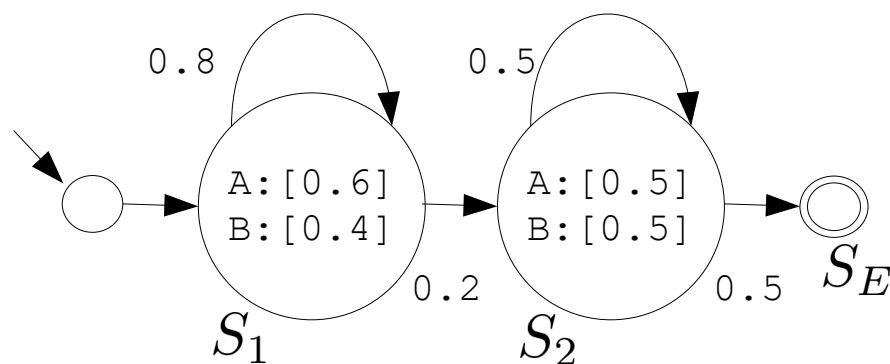
## 10.3 音響モデルの使い方

- HMM における確率計算
  - 音声認識用の HMM の構成
    - 状態で信号を出力するものとする
    - 信号を出力しない初期状態と終了状態を持つ
  - 正確には、可能な全ての遷移系列に対応する確率の和を計算
  - 実際には、各時点での最大確率のみを掛け合わせるビタビアルゴリズムで近似

# 10.3.1 HMM における確率計算

- 例題 10.1 より

- この HMM が系列 AAB を出力する確率



経路は以下の 2 通り

A	A	B
$S_1$	$S_1$	$S_2$
$S_1$	$S_2$	$S_2$

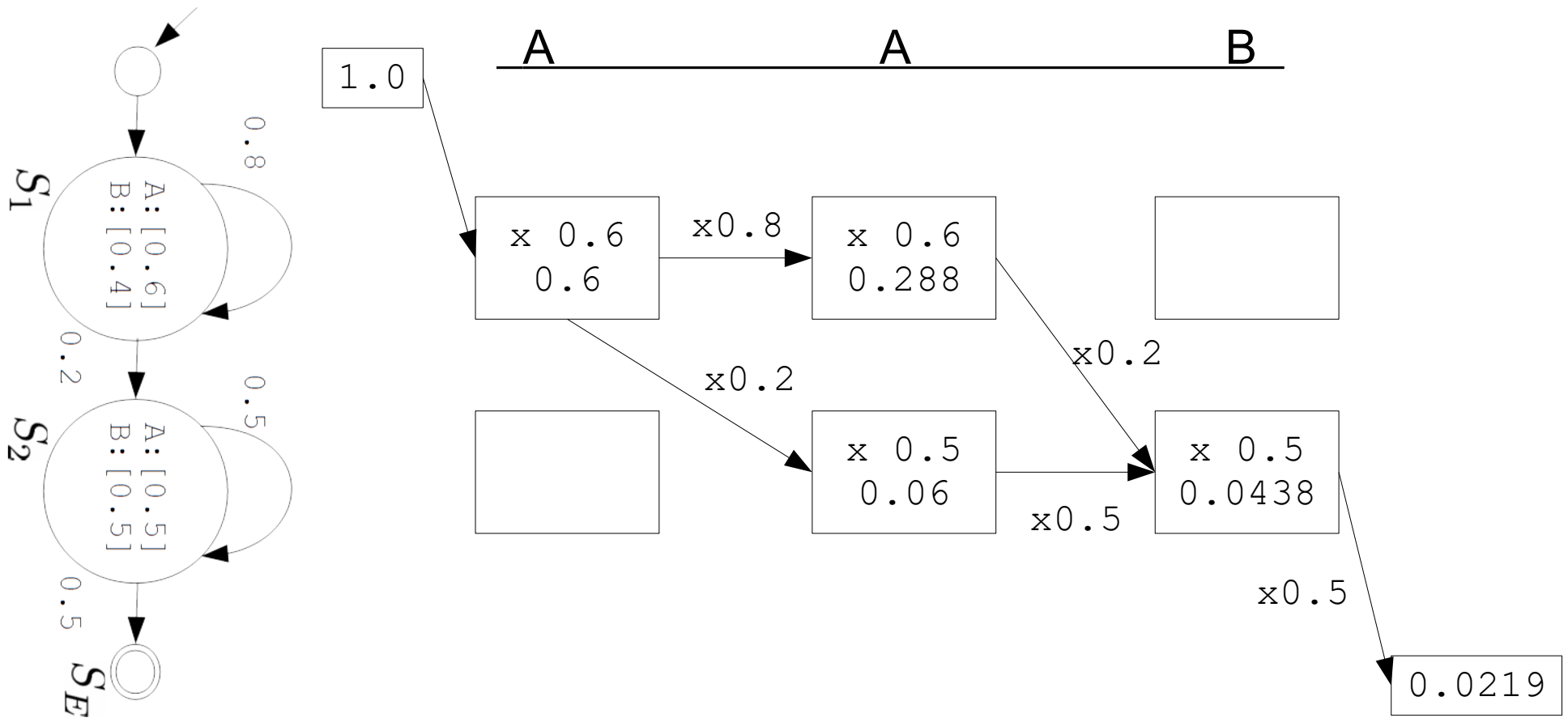
$$p(A|S_1) \cdot p(S_1 \rightarrow S_1) \cdot p(A|S_1) \cdot p(S_1 \rightarrow S_2) \cdot p(B|S_2) \cdot p(S_2 \rightarrow S_E) \\ = 0.6 \times 0.8 \times 0.6 \times 0.2 \times 0.5 \times 0.5 = 0.0144$$

