東京大学 工学部 計数工学科/物理工学科

応用音響学：音声認識

環境依存モデル、環境適応

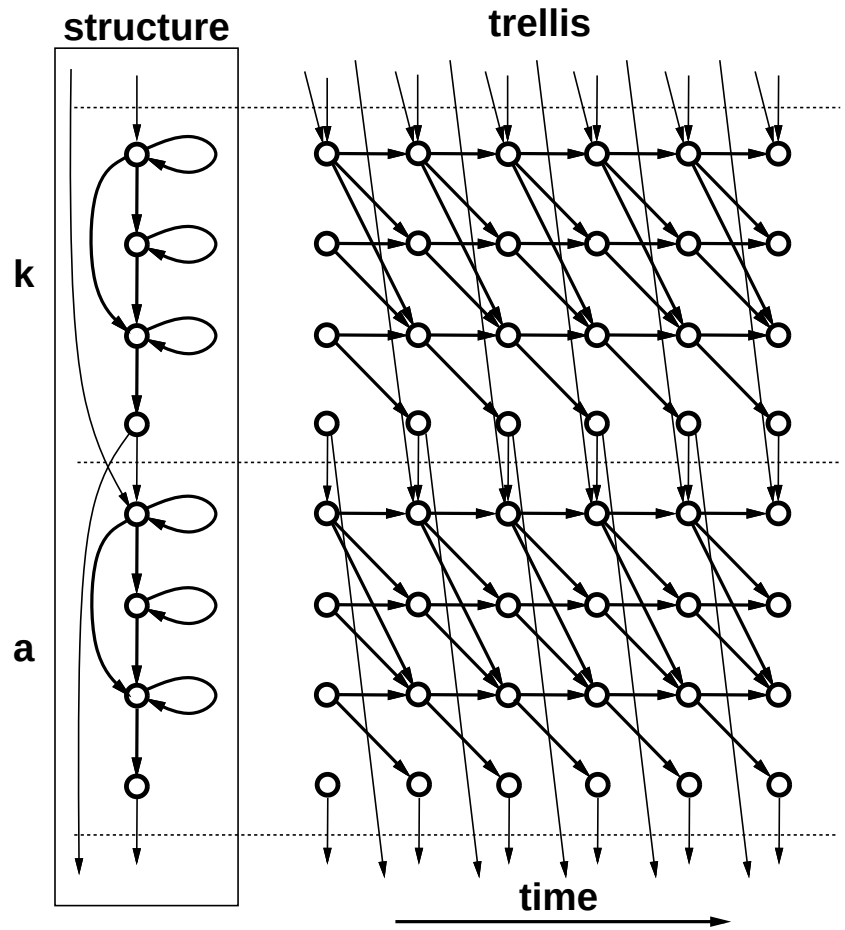
嵯峨山 茂樹 ＜sagayama@hil.t.u-tokyo.ac.jp＞
東京大学 工学部 計数工学科 http://hil.t.u-tokyo.ac.jp/

- 環境依存音素モデル
- 話者適応
- 雑音環境適応

サブワード単位

- 言語音声を精度良くかつ効率良く表わせる表現単位: サブワード単位
- 最も一般的なのは音素: 24種類ほど
  母音: /a/, /i/, /u/, /e/, /o/ の5個、
  子音: /k/, /s/, /t/, /n/, /h/, /m/, /r/, /w/, /g/, /z/, /d/, /b/, /p/, ...等々
- 音素コンテキスト
  「は」と「ひ」と「ふ」: 音素 /h/ の発音が後続音素/a/, /i/, /u/ に依存
  調音結合
- 異音(allophone)

HMM の Trellis 計算



Trellis Computation

図1. HMM トレリス計算 (縦方向は状態、横方向は時間)

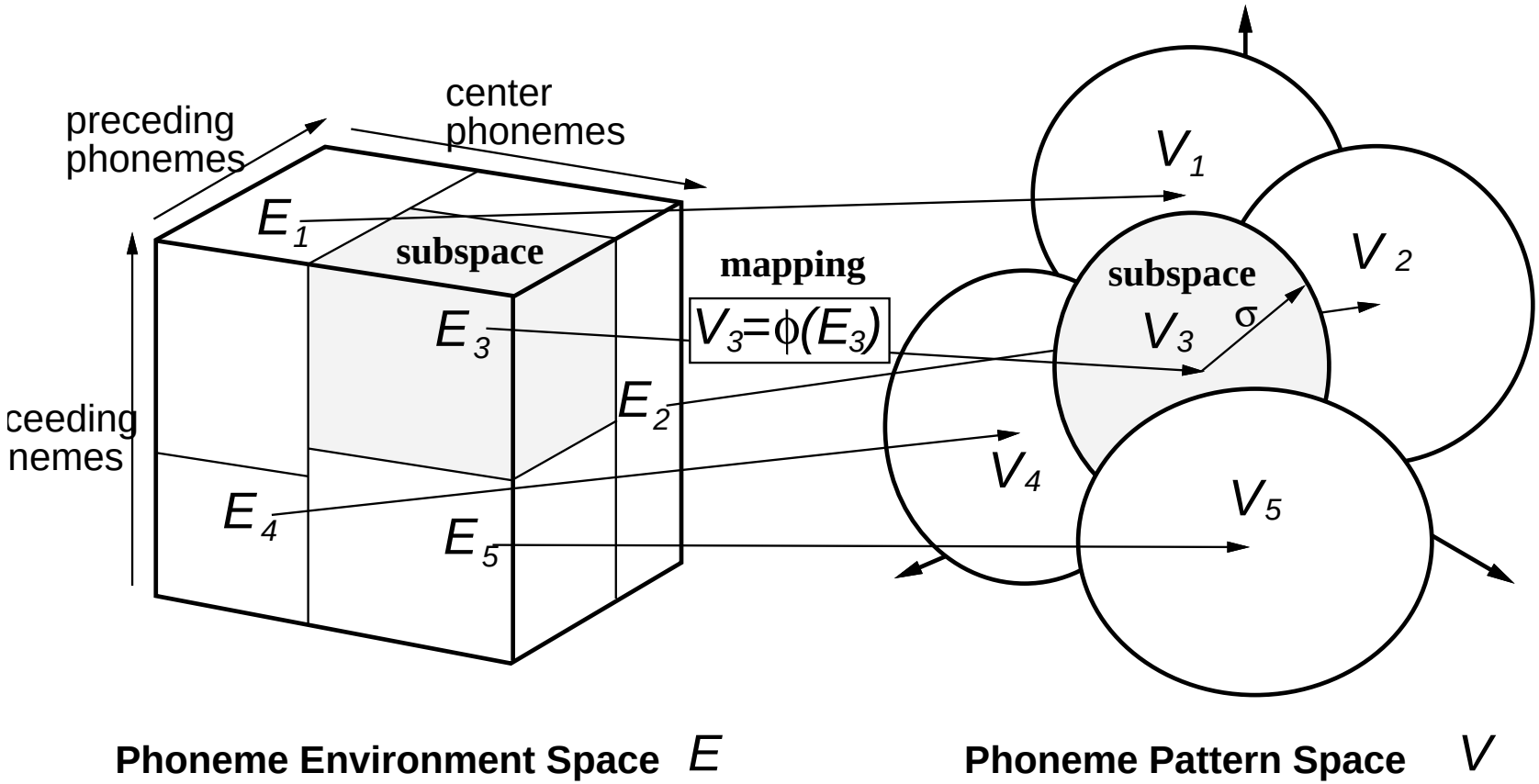
嵯峨山茂樹 (東京大学大学院情報理工学系研究科システム情報学専攻)

