日本語ディクテーションにおける基礎技術

Vxa-vxixeの裏側で何が行なわれているのか？

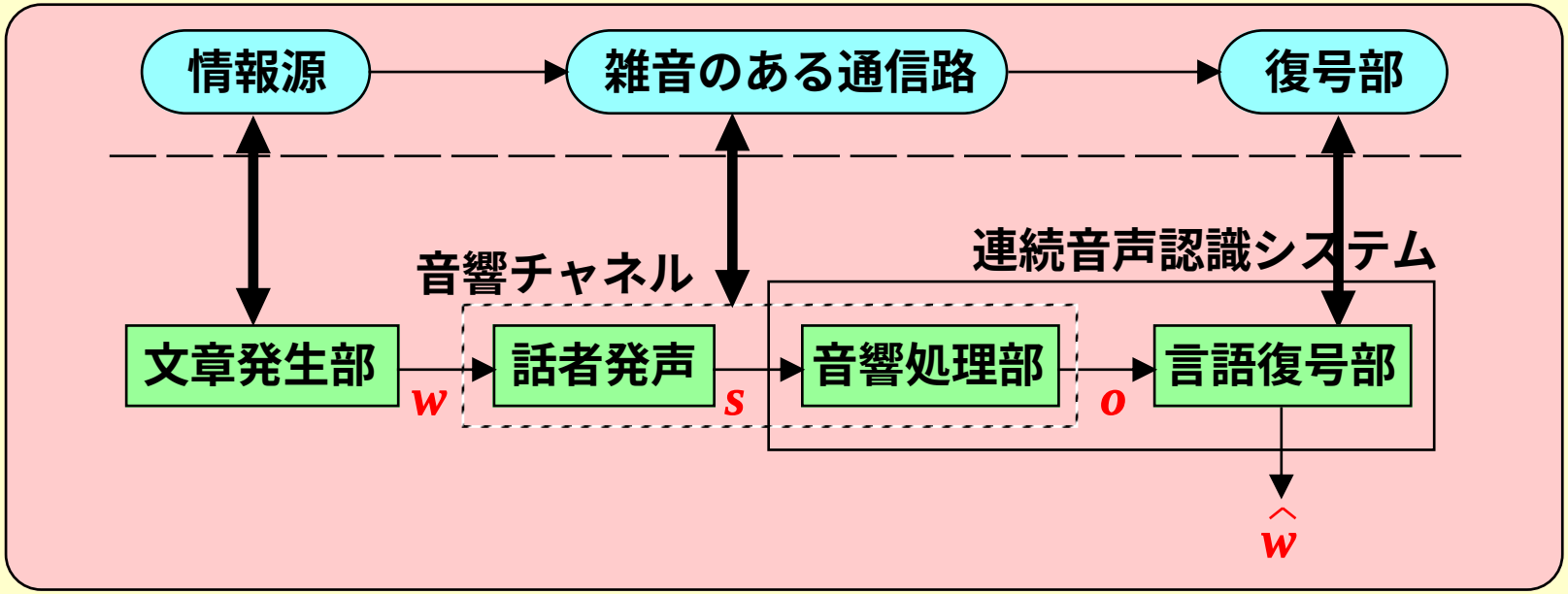
おおよそ

峯松 信明
(豊橋技術科学大学)

本講演の流れ

- 音声認識の情報理論的な枠組み
- 音声データベース / テキストデータベース
- 音響モデル / 「音」の統計的なモデリング
- 言語モデル / 「ことば」の統計的モデリング
- デコーダ / 2つのモデルと入力音声との照合，認識エンジン
- 現状の音声認識精度 / 種々の条件下での認識結果の比較
- まとめと今後の課題

音声認識の情報論的な定式化



❒ ***音声認識の情報論的な定式化***

✔ 音声認識 = $\arg\max_i P(w_i|o)$

✔ $P(w_i|o) = \dfrac{P(o|w_i)P(w_i)}{P(o)}$ ，　P(wi) : wi の事前確率，P(o) : o の事前確率

✔ 音声認識 = $\arg\max_i P(o|w_i)P(w_i)$

✔ P(wi) : wi の事前確率=**言語モデル**，P(o|wi) : wi が与えられた時の o の事後確率 = **音響モデル**

✔ **言モ** = テキストデータベースより，**音モ** = (書き起こし付きの)音声データベースより構築

音声データベース / テキストデータベース

❒ *言語モデルP(wi)　= あるドメインにおいて如何なる言語事象(単語)が出現し易いのか？*
- ✔ 以前は，規則で書くことが多かった(CFG)。

❒ *音響モデルP(o|wi) = ある言語事象を前提とすると，如何なる音響現象が観測され易いのか？*
- ✔ 以前は，テンプレートとして用意することが多かった(DPマッチング)。

❒ *数理統計的な枠組みで各モデルを構築する方式へ(N-gram & HMM)*
- ✔ まず，データベースありき，，，，，の時代
- ✔ テキストデータベース = 新聞データベース
- ✔ 音声データベース = 各研究機関が独自で収録 / 公開を目的として複数機関が協力して収録
- ✔ JNAS = { 男性150人，女性150人 } x { 新聞記事 100 文 + 音素バランス文 50 文 }
- ✔ IxM社では，一社で更に大きな音声データベースを持っているらしい，，，。

- ✔ The larger, the better.....

音響モデル --- HMM ---(#1)