$$p(A|S_1) \cdot p(S_1 \rightarrow S_2) \cdot p(A|S_2) \cdot p(S_2 \rightarrow S_2) \cdot p(B|S_2) \cdot p(S_2 \rightarrow S_E) \\ = 0.6 \times 0.2 \times 0.5 \times 0.5 \times 0.5 \times 0.5 = 0.0075$$

$$p(AAB|\omega_1) = 0.0144 + 0.0075 = 0.0219$$

## 10.3.2 トレリスによる効率のよい計算

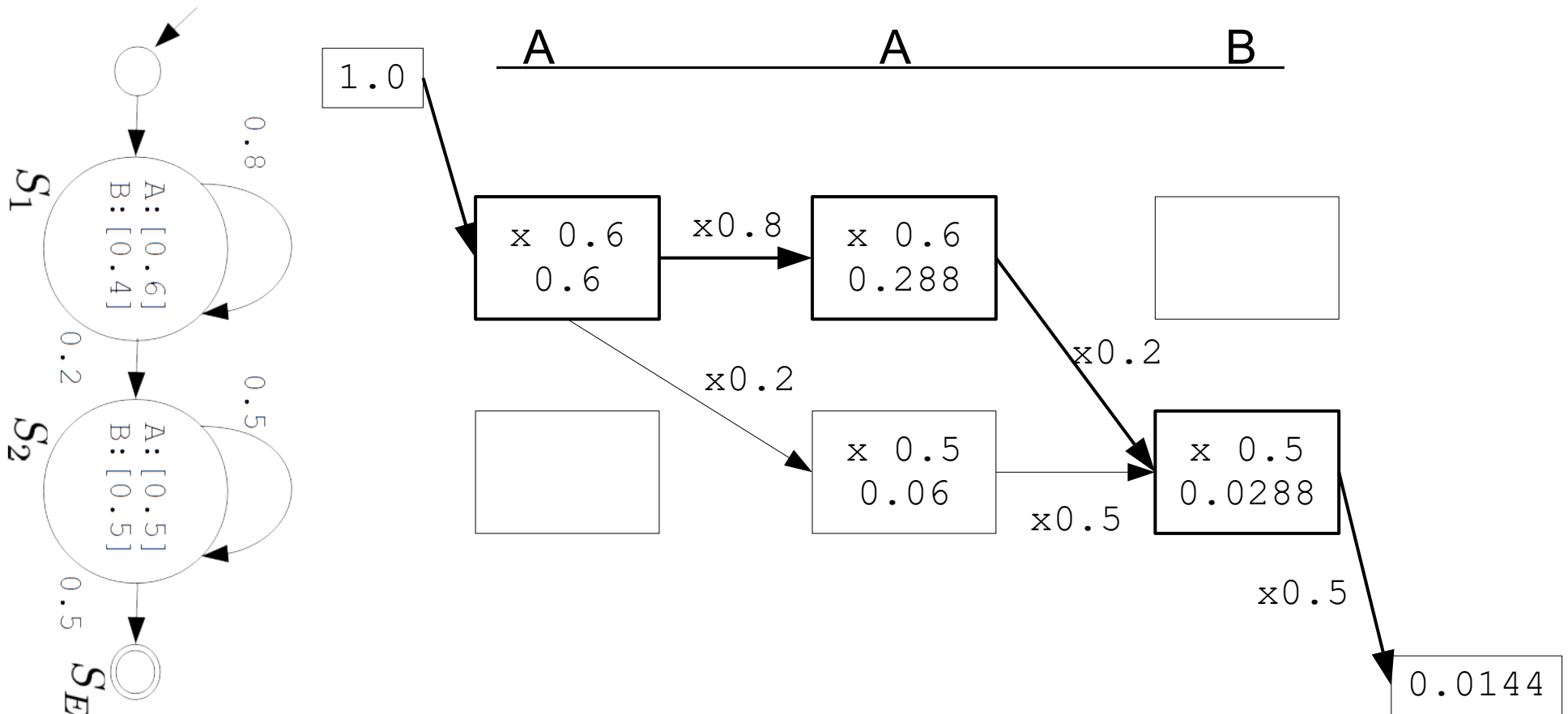