音素の環境依存性 – 異音(allophone)の例

- 先行音素，後続音素など phoneme context の影響
- 子音の例: 音素 /h/ は後続母音により別の allophone
  - は /ha/ [ha, xa], ひ /hi/ [hi, çi], ふ /hu/ [hɯ, ɸɯ], へ /he/ [hɛ, xɛ], ほ /ho/ [ho, xo]
- /i/ に後続する子音は 「硬口蓋化」する．
  - き /ki/, ぴ /pi/ は 摩擦性
- 「撥音」の例: 撥音 /ɴ/ は後続音素により別の調音
  - 後続音素が /a, i, u, e, o, k, h, j, w, g/ の場合 [ŋ]
  - 後続音素が /t, ʧ, ts, d, n, r, ʒ, dʒ/ の場合 [n]
  - 後続音素が /m, b, p/ の場合 [m]
  - 後続音素が /s, ʃ/ の場合 [$\tilde{ɨ}$]
- 「促音」の例: 撥音 /Q/ は後続音素が1モーラ長くなる．
  - 後続音素により別の調音
  - 後続音素が摩擦音 /s, ʃ, z, ʒ/ の場合，[s, ŋ, z, ʒ]
  - 後続音素が無声破裂音 /k, t, p, ʧ, ts/ の場合，無音
  - 後続音素が有声破裂音 /b, d, g/ の場合(外来語)，buzz bar

嵯峨山茂樹 (東京大学大学院情報理工学系研究科システム情報学専攻)

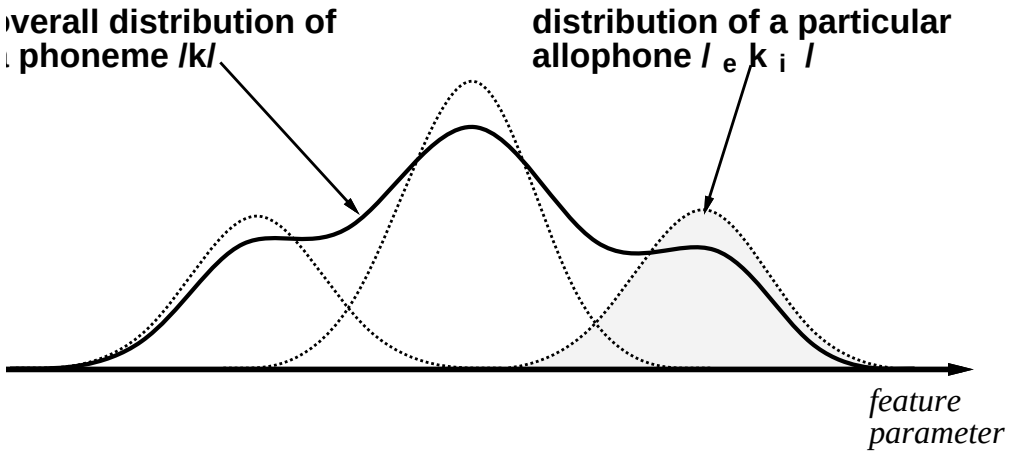
音素環境クラスタリング (PEC) (1987)



The Basic Idea of Phoneme Environment Clustering

図2. PECの原理 — 音響空間の像の分散が小さくなるように音素環境空間を逐次2分割する

音素環境クラスタリング (PEC) (1987)



Assersion: environment dependent phones follow Gaussian distributions

phonemes + mixture density vs. allophones + single Gaussian density

図3. PECの狙い

■ 木構造クラスタリング (トップダウン)

Cf. Generalized Triphones: ボトムアップ

音素(異音)モデル：隠れマルコフ網 (HMnet)

- 前後の音素の影響を詳細に反映した異音(allophone)モデル → 高精度
- 少量の学習データから良い音響モデルが作れる統計的な頑健さ → 頑健
- 各異音クラスタを表現する状態遷移経路が状態を共有する構造 → 高効率
- 各状態を単一ガウス分布で表現 → 計算量少、話者適応に有利
- たとえば ..... あらゆる異音は約1600種のクラスタに分類、わずか600状態(=600ガウス分布)間の状態遷移経路で表現