❒ 音響モデリングの単位

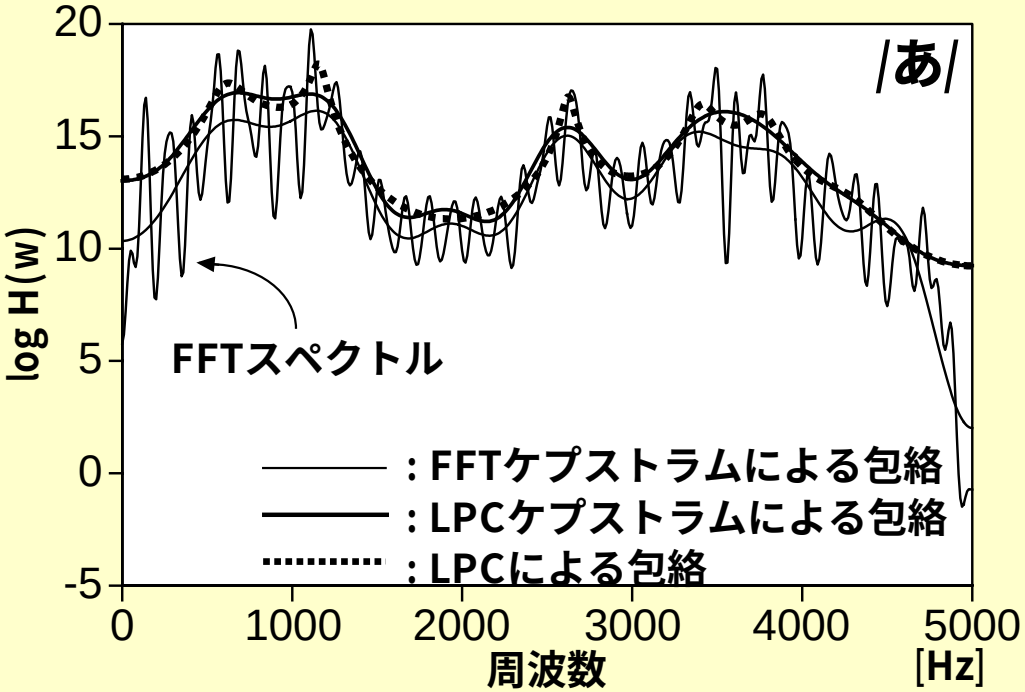
✔ 単語を単位とすると単位数が爆発 → 音素，音節を単位に

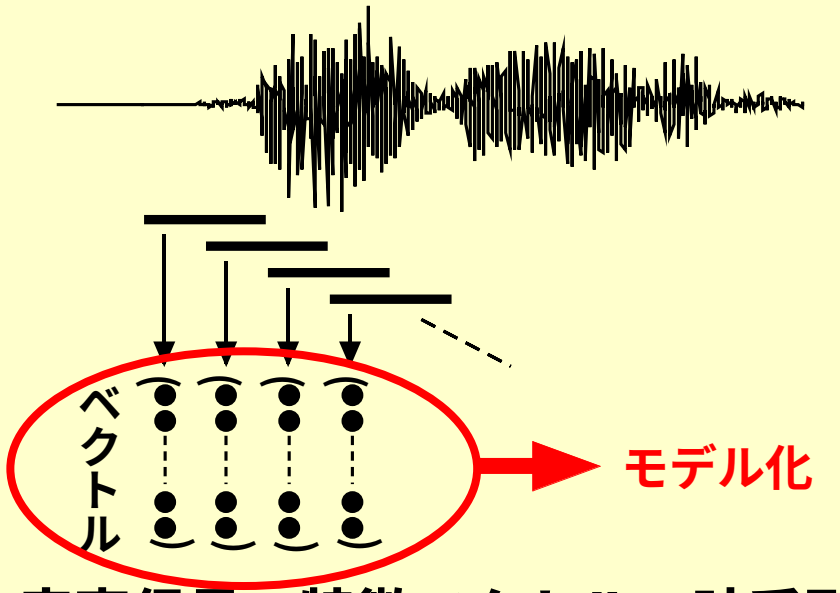
a i u e o a: i: u: e: o: N w y j my ky by gy dy ny hy ry py
p t k ts ch b d g z m n s sh h f r q sp silE silB

❒ 音響特徴パラメータ

✔ /あ/い/う/え/お/　の音響的差異はどこに観測されるのか？ → 音声のスペクトル包絡

✔ スペクトル包絡の情報をどう符号化するのか？ → ケプストラム





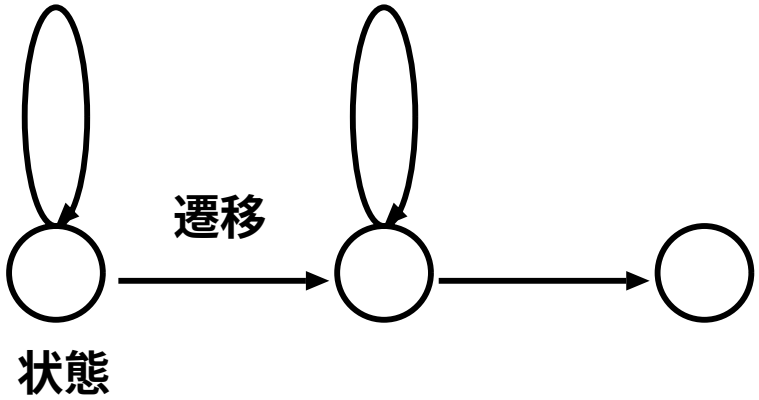
音声信号→特徴ベクトルの時系列

マルコフプロセス

$$P(x_n|x_{n-1},\cdots,x_1) = P(x_n|x_{n-1})$$

- 現在の信号が決定されれば過去の信号は未来の信号に影響を与えない
- 過去の全ての情報が現在の信号に集約されている

隠れマルコフモデル(HMM)



$$P\big(x_n|\ \underbrace{x_{n-1},\cdots,x_1}\ \big) = P\big(x_n|\ \underbrace{S_n}\ \big)$$

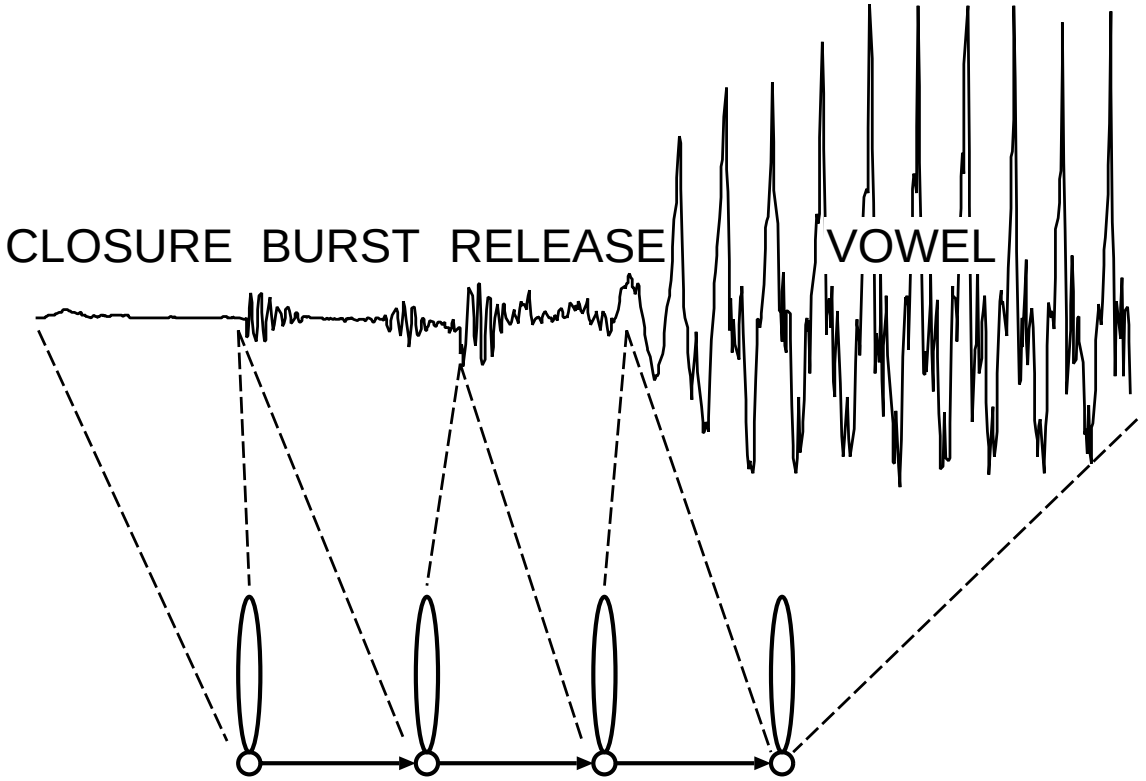
観測信号系列　　現在の状態

観測信号系列：$x_1, x_2, \cdots, x_n, \cdots$

状態系列：$S_1, S_2, \cdots, S_n, \cdots$

- 観測信号系列から現在の状態を決定できない
- 信号系列：観測可能、状態系列：隠れている (Hidden)

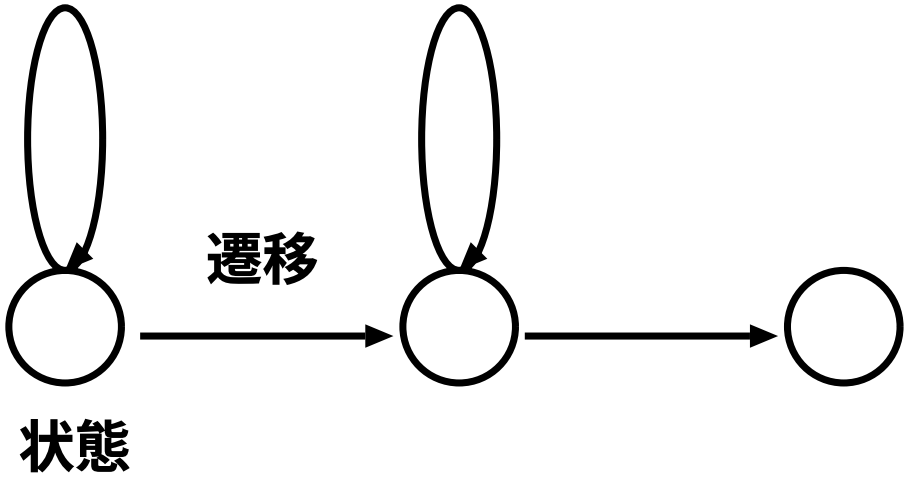
音声の生成モデルとしてのHMM



確率的生成モデル

状態間の境界 (遷移確率)　状態毎の出力信号 (出力確率)