- トレリス計算
  - 共通する計算をまとめる



## 10.3.3 ビタビアルゴリズムによる近似計算

- ビタビアルゴリズム

- 最大値のみ保存して確率を計算
- 最適解の経路が確率計算と同時に求まる



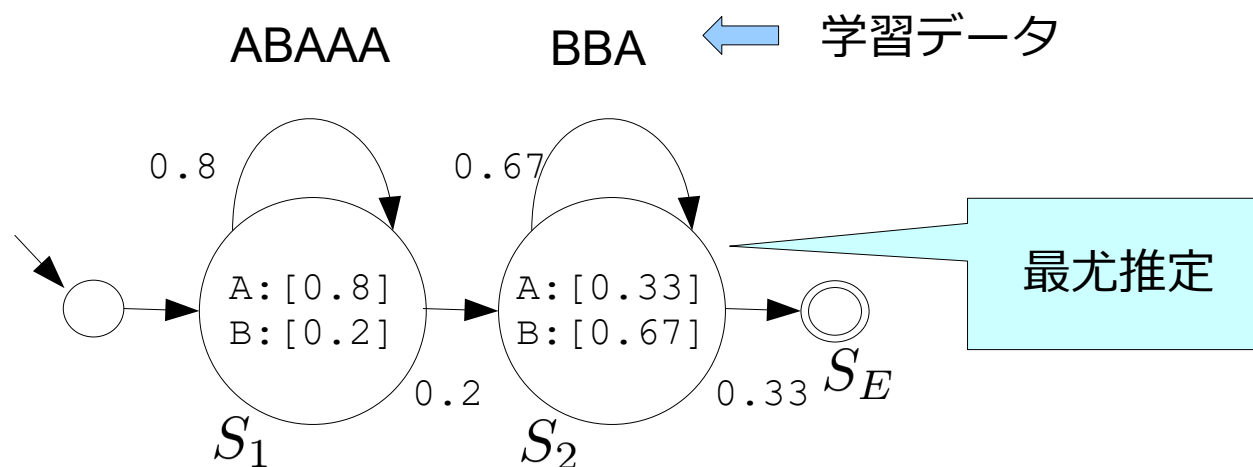


## 10.4 音響モデルの鍛え方

- HMM の学習
  - パラメトリックな学習
  - 確率密度関数の平均と分散、状態遷移確率を学習
- 学習における問題点
  - 学習データに対して状態遷移系列がわからない