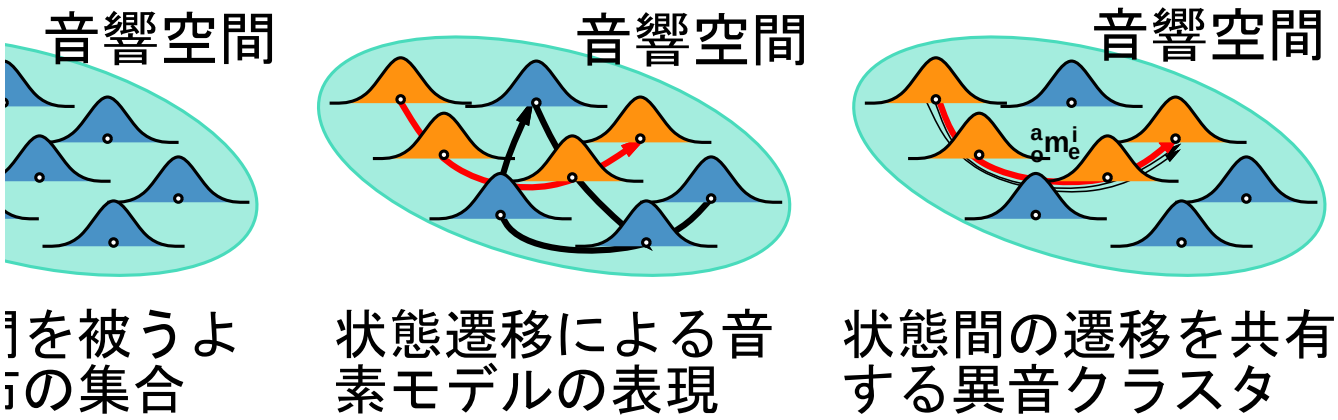


図4. 隠れマルコフ網の概念 — 音響空間を覆う単位分布群、分布を結んで音素パターンを表現する経路群、環境依存音素パターンのクラスタ群 ... の3つを同時に持つモデル

逐次状態分割法(SSS: Successive State Splitting)

- HMnet を生成するアルゴリズム
- 状態セット、経路セット、異音クラスタセットを同時に準最適決定
- 1状態から始めて、状態を異音方向と時間方向に逐次分割する
- tree構造の音素環境クラスタ