HMM パラメーター



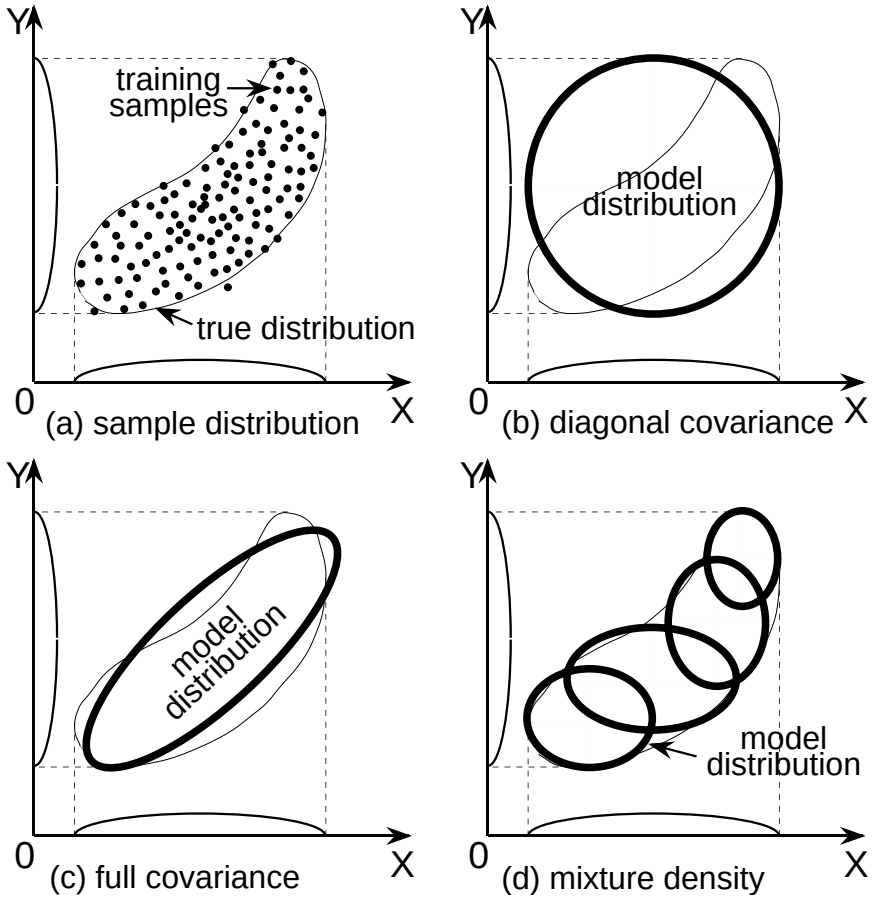
- **状態**から**状態**へ**遷移**する確率 (**遷移確率** a_{ij})
- **状態**からベクトルを**出力**する確率 (**出力確率** $b_i(o)$)

$$\alpha_j(t) = \sum_i \alpha_i(t-1) a_{ij} b_j(o_t) \quad (\textbf{前向き確率})$$

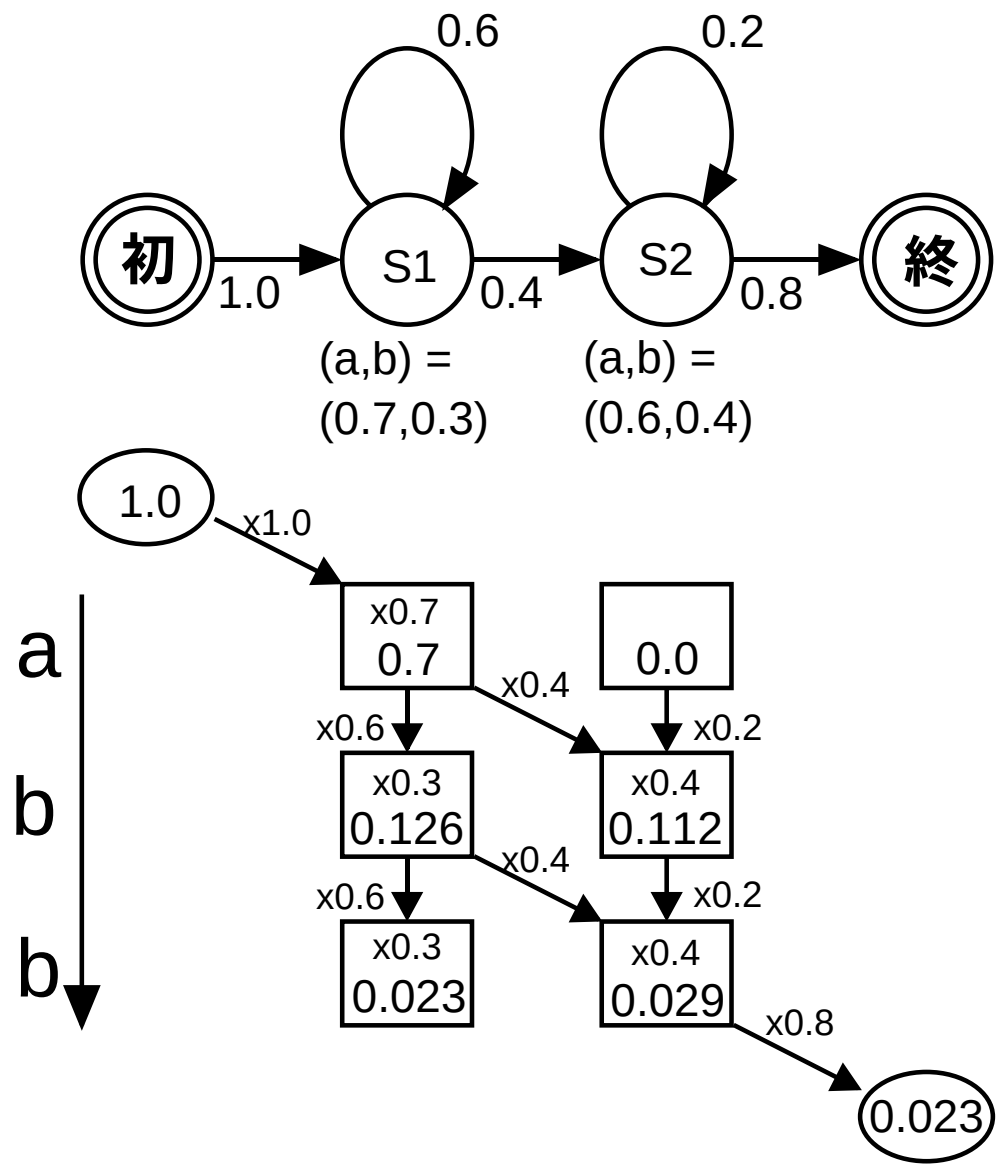
$$\beta_j(t) = \sum_i a_{ji} b_i(o_{t+1}) \beta_i(t+1) \quad (\textbf{後向き確率})$$