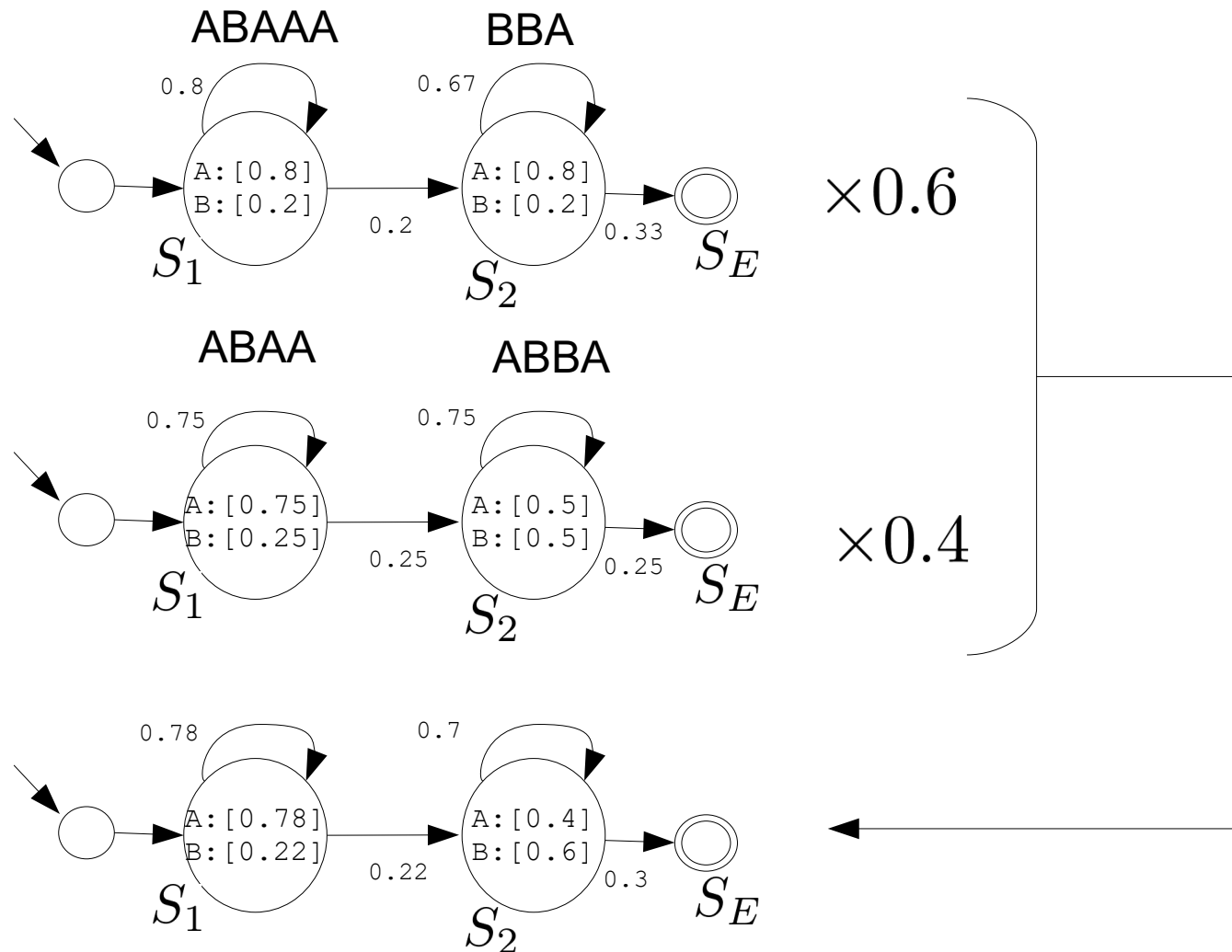
## 10.4 音響モデルの鍛え方

- 状態遷移系列が既知であれば
  - 状態遷移確率
    - 状態からの遷移を数え上げることによって学習可能
  - 信号出力確率
    - 状態ごとに平均・分散を計算することで学習可能



