# 第4章　統計的音声認識

## 4.1　枠組み

音声認識システムでは，音声による通信を，話者のメッセージ $M$ が記号列 $\boldsymbol{W}$ に符号化され，それが音声波形 $\boldsymbol{S}$ に形を変えて通信路を経由し，聴き手に $\tilde{\boldsymbol{S}}$ として伝わり，$\tilde{\boldsymbol{S}}$ の特徴分析で得られた観測時系列 $\boldsymbol{O}$ から記号列 $\tilde{\boldsymbol{W}}$ を復元し，その記号列から発話者のメッセージ $\tilde{M}$ を推定する過程と捉えます（図 4.1).

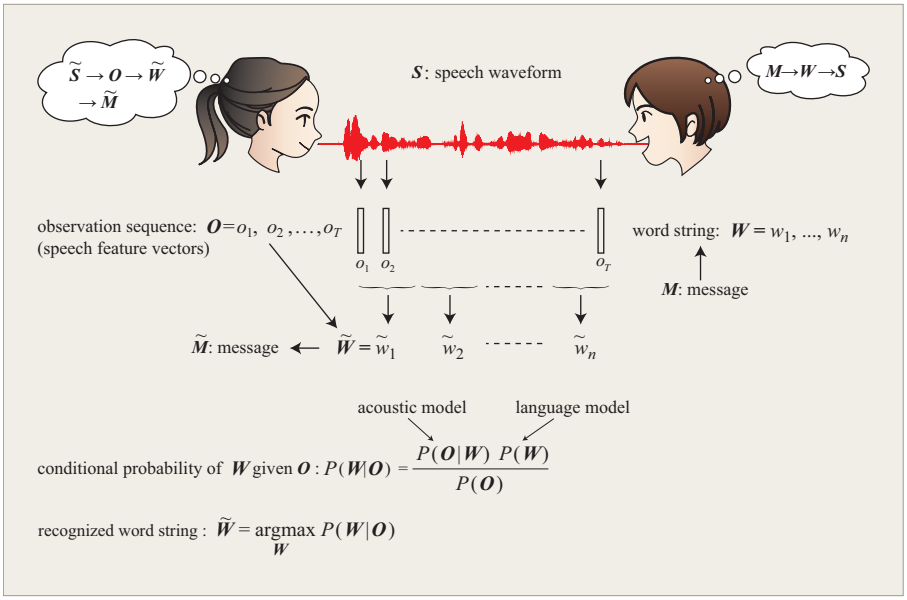


図 4.1: 音声通信の統計的解釈．$\boldsymbol{M}$：メッセージ（話者が伝えたい内容），$\boldsymbol{W}$：単語列（発話），$\boldsymbol{S}$：音声波形，$\tilde{\boldsymbol{S}}$：聴き手が受け取った音声波形，$\boldsymbol{O}$：観測系列（聴き手の受け取った音声特徴の系列），$\tilde{\boldsymbol{W}}$：推定単語列，$\tilde{\boldsymbol{M}}$：推定メッセージ

音声認識の究極の目的は，音声波形 $\tilde{\boldsymbol{S}}$ から話者のメッセージ $\boldsymbol{M}$ の推定値 $\tilde{\boldsymbol{M}}$ を得ることであるともいえます．たとえば，部屋に人間とロボットが居るとして，人間が室温を下げて欲しいと思い，ロボットに「暑いですね.」と話したとします．その発話を聴いてロボットがエアコンの温度を調節して室温を下げてくれたならば，ロボットは人間のメッセージ（意図）を理解したということができます．ロボットにこのような対応をさせるためには，単に聴いた音声波形を「暑いですね.」という文に変換することができるだけではダメです．発話の字面の背後にある意図を理解することが必要であり，**音声理解（speech understanding）**と呼ばれています．現在，一般に音声認識と呼ばれている技術は，意図理解までは行わず，話者の発した単語列 $\boldsymbol{W}$ の推定値 $\tilde{\boldsymbol{W}}$ を求めることを指しています．

## 4.2 基本原理

最大事後確率基準に基づく**ベイズ（Bayes）の識別規則**に基づいた音声認識の原理を説明します.

音響特徴量の系列を $\boldsymbol{O}$ とします. $\boldsymbol{O}$ は，$m$ 次元の実数値ベクトル $\boldsymbol{o}_t \in R^m$ の長さ $T$ の系列

$$\boldsymbol{O} \triangleq \boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T \tag{4.1}$$

とします．一般に，$\boldsymbol{O}$ は音声波形サンプル $\boldsymbol{X} = (x_1, \ldots, x_N)$ に対して短時間フレーム分析を行うことにより得ます. 発話された単語列を

$$\boldsymbol{W} \triangleq w_1, w_2, \ldots, w_n, \quad w_i \in \mathcal{W} \tag{4.2}$$

とします. $\boldsymbol{W}$ は，ある有限の語彙 $\mathcal{W}$ の要素 $w_i$ からなる長さ $n$ の単語列です.