HMMの分類

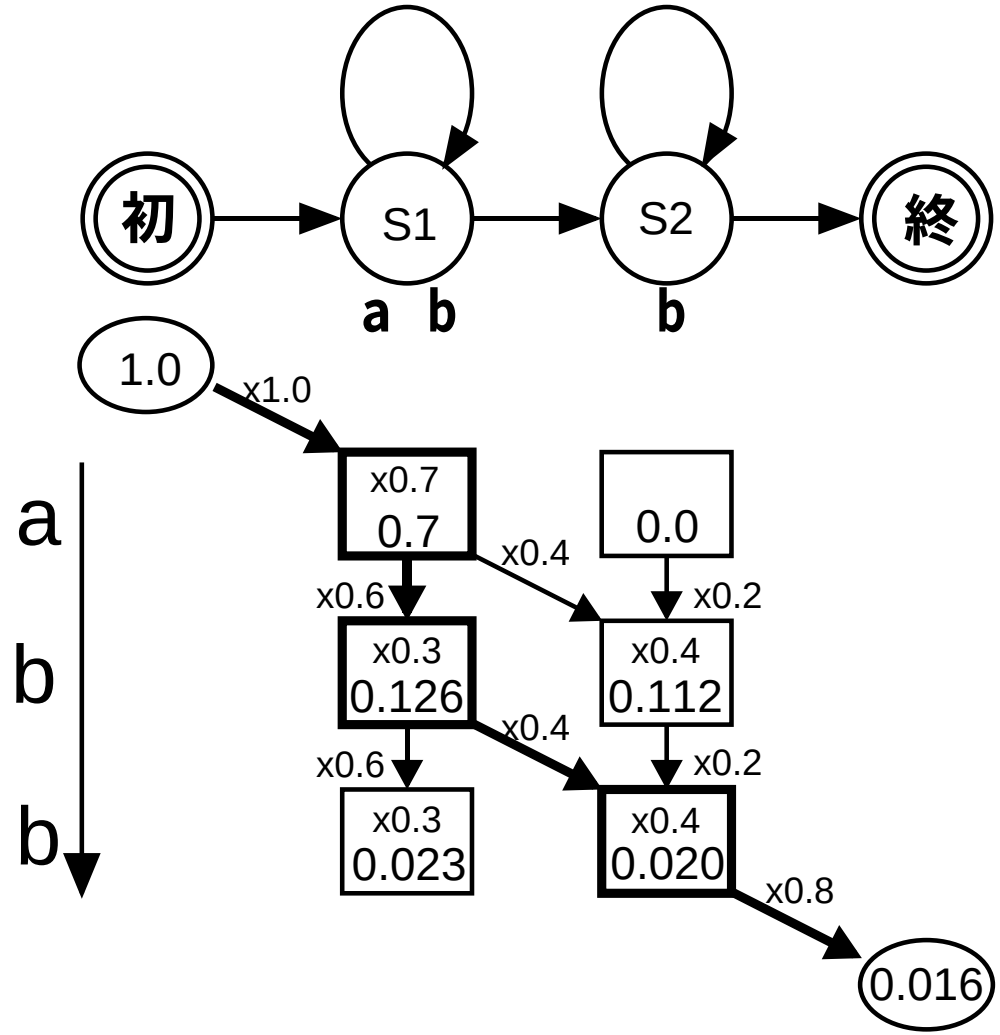
- 離散分布 HMM
- 連続分布 HMM
  - 単一分布型 HMM
    - 全角行列
    - (対角行列)
  - 混合分布型 HMM
    - (全角行列)
    - 対角行列
- 半連続 HMM



HMMの確率計算法(トレリス)



HMMの確率計算法(Viterbi)



最大の確率を与える経路のみを考慮する

EMアルゴリズムによるパラメータの推定1

- 前向確率

$$\alpha_j(t) = P(o_1, \cdots, o_t, s(t) = j|M) \qquad = \sum_i \alpha_i(t-1) a_{ij} b_j(o_t)$$

- 後向確率

$$\beta_j(t) = P(o_{t+1}, \cdots, o_T|s(t) = j, M) \quad = \sum_i a_{ji} b_i(o_{t+1}) \beta_i(t+1)$$

$$\rightarrow \quad \alpha_j(t)\beta_j(t) = P(O, s(t) = j|M)$$

$$\rightarrow \quad P(s(t) = j|O, M) = \frac{\alpha_j(t)\beta_j(t)}{P(O|M)} = \frac{\alpha_j(t)\beta_j(t)}{\alpha_N(T)} = L_j(t)$$

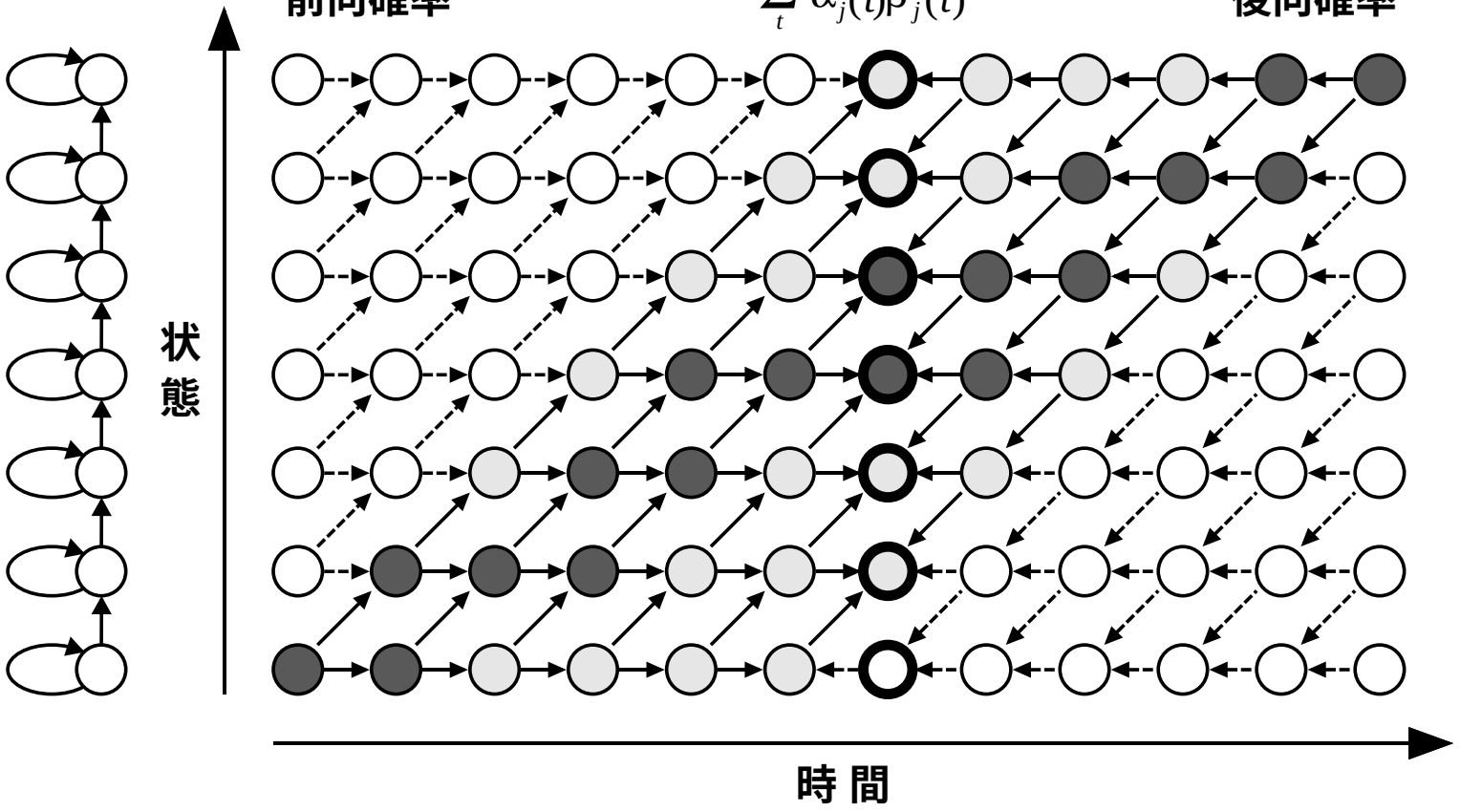
$\rightarrow$ ベクトル o_t と状態 j との「結び付き」の度合い

$$\rightarrow \quad \hat{\mu}_j = \frac{\sum_t L_j(t) \cdot o_t}{\sum_t L_j(t)} = \frac{\sum_t \alpha_j(t)\beta_j(t) \cdot o_t}{\sum_t \alpha_j(t)\beta_j(t)}$$

EMアルゴリズムによるパラメータの推定2

$$\mu_j = \frac{\sum_t \alpha_j(t)\beta_j(t)\, o_t}{\sum_t \alpha_j(t)\beta_j(t)}$$

$$\Sigma_j = \frac{\sum_t \alpha_j(t)\beta_j(t)(o_t - \mu_j)(o_t - \mu_j)^t}{\sum_t \alpha_j(t)\beta_j(t)}$$