## 10.4 音響モデルの鍛え方

- 状態遷移系列の確率がわかっている
  - 学習結果の重み付き加算

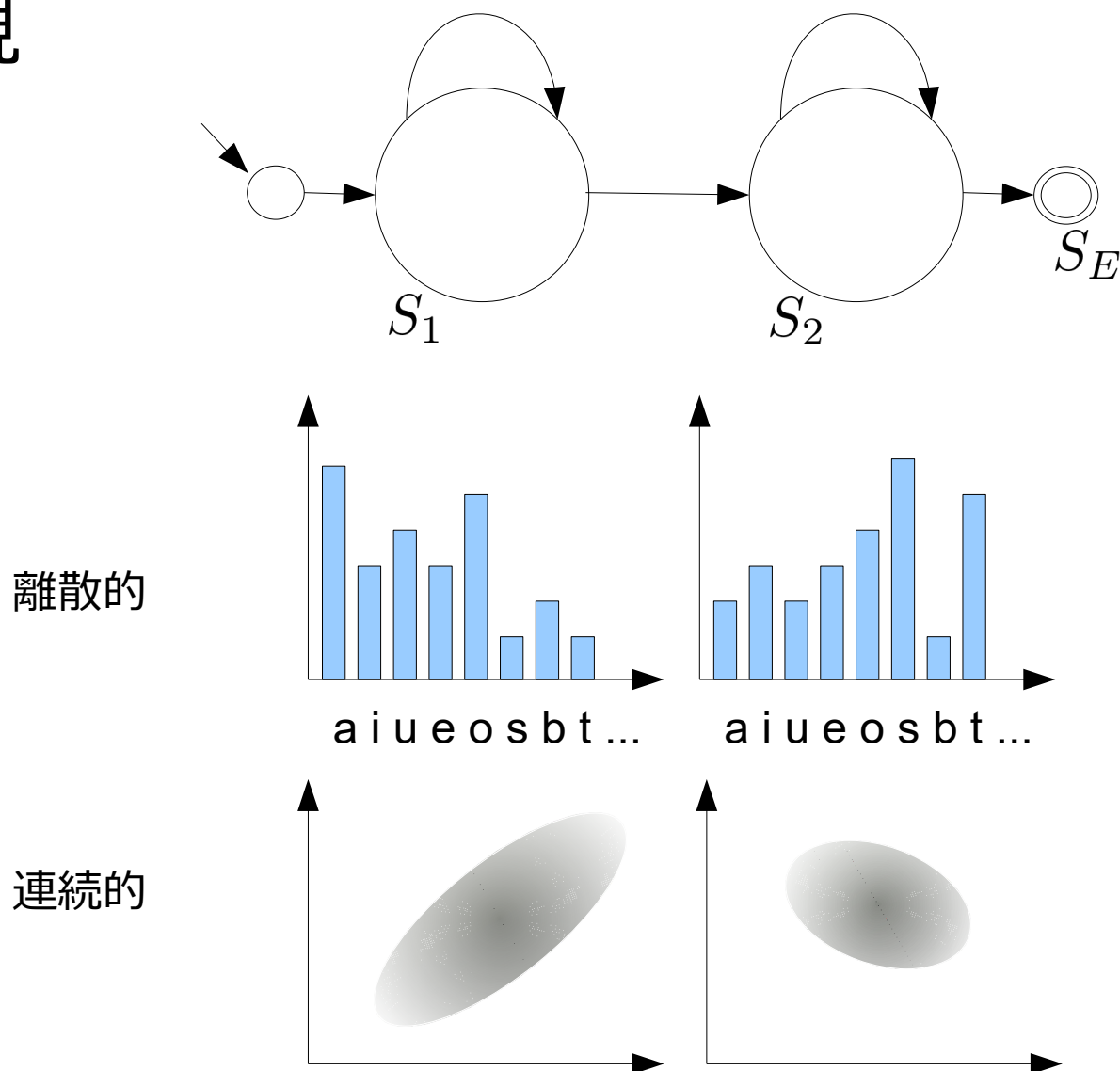


## 10.4 音響モデルの鍛え方

- Baum-Welch 法による HMM の学習
  - HMM のパラメータを適当な初期値に設定
  - E(Expectation) ステップ
    - 学習データ ( 入力 ) に対して、状態遷移を与えたときの確率を現在の HMM を用いて計算
    - それを全ての可能な状態遷移について求める ( 実際は動的計画法を用いて効率的に計算 )
  - M(Maximization) ステップ
    - E ステップで得られたデータから HMM のパラメータを最尤推定
  - E,M ステップをパラメータの変化量が一定値以下になるまで繰り返し