$P(\boldsymbol{W}|\boldsymbol{O})$ は音響特徴量系列 $\boldsymbol{O}$ が与えられたとき単語列 $\boldsymbol{W}$ が発話された条件付き確率であり，**事後確率（posterior probability）**と呼ばれます．ベイズの定理により，$P(\boldsymbol{W}|\boldsymbol{O})$ は,

$$P(\boldsymbol{W}|\boldsymbol{O}) = \frac{P(\boldsymbol{O}|\boldsymbol{W})P(\boldsymbol{W})}{P(\boldsymbol{O})} \tag{4.3}$$

と書き直すことができます.

$P(\boldsymbol{W})$ は単語列 $\boldsymbol{W}$ の生起確率です．たとえば，単語列 $\boldsymbol{W}$ が 0.7 の確率で生起し，他の単語列をすべて合わせても 0.3 の確率でしか生起しなければ，どの観測データ $\boldsymbol{O}$ に対してもそれが単語列 $\boldsymbol{W}$ に属すると答えれば，0.7 の確率で正答となります．観測する前からわかっている確率という意味で，これを**事前確率（prior probability）**と呼びます.

$P(\boldsymbol{O}|\boldsymbol{W})$ は単語列 $\boldsymbol{W}$ が生起したという条件の下で音声特徴量時系列 $\boldsymbol{O}$ が観測される確率を表わしていて，**クラス条件付き確率（class conditional probability）**です．$\boldsymbol{O}$ が $\boldsymbol{W}$ に属しているのが尤もらしいと考えられる確率と解釈することができることから**尤度（likelihood）**とも呼ばれます[1].

$P(\boldsymbol{O})$ は $\boldsymbol{O}$ が観測される確率であり，$\boldsymbol{O}$ と $\boldsymbol{W}$ の同時確率 $P(\boldsymbol{O}, \boldsymbol{W})$ から

$$P(\boldsymbol{O}) = \sum_{\boldsymbol{W}} P(\boldsymbol{O}, \boldsymbol{W}) = \sum_{\boldsymbol{W}} P(\boldsymbol{O}|\boldsymbol{W})P(\boldsymbol{W}) \tag{4.4}$$

のように可能な単語列全体の関する和をとることで得られます．このような操作を**周辺化**といい，$P(\boldsymbol{O})$ を**周辺確率**と呼びます.

式 4.3 を眺めてみると，事後確率 $P(\boldsymbol{W}|\boldsymbol{O})$ は事前確率 $P(\boldsymbol{W})$ をクラス条件付き確率と周辺確率の比 $P(\boldsymbol{O}|\boldsymbol{W})/P(\boldsymbol{O})$ で修正したものとみることができることに気が付くでしょう．すなわち，単語列 $\boldsymbol{W}$ が与えられたときに観測値 $\boldsymbol{O}$ が得られる尤度 $P(\boldsymbol{O}|\boldsymbol{W})$ が，単語列 $\boldsymbol{W}$ が与えられない場合の確率 $P(\boldsymbol{O})$ よりも大きければ，事後確率は事前確率よりも大きくなり，そうでなければ事後確率が事前確率よりも小さくなります.

[1] ただし，尤度の和は $\sum_{\boldsymbol{W}} P(\boldsymbol{O}|\boldsymbol{W}) \neq 1$ なので，厳密な意味で確率とはいえません.

単語列 $\boldsymbol{W_i}$ と $\boldsymbol{W_j}$ の**識別境界（discrimination boundary）**は，事後確率が等しくなるところ，すなわち，

$$P(\boldsymbol{W_i}|\boldsymbol{O}) = \frac{P(\boldsymbol{O}|\boldsymbol{W_i})P(\boldsymbol{W_i})}{P(\boldsymbol{O})} = \frac{P(\boldsymbol{O}|\boldsymbol{W_j})P(\boldsymbol{W_j})}{P(\boldsymbol{O})} = P(\boldsymbol{W_j}|\boldsymbol{O}) \tag{4.5}$$

が成立するところです．式 4.5 からわかるように，周辺確率 $P(\boldsymbol{O})$ は単語列に共通に現れているので，識別規則に含める必要はありません．したがって，音声認識器は $P(\boldsymbol{O}|\boldsymbol{W})P(\boldsymbol{W})$ を最大化する単語列 $\tilde{\boldsymbol{W}}$，

$$\tilde{\boldsymbol{W}} \triangleq \underset{\boldsymbol{W}}{\operatorname{argmax}}\, P(\boldsymbol{O}|\boldsymbol{W})P(\boldsymbol{W}) \tag{4.6}$$

を求めることになります．$P(\boldsymbol{O}|\boldsymbol{W})$ を**音響モデル（acoustic model）**，$P(\boldsymbol{W})$ を**言語モデル（language model）**と言います．