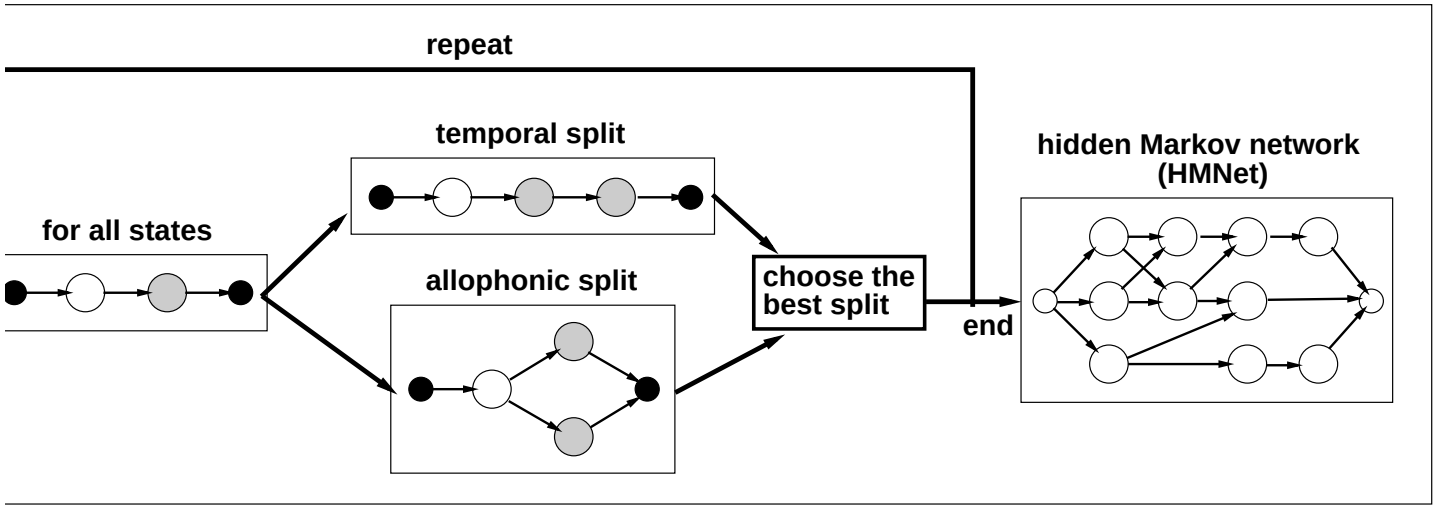


図5. 逐次状態分割法(SSS)の概略 — 大量の音素パターンのデータを用いて自動的にHMnetを生成する

逐次状態分割法(SSS)の例

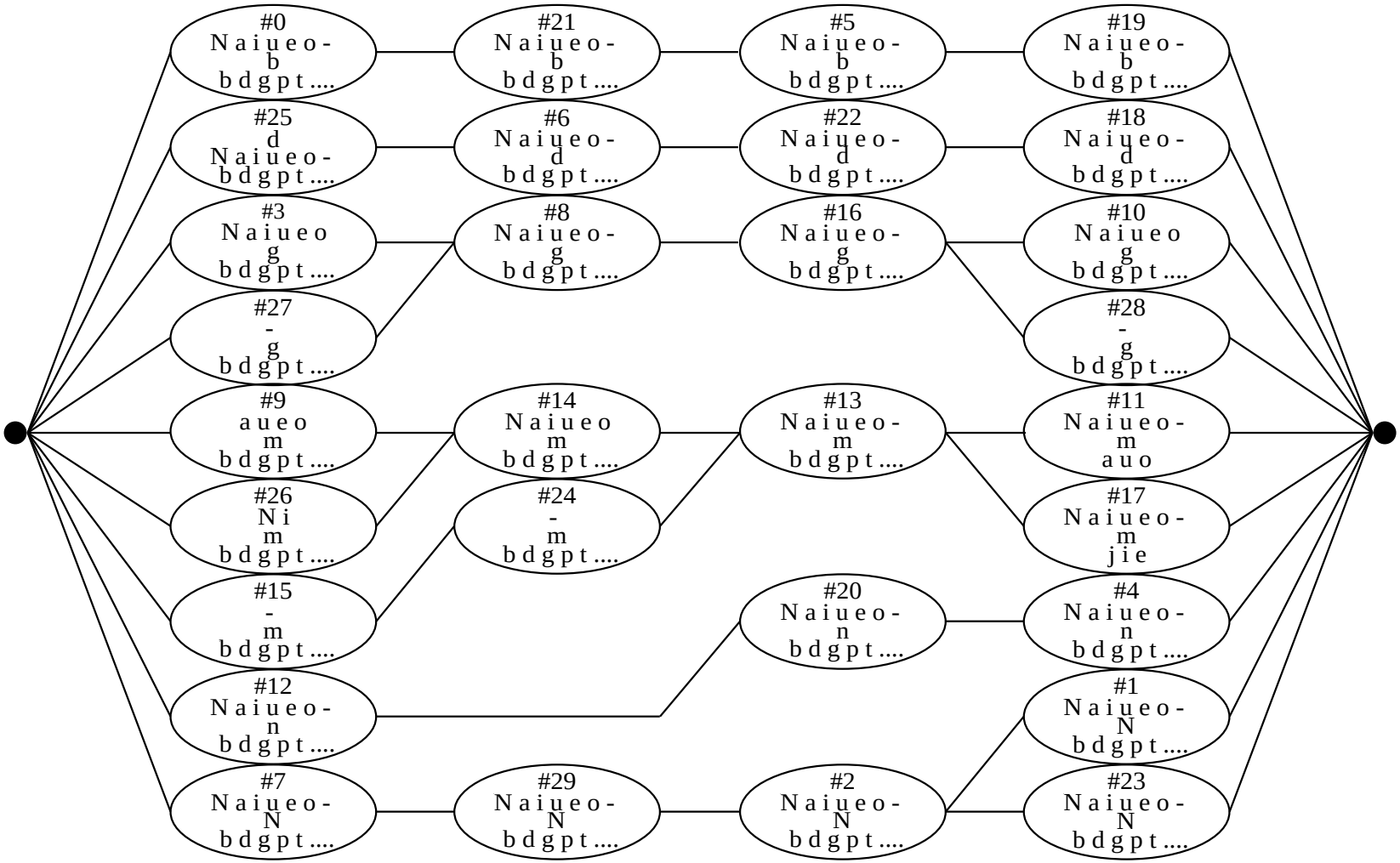
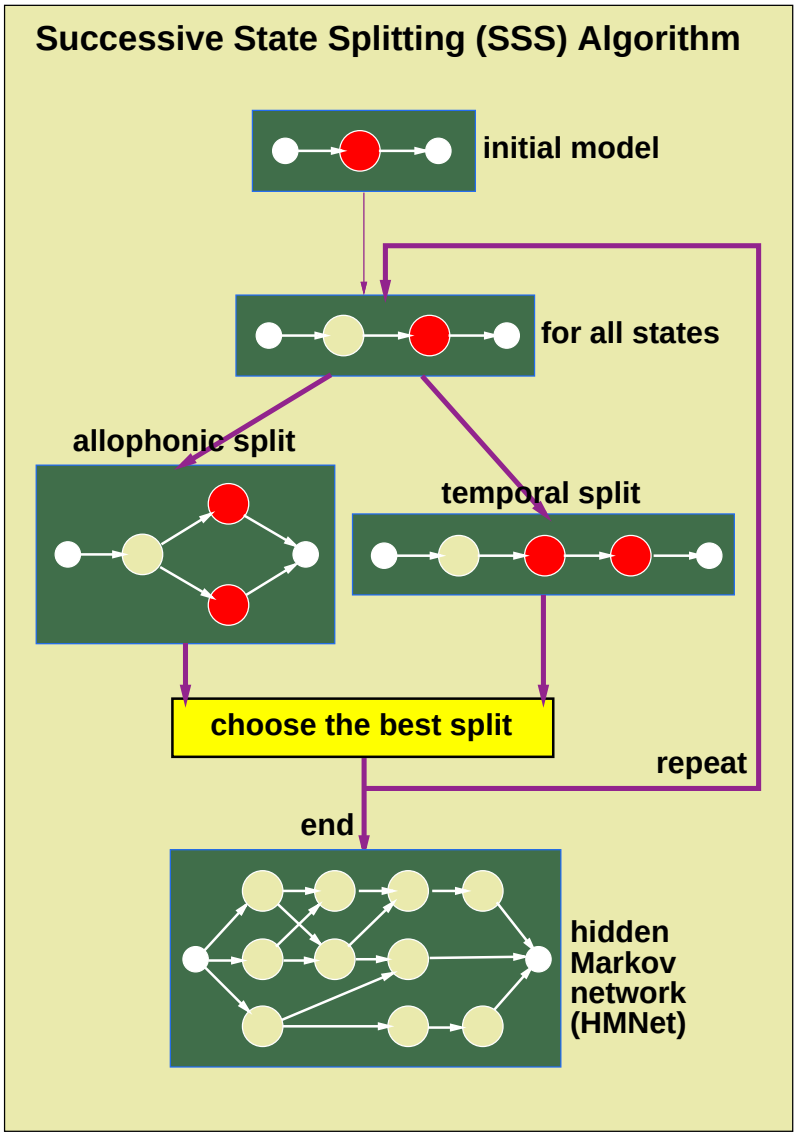