EMアルゴリズムによるパラメータの推定3

- 学習データ数 $= 1$ 個

$$\hat{\mu}_j = \frac{\sum_t L_j(t) \cdot o_t}{\sum_t L_j(t)}, \quad \hat{\Sigma}_j = \frac{\sum_t L_j(t) \cdot (o_t - \mu_j)(o_t - \mu_j)^t}{\sum_t L_j(t)}$$

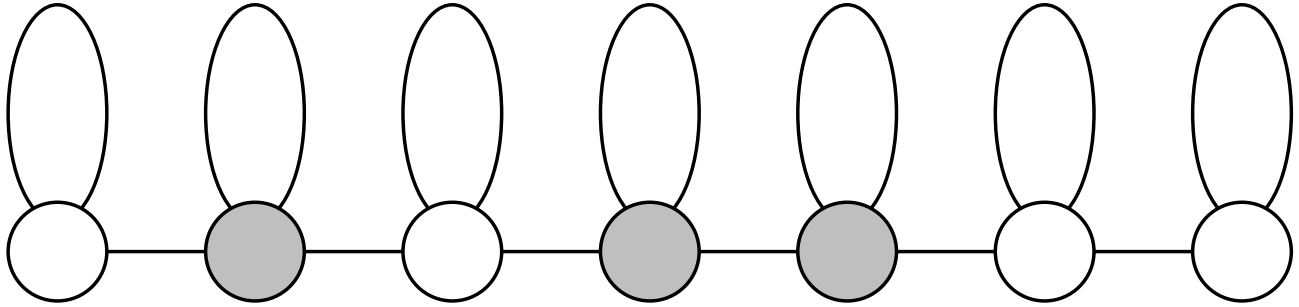
- 学習データ数 $= R$個

$$\hat{\mu}_j = \frac{\sum_r \left[\sum_t L_j^r(t) \cdot o_t^r\right]}{\sum_r \left[\sum_t L_j^r(t)\right]} = \frac{\sum_r \frac{1}{P^r} \left[\sum_t \alpha_j^r(t)\beta_j^r(t) \cdot o_t^r\right]}{\sum_r \frac{1}{P^r} \left[\sum_t \alpha_j^r(t)\beta_j^r(t)\right]}$$

$$\hat{\Sigma}_j = \frac{\sum_r \left[\sum_t L_j^r(t) \cdot (o_t^r - \mu_j)(o_t^r - \mu_j)^t\right]}{\sum_r \left[\sum_t L_j^r(t)\right]} = \cdots$$

EMアルゴリズムによるパラメータの推定4(結び)

- 複数の状態に，同一の平均値・分散行列を持たせる



$$\hat{\mu}_S = \frac{\sum_r \sum_{j \in S} \left[\sum_t L_j^r(t) \cdot o_t^r \right]}{\sum_r \sum_{j \in S} \left[\sum_t L_j^r(t) \right]}$$

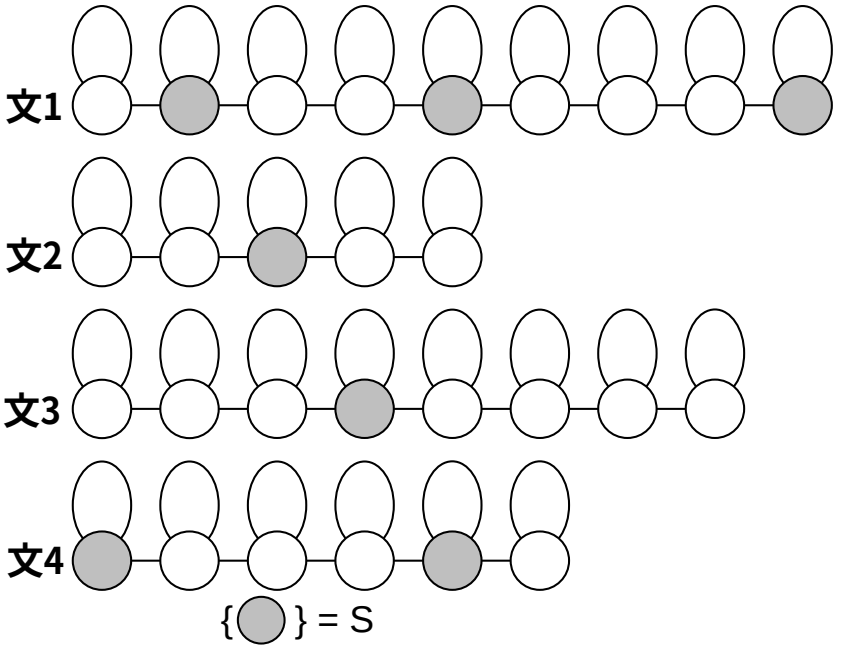
$$\hat{\Sigma}_S = \frac{\sum_r \sum_{j \in S} \left[\sum_t L_j^r(t) \cdot (o_t^r - \mu_j)(o_t^r - \mu_j)^t \right]}{\sum_r \sum_{j \in S} \left[\sum_t L_j^r(t) \right]}$$

EMアルゴリズムによるパラメータの推定5(連結学習)

- 書き起こしのみが存在する学習データに対する対処

→ HMM を連結して，発話 (文) 単位の HMM を構成

→ 文 HMM 間で,「同一種類」の状態を「結びの関係を持った」状態として捉える。



$$\hat{\mu}_S = \frac{\sum_r \sum_{j^r \in S} \left[\sum_t L^r_{j^r}(t) \cdot o^r_t \right]}{\sum_r \sum_{j^r \in S} \left[\sum_t L^r_{j^r}(t) \right]}$$

$$\hat{\Sigma}_S = \frac{\sum_r \sum_{j^r \in S} \left[\sum_t L^r_{j^r}(t) \cdot (o^r_t - \mu_{j^r})(o^r_t - \mu_{j^r})^t \right]}{\sum_r \sum_{j^r \in S} \left[\sum_t L^r_{j^r}(t) \right]}$$