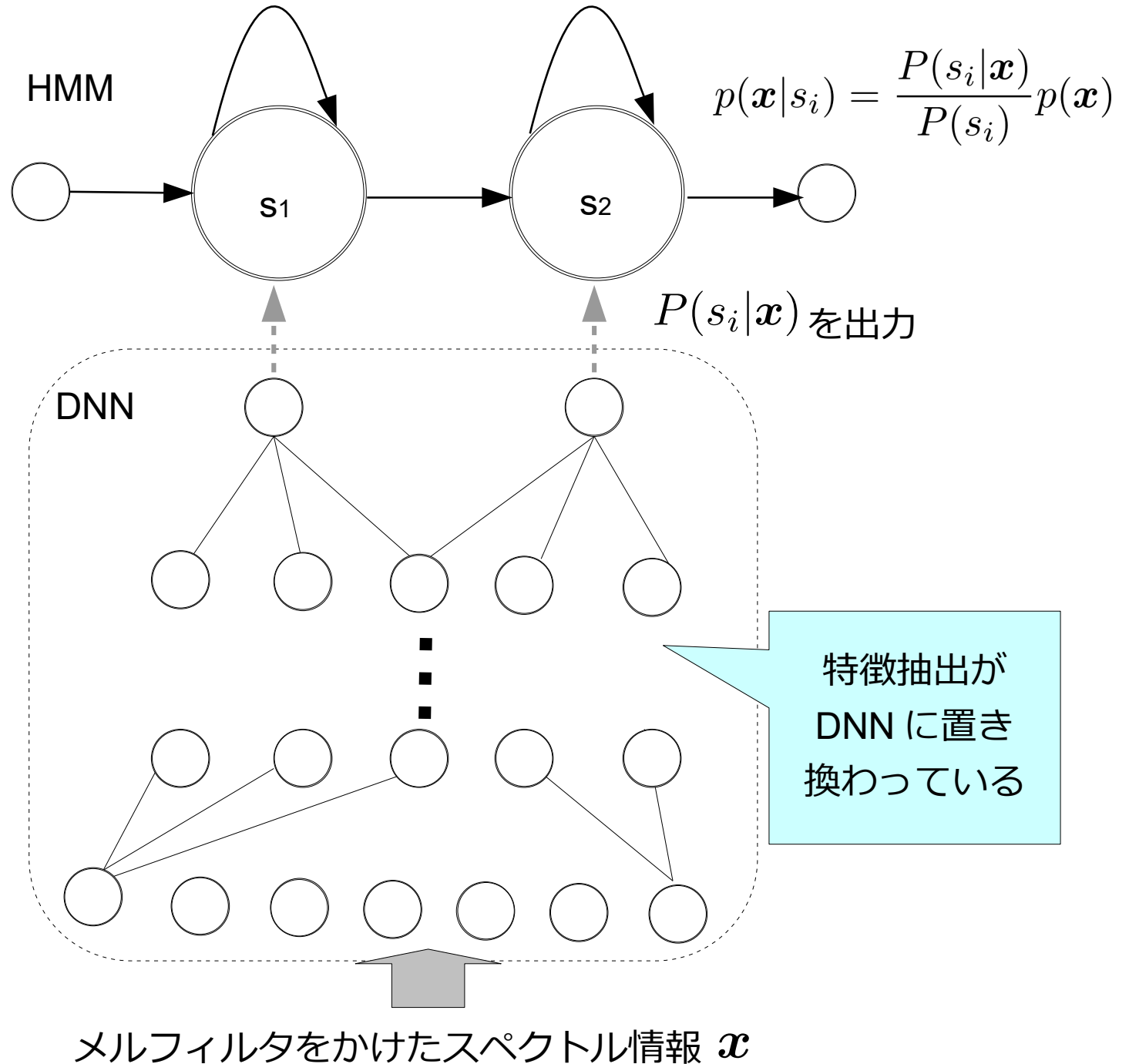
## 10.5 実際の音響モデル

- 各状態の出力確率は確率密度関数（連続値）で表現



# ディープニューラルネットを用いた音声認識

DNN-HMM 法



# 10.5 実際の音響モデル

- HMM での表現単位
  - 音素 (monophone)
    - 日本語の場合： 23 種類
  - 文脈依存音素 (triphone)
    - 前後の音による変化を捉えた高精度な音響モデル
    - 日本語の場合： 約 8000 種類
  - 単語は音素単位の HMM を繋いで表現
    - 無声化などの現象に注意 (例： ashta)

## 10.5 実際の音響モデル

- 音素文脈の考慮
  - 音素は前後の音素の影響を受けて大きく変化する  
( 調音結合 )
  - 前後の音素を考慮した 3 つ組音素 ( トライフォン )  
を音素単位とすると性能が向上する。
    - テキスト : あらゆる現実を ...
    - 音素系列 : arayurugeN...
    - トライフォン系列 :  
a+r a-r+a r-a+y a-y+u y-u+r u-r+u r-u+g ...