図6. 逐次状態分割法(SSS)により生成されたの概略 — 大量の音素パターンのデータを用いて自動的にHMnetを生成する

異音モデルの自動生成

逐次状態分割法(SSS)による異音モデルの自動生成



異音モデルの自動生成

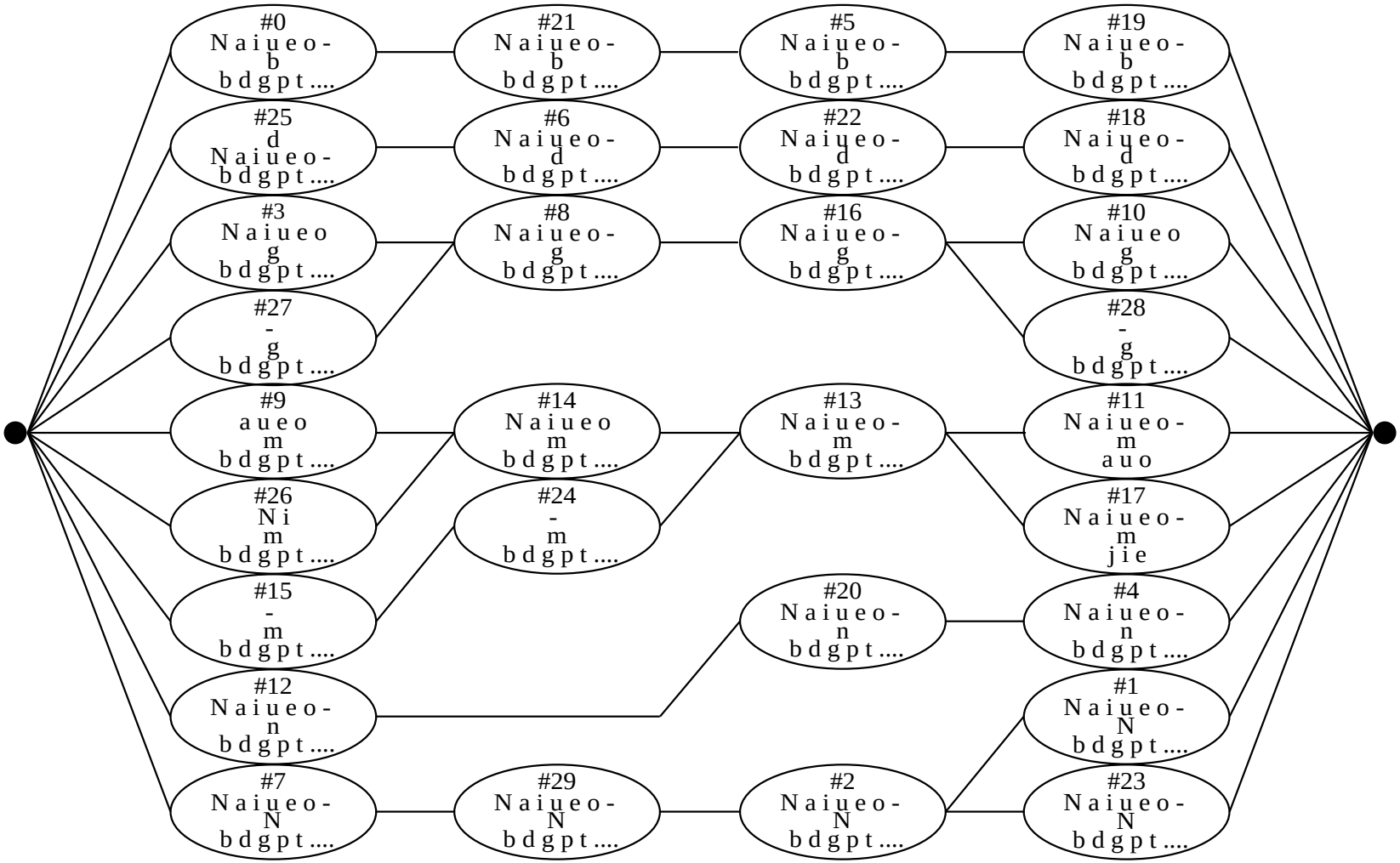


図7. 逐次状態分割法(SSS)による異音モデルの自動生成結果例

話者適応 : 移動ベクトル場平滑化(VFS)方式 (1991)

- 話者適応
  - 話者ごとの特性の違い → 不特定話者モデルを向上させる
  - 教師あり/なし
- 移動ベクトル場の原理
  - 話者差を特徴空間中のベクトル場としてモデル化
  - 特性が滑らかに変化するフィルタで話者差を説明
- 移動ベクトル場平滑化(VFS)方式
  - 少量の学習データで話者適応が可能
  - 広い適用可能性 – 離散・連続・HMnet いずれにも
  - 話者差のベクトル場をDPマッチング(内容固定)・連結学習(内容既知)により推定、空間フィルタによりベクトル場を平滑化