孤立単語音声の認識

$$\arg\max_{M} P(O|M) = \arg\max_{M} \left\{ \sum_{X} P(O, X|M) \right\}$$

$$\downarrow \quad (X\text{=経路})$$

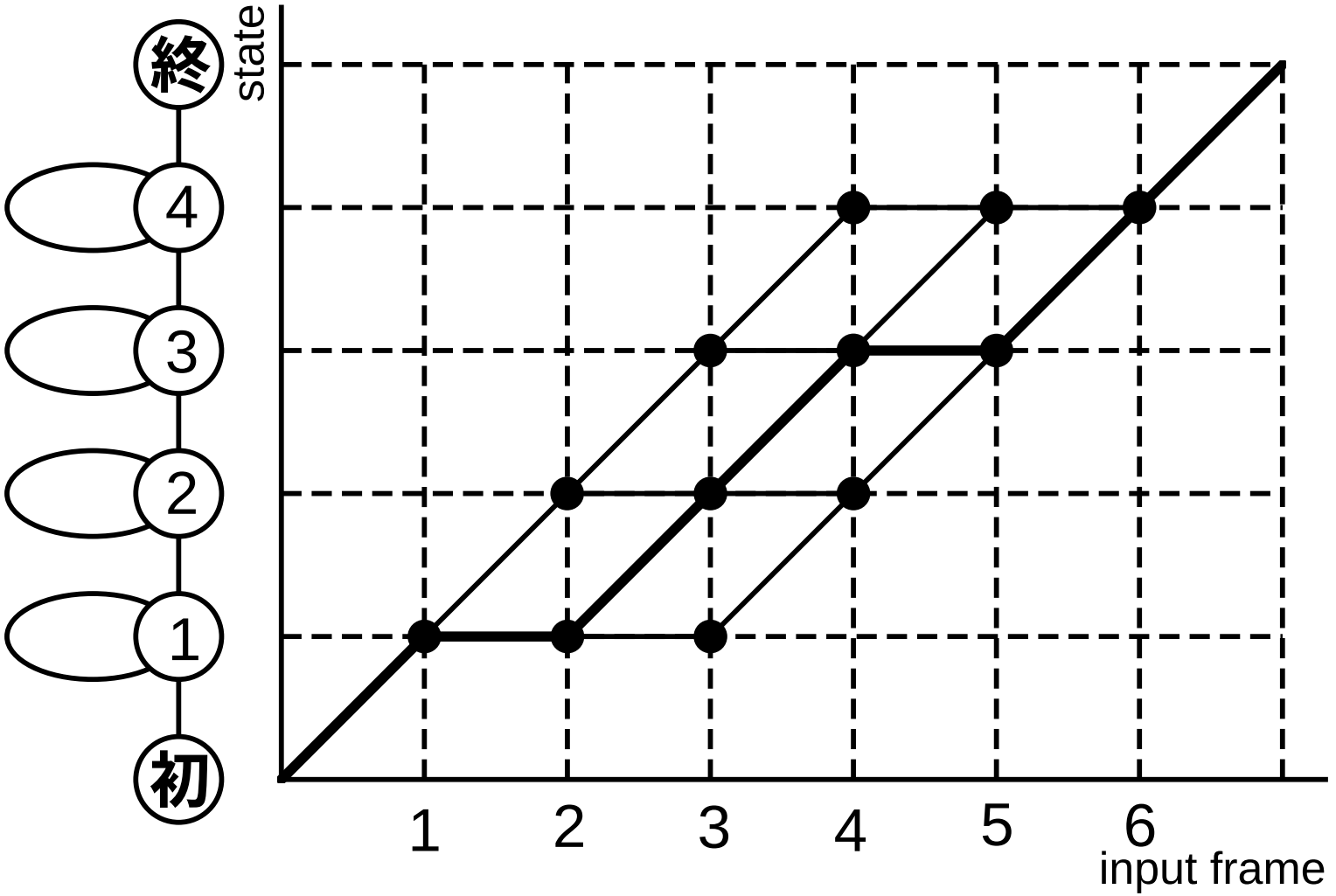
$$\arg\max_{M} \hat{P}(O|M) = \arg\max_{M} \left\{ \max_{X} P(O, X|M) \right\}$$

$$\alpha_j(t) = \sum_{i} \alpha_i(t-1) a_{ij} b_j(o_t), \quad (\alpha_N(T) \equiv P(O|M))$$

$$\downarrow$$

$$\phi_j(t) = \max_{i} \phi_i(t-1) a_{ij} b_j(o_t), \quad (\phi_N(T) \equiv \hat{P}(O|M))$$

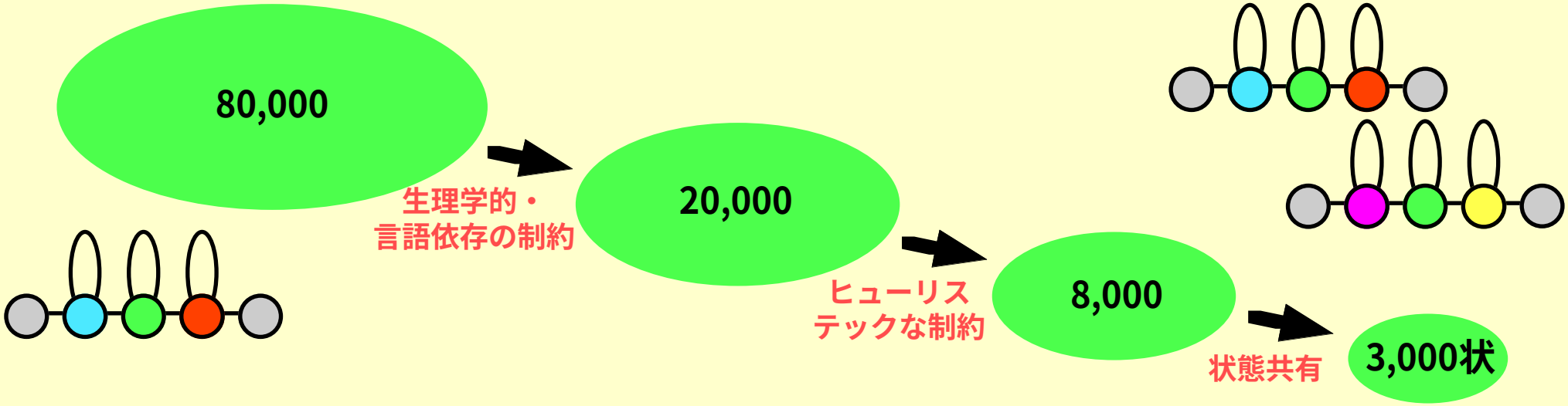
孤立単語音声の認識 (contd.)



音響モデル --- HMM ---(#?)

❒ *対象音韻の音響パラメータが持つ前後音韻への依存性*

- ✔ 音響パラメータの分布は前後音韻に影響を受ける → 前後音韻別のモデル
- ✔ kAr，mEi，sAz，，，，→ triphone
- ✔ #triphone = [#monophone]^3 = approx. 80,000 ?



❒ *モデル精度向上に向けて*

- ✔ 混合分布化
- ✔ 状態数，モデルのトポロジーの最適な配置
- ✔ 継続時間長制御の導入
- ✔ 特徴パラメータの速度成分，加速度成分の導入
- ✔ セグメント化
- ✔ その他色々色々色々色々色々色々色々色々，，，，，，，，，

言語モデル --- N-gram ---(#1)

❒ *現在までの単語履歴から次の単語を予測するモデル*

→ P(wi|w1,...,wi-1)

→ i の増加と共に求めるべきパラメータ数が激増，N-1重マルコフ過程を仮定

→ P(wi|w1,...,wi-1) = P(wi|wi-N+1,...,wi-1)，過去のN-1個の単語履歴にのみ依存

→ P(wi|wi-N+1,...,wi-1) = C(wi-N+1,...,wi) / C(wi-N+1,...,wi-1)　：N-gram の最尤推定値

→ P(wi|wi-1) = C(wi-1, wi) / C(wi-1)　：バイグラム

→ P(wi|wi-2, wi-1) = C(wi-2, wi-1, wi) / C(wi-2, wi-1)　：トライグラム

❒ *N-gram 推定時の問題 --- データが足りない ---*

✔ 全語彙を対象とするのではなく，頻出M単語を対象語彙として限定する

✔ 単語のトライグラム(語彙数20K)を精度良く推定するためには数年間の新聞データでは不足

→ C(wi-2, wi-1, wi) = 0，あるいは，C(wi-2, wi-1) = 0 の場合どうする？？？

❒ *Back off スムージングによる N-gram 確率の推定(未知N-gramへの対処)*

✔ C(wi-2, wi-1) = 0 の時

→ P(wi|wi-2, wi-1) = P(wi|wi-1)　：バイグラム確率を流用

✔ C(wi-2, wi-1, wi) = 0 (但し，C(wi-2, wi-1) > 0) の時

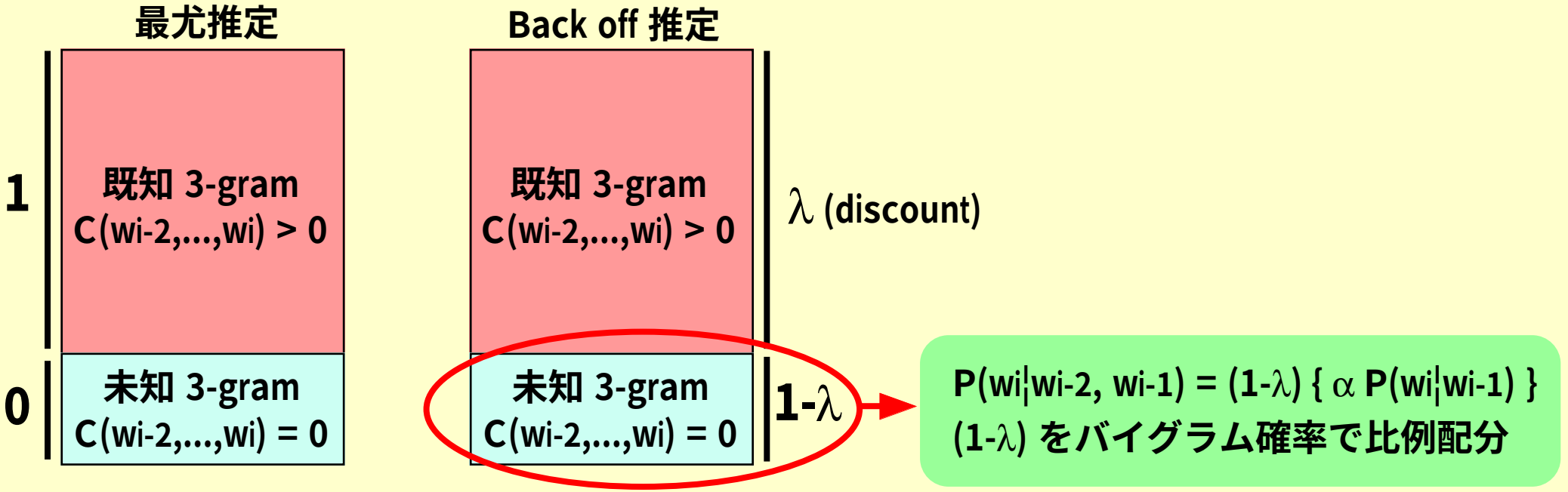
→ 既存の N-gram 確率の一部を未知の N-gram の確率として分配する

言語モデル --- N-gram ---(#2)

❒ *Back off スムージングによる 3-gram 確率の推定(未知3-gramへの対処)*

✔ C(wi-2, wi-1, wi) = 0 (但し，C(wi-2, wi-1) > 0) の時

→ 既存の 3-gram 確率の一部を未知の 3-gram の確率として分配する



❒ *種々の Back off スムージング*

✔ λ を wi に依存させ，様々な知見に基づいた Back off スムージングが提案されている

❒ *カットオフの導入*

✔ C(wi-2, wi-1, wi) = 0 のみならず，C(wi-2, wi-1, wi) <= C_0 時にも Back off により推定する

❒ *単語クラス(品詞，上位概念)を利用したN-gram*

言語モデル --- 辞書 / 語彙サイズ ---(#3)

❒ *認識対象語彙 = テキストデータベースにおける頻出単語*

❒ *語彙サイズと被覆率(coverage)*

✔ 45 か月分の新聞データを元に算出

✔ 「句読点」も語彙の一部。該当する音素表記は無音記号(sp, silB, silE)

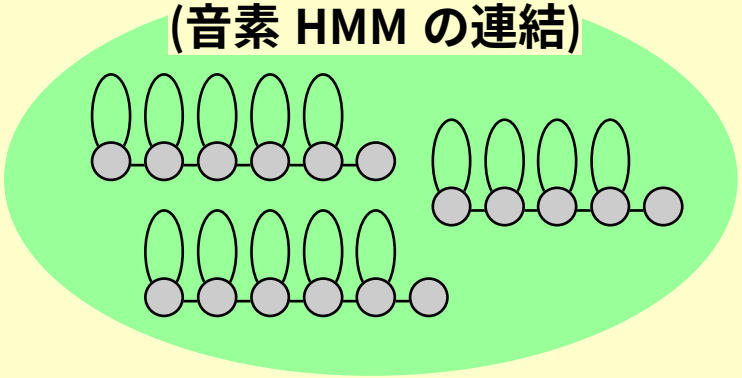
語彙サイズ	被覆率
5,000	88.2 %
6,135	90.0 %
20,000	96.5 %
22,959	97.0 %

デコーダ(認識エンジン，#1)

❒ *基本的なアルゴリズム(one-pass アルゴリズム)*

✔ バイグラム言語モデルを使ったデコーディング

単語 HMM セット
(音素 HMM の連結)

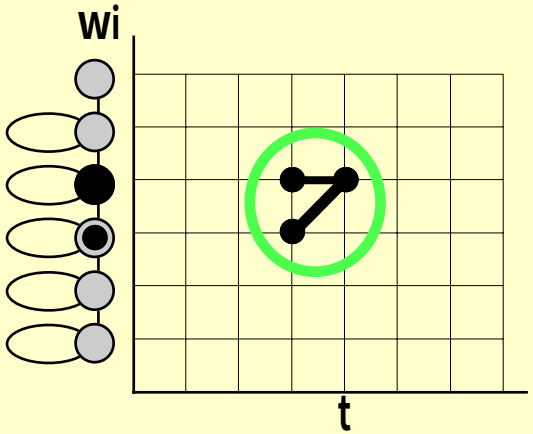


時刻 t の入力フレームは何れの単語の何れの状態とも照合を行なわなければならない

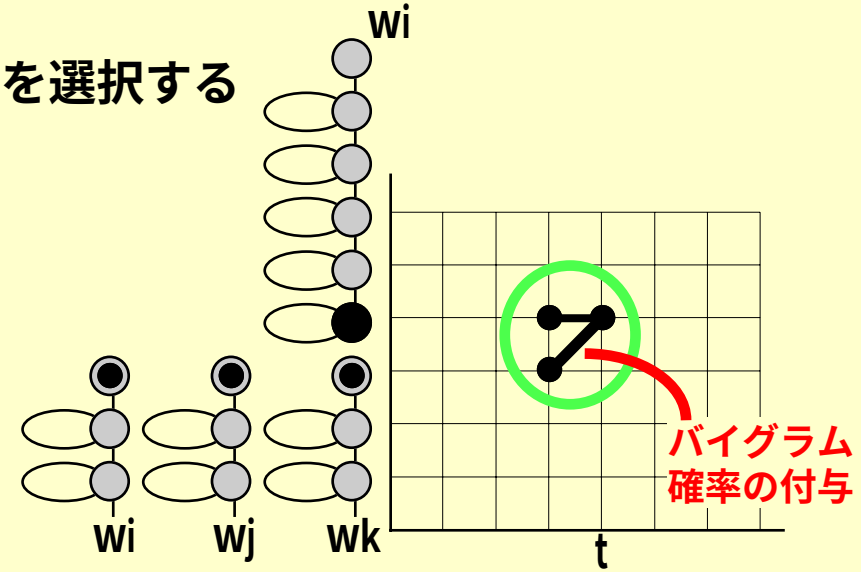
入力フレーム t は全状態と照合される

[時刻 t が wi の状態 s と対応する場合の最適パス] を各単語，各状態ごとに保持

○ : 累積スコアが高いパスを選択する

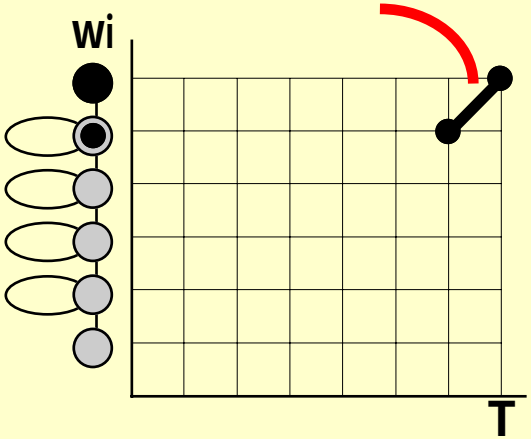


単語の非第一状態との照合



単語の第一状態との照合

最適パスを文尾から
バックトレースする



時刻Tのフレームの照合

デコーダ(認識エンジン，#2)