不特定話者と話者適応化

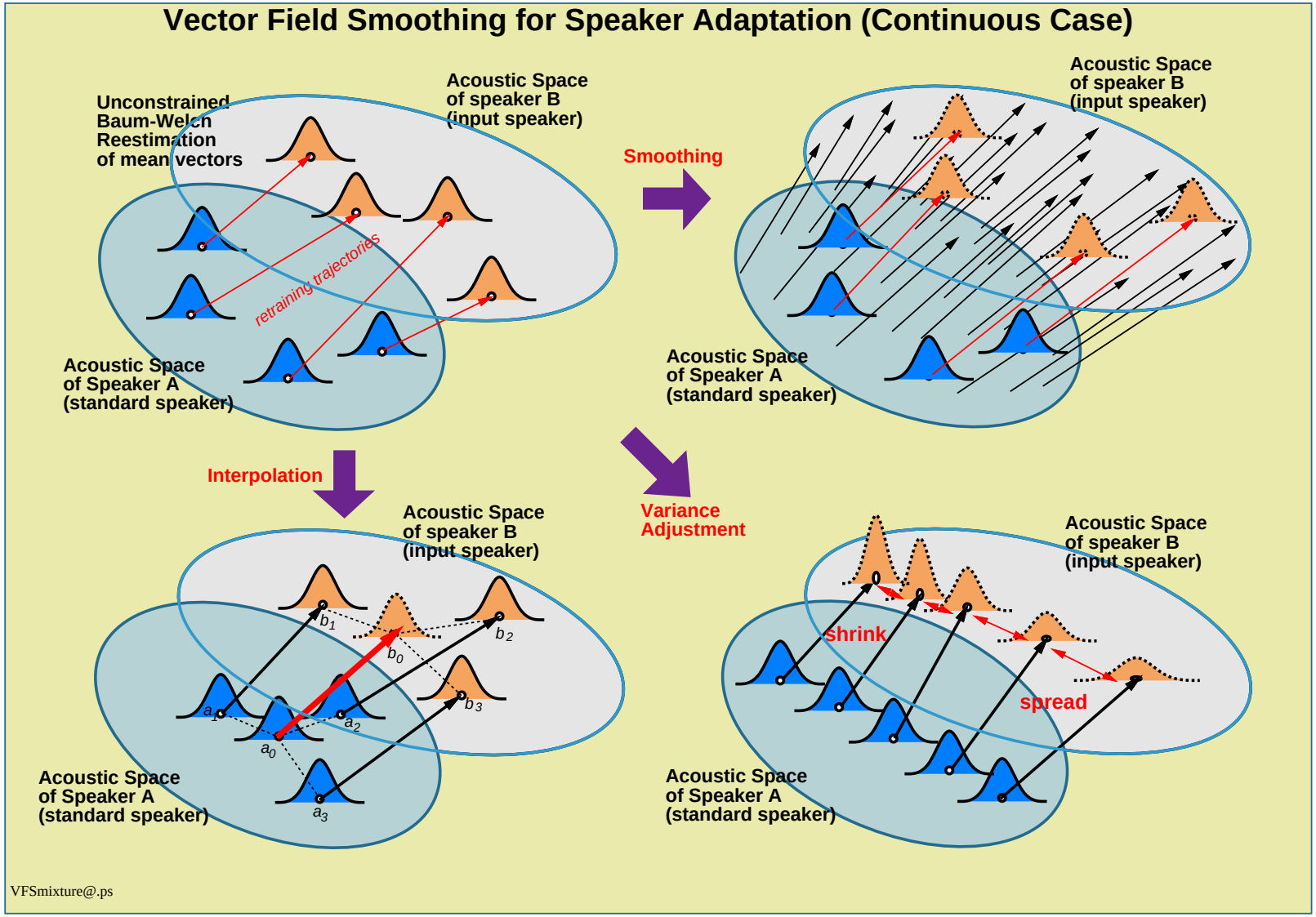
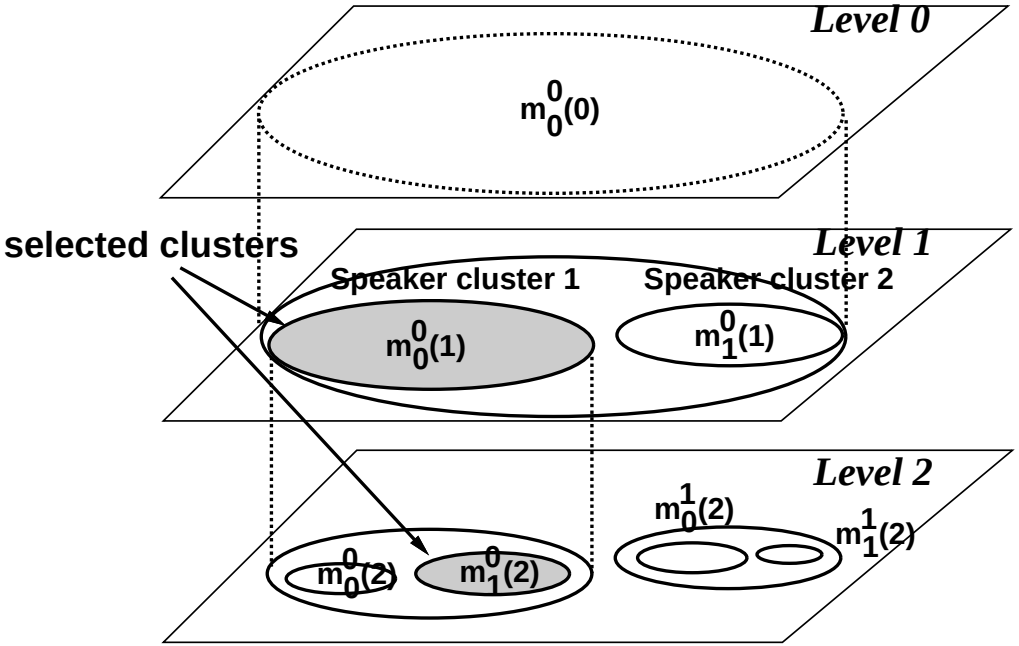
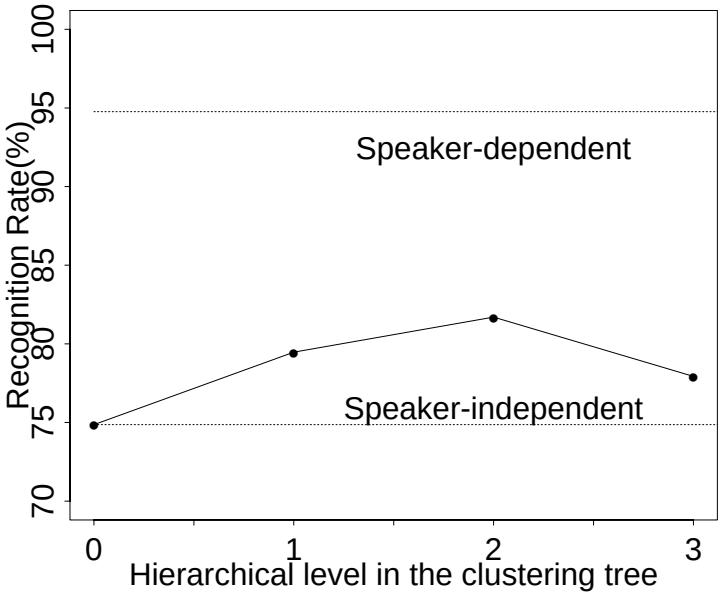


図8. ベクトル場平滑化法(VFS)によるHMMの話者適応

話者木構造



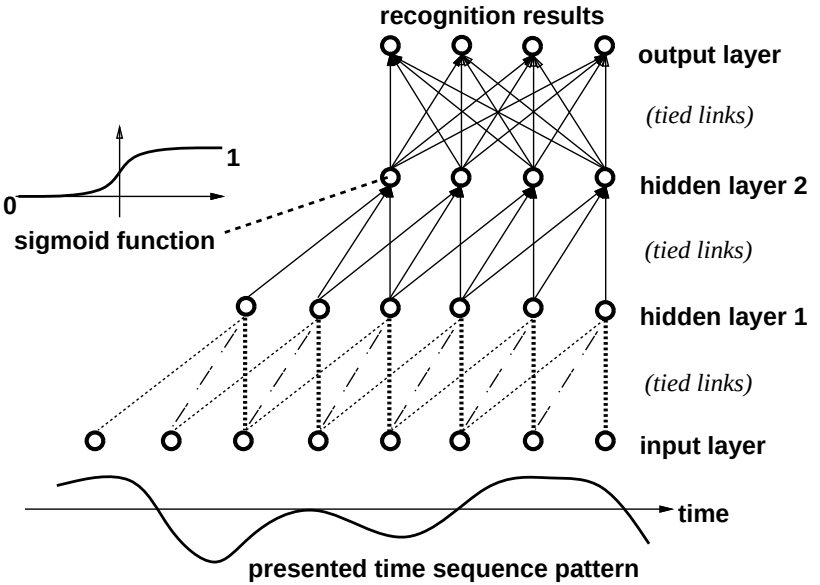


木構造話者クラスタリングの階層選択 (原理，性能)

ニューラルネット(TDNN)による音声認識

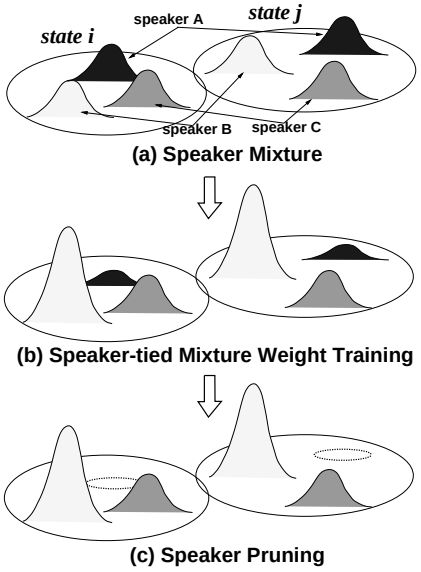
- 「過学習」問題 — 頑健さ
- 過学習緩和方法
  - ニューロン出力平滑化
  - 連続値教師信号による学習
  - 分布を平滑化した学習
  - ニューラルファジー学習
  - 時空間ブロック統合TDNN
  - 対判別型TDNN

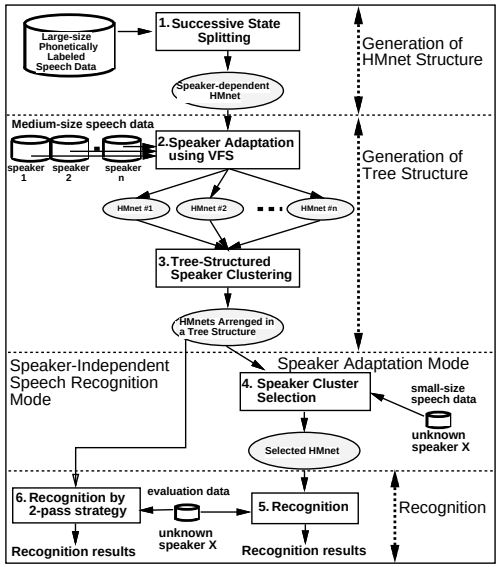
Partially Connected Neural Network
(Basic Idea of Time-Delay Neural network (TDNN))

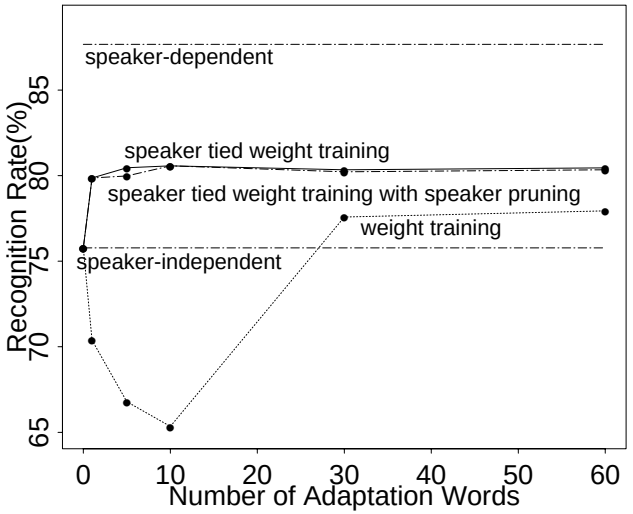


話者混合モデル (1992)

- 単一ガウス分布型のHMnet を n 話者(クラスタ)について作成.
- n 混合ガウス分布型のHMnet にする．話者確率は等確率.
- 少量のサンプルから混合重み(n個)をBaum-Welchアルゴリズムで学習.







話者混合による高速話者適応 (原理，手順，性能)

木構造の波及

- 音声学的規則による音素環境の木構造クラスタリング (CMU, U Cambridge, IBM, etc.)
- ML-SSS (ATR, Boston U)

雑音環境音声認識

- 雑音を取り込まない
  - アレイマイクロフォン
  - 複数マイクロフォンで零感度方向を作る
- 雑音を軽減する
  - SS (spectrum Subtraction)
- 雑音を含むモデルを用いる
  - モデル分解
  - モデル合成: PMC (Parallel Model Combination)
  - モデル適応: JA (Jacobian Adaptation)

嵯峨山茂樹 (東京大学大学院情報理工学系研究科システム情報学専攻)

Jacobian による音響モデルの雑音適応 (1996)

■ 雑音適応：非線形な適応

$$\overbrace{\boldsymbol{C}_{S+N}}^{\text{複合ケプストラム}} = \overbrace{\boldsymbol{F}^{*}}^{\text{逆フーリエ変換}} \left[\log \left\{ \underbrace{\overbrace{\exp\left(\boldsymbol{F}\, C_S\right)}^{\text{音声スペクトル } \boldsymbol{S}} + \overbrace{\exp\left(\boldsymbol{F}\, C\right.}^{\text{重畳雑音スペク}}}_{\text{複合スペクトル } \boldsymbol{S}+\boldsymbol{N}} \right.\right.$$

■ 着眼点

- ■ 雑音 A から雑音 B への適応
- ■ 微分量による雑音適応 → 線形計算 (Taylor 展開の1次)．ヒント:

$$df = f_x dx + f_y dy$$

■ Jacobian 適応

$$\underbrace{\Delta \boldsymbol{C}_{S+N}}_{\boldsymbol{C}_{S+N}\text{の微小変化}} = \underbrace{\frac{\partial \boldsymbol{C}_{S+N}}{\partial \boldsymbol{C}_S}}_{\textit{Jacobi}\text{ 行列}} \underbrace{\Delta \boldsymbol{C}_S}_{\boldsymbol{C}_S\text{の微小変化}} + \underbrace{\frac{\partial \boldsymbol{C}_{S+N}}{\partial \boldsymbol{C}_N}}_{\textit{Jacobi}\text{ 行列}} \underbrace{\Delta \boldsymbol{C}_N}_{\boldsymbol{C}_N\text{の微小変化}}$$

Jacobian Matrix of Cepstrum

■ ケプストラムの Jacobi 行列:

$$\frac{\partial \boldsymbol{C}_{S+N}}{\partial \boldsymbol{C}_N} = \underbrace{\frac{\partial \boldsymbol{C}_{S+N}}{\partial \log(\boldsymbol{S}+\boldsymbol{N})}}_{=F^*} \underbrace{\frac{\partial \log(\boldsymbol{S}+\boldsymbol{N})}{\partial(\boldsymbol{S}+\boldsymbol{N})}}_{=\frac{1}{S+N}} \underbrace{\frac{\partial(\boldsymbol{S}+\boldsymbol{N})}{\partial \boldsymbol{N}}}_{=1} \underbrace{\frac{\partial \boldsymbol{N}}{\partial \log \boldsymbol{N}}}_{=N} \underbrace{\frac{\partial \log \boldsymbol{N}}{\partial \boldsymbol{C}_N}}_{=F}$$

$$= \boldsymbol{F}^* \frac{\boldsymbol{N}}{\boldsymbol{S}+\boldsymbol{N}} \boldsymbol{F} = \left(\sum_k F^{-1}_{ik} \frac{N_k}{S_k+N_k} F_{kj} \right)_{ij}$$

■ デルタケプストラムの Jacobi 行列:

$$\frac{\partial \dot{\boldsymbol{C}}_{S+N}}{\partial \boldsymbol{C}_N} = \frac{\partial}{\partial \boldsymbol{C}_N}\left(\frac{\partial \boldsymbol{C}_{S+N}}{\partial t}\right) = \frac{\partial}{\partial t}\left(\frac{\partial \boldsymbol{C}_{S+N}}{\partial \boldsymbol{C}_N}\right) = \boldsymbol{F}^* \frac{\partial}{\partial t}\left(\frac{\boldsymbol{N}}{\boldsymbol{S}+\boldsymbol{N}}\right)\boldsymbol{F}$$

$$= \boldsymbol{F}^* \left(\frac{\dot{\boldsymbol{N}}\boldsymbol{S} - \boldsymbol{N}\dot{\boldsymbol{S}}}{(\boldsymbol{S}+\boldsymbol{N})^2}\right)\boldsymbol{F} = \left(\sum_k F^{-1}_{ik} \frac{\dot{N}_k S_k - N_k \dot{S}_k}{(S_k+N_k)^2} F_{kj} \right)_{ij}$$

where the time-derivative of speech spectrum $\dot{S}$ is given by:

$$\dot{\boldsymbol{S}} = \frac{\partial}{\partial t}\{\exp(\log \boldsymbol{S})\} = \exp(\log \boldsymbol{S})\frac{\partial}{\partial t}(\log \boldsymbol{S}) = S\frac{\partial}{\partial t}(F\boldsymbol{C}_S) = \boldsymbol{S}\boldsymbol{F}\dot{\boldsymbol{C}}_S$$

Jacobian 適応法の計算量

- CPU time for adaptation (acoustic analysis not included) (measured on Sun SPARCstation20)

phase	JA		NOVO	ratio JA/NOVO
	cep	cep+△cep		
training	2,216 ms	8,033 ms	4,416 ms	1/2
recognition	149 ms	349 ms	5,066 ms	1/34

Noise Adaptation (Exhibition Hall ← Crowd)