❒ *語彙数の増加 / 状態数の増加 → 記憶容量 / 計算量の増加*

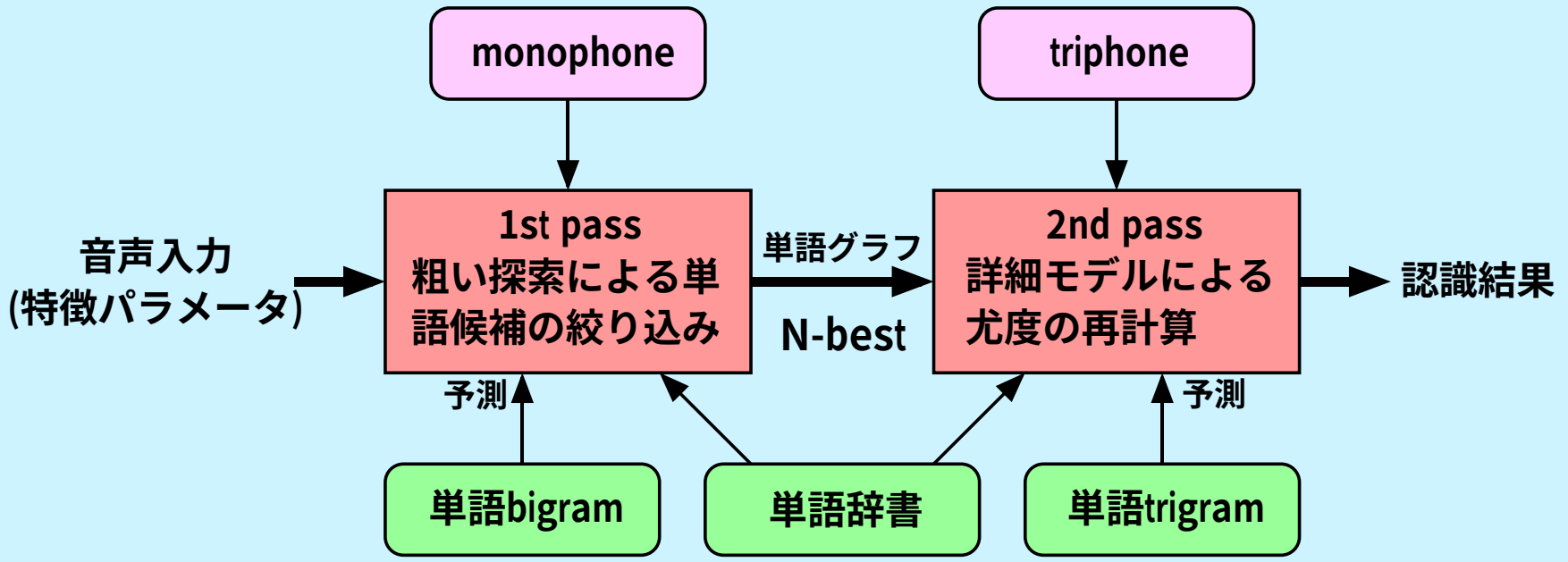
✔ ビームサーチ

→ 各時刻において累積尤度の上位 M 個を残し，それ以外は切り捨てる

❒ *音響モデル / 言語モデル / デコーダの統合*

✔ マルチパス探索

→ 1st pass で粗いモデルによる予備探索，2nd pass で詳細なモデルによる再評価



デコーダ(認識エンジン，#3)

❒ *種々のテクニック*

✔ 枝刈りに関して
探索時点 t までの累積尤度と，それ以降の時刻 T までのフレームに対する評価値を使って刈る枝を決定する。時刻 T までのフレームに対する評価値は heuristics により決定する。

✔ 木構造化辞書
prefix が同じ単語を木構造化し，照合処理の効率を図る。

✔ 言語モデル確率のファクタリング
単語尾と単語頭の接続部において言語モデル確率は付与されるが，これが単語内部での照合においても付与されるよう，言語モデル確率を分割して各ノードに割り振る。

✔ 言語尤度重み
音響尤度に比べて言語尤度はそのレンジが小さいため，言語尤度に重みをかけて両者を足し合わせることにより，認識精度の向上を図る。

✔ インサーションペナルティー
短い単語の挿入誤りを防ぐため，照合中に新たな単語へと遷移する度にペナルティーを付与する。言語尤度重みの大きさに依存して設定される。

音声認識性能(IPA'98)

❒ *連続音節認識結果*

- ✔ 音響モデル：状態共有の triphone，総状態数 = 3,000
- ✔ 言語モデルは使用せず。純粋に音響モデルの性能評価
- ✔ 評価文数 = 約 930

特徴パラメータ	Acc.	Cor.	Sub.	Ins.	Del.
MFCC+ΔMFCC+Δpow	66.5	80.6	16.7	14.1	2.8
上記+Δ^2MFCC+Δ^2pow	70.8	83.6	14.4	12.8	2.0

正解率(Cor.) + 置換率(Sub.) + 脱落率(Del.) = 100
正解精度(Acc.) = 正解率(Cor.) - 挿入率(Ins.)

❒ *大語彙連続音声認識(連続単語認識)結果*

- ✔ 音響モデル：状態共有の triphone，総状態数 = 3,000
- ✔ 言語モデルは45ケ月分の新聞記事から作成 / 語彙数 = 20 K
- ✔ 評価文数 = 約 100

デコーダ	音響モデル	言語モデル	計算時間	Acc.	Cor.
高速版	monophone	圧縮triphone	2.2 x RT	84.6	85.7
高精度版	triphone	triphone	8.5 x RT	92.6	93.7

種々の条件下における認識結果例

❒ *連続音韻認識結果(triphone の任意連結)*

SIL Q b e: k o k u d e o N o b e t o n a m u k i t a: N h e: e i n o k o k u m i N n o m e w a c h i m e t a k u SIL SIL Q a y a g a d o o: o j o: o: w a ts u n e r u m a d e i n i w a SIL SIL ts u k a n a r i n o s a i g e ts o h i ch i o o t o sh i t a

❒ *連続音節認識結果(上記＋音節構造の知識導入)*

SILげいこくでおんおべとなむきたんへいのこくみんのめわちめたくSIL SILっあやがどおじょおわつねるまでいにわSIL SILつかなりのさいげつおひちおおとしたSIL

❒ *連続単語認識結果(上記＋単語の知識導入，語彙数=20K)*

1st pass 米穀 ネオン ベトナム 機関 平 残っ 区民 度目 月 目立っ 句 。 ？ カヤ 花道 王女 大和 詰める まで なり なさい えっ 消費 治療 落とし 他

2nd pass 米穀 ネオン ベトナム 帰還 平 残っ 区民 度目 月 目立っ く 、 、 カヤ 門 王女 大和 詰める まで 庭 り なさい れ 曹 陽 治療 落とし た

❒ *大語彙連続音声認識結果(上記＋単語間の連鎖知識導入)*

1st pass 米国のベトナム帰還兵の国民の目が冷たく、彼らは同情を集めるまでには、かなりの歳月を必要 落とし た。

2nd pass 米国のベトナム帰還兵の国民の目は冷たく、彼らが同情を集めるまでには、かなりの歳月を必要 と し た。

❒ *正解文*

米国でもベトナム帰還兵への国民の目は冷たく、彼らが同情を集めるまでには かなりの歳月を必要とし た。

まとめ

❒ *音声認識の情報論的な定式化*

❒ *音声言語データベース / テキストデータベース*
- ✔ 現在は「まず，データありき」のご時世，，，，

❒ *音響モデル P(o|w)　--- HMM ---*
- ✔ 音声の生成モデルとしての HMM，特徴ベクトル = observed，状態系列 = hidden
- ✔ Viterbi 演算による尤度 P(o|w) 計算
- ✔ 環境依存型の triphone

❒ *言語モデル P(w)　--- N-gram ---*
- ✔ マルコフ過程を仮定した単語予測モデル
- ✔ Back off スムージングによるトライグラムの推定

❒ *デコーダ(認識エンジン)*
- ✔ one-passアルゴリズム
- ✔ 音響モデル / 言語モデル / デコーダの統合：マルチパスによる構成

❒ *現状の認識精度(技術レベル)とその例*
- ✔ IPA'98 の性能評価
- ✔ 種々の条件下における認識結果の比較

今後の課題

- 音響モデルの迅速な環境適応(ノイズ，マイク特性)
- 音響モデルの迅速な話者適応
- 言語モデルの種々ドメインへの適応化
- 未知語処理 / 不要語処理
- 非母語話者に対する対応
- 局所的モデリングから大局的モデリング(言語モデル)
- データベース収録(The larger, the better)
- などなど，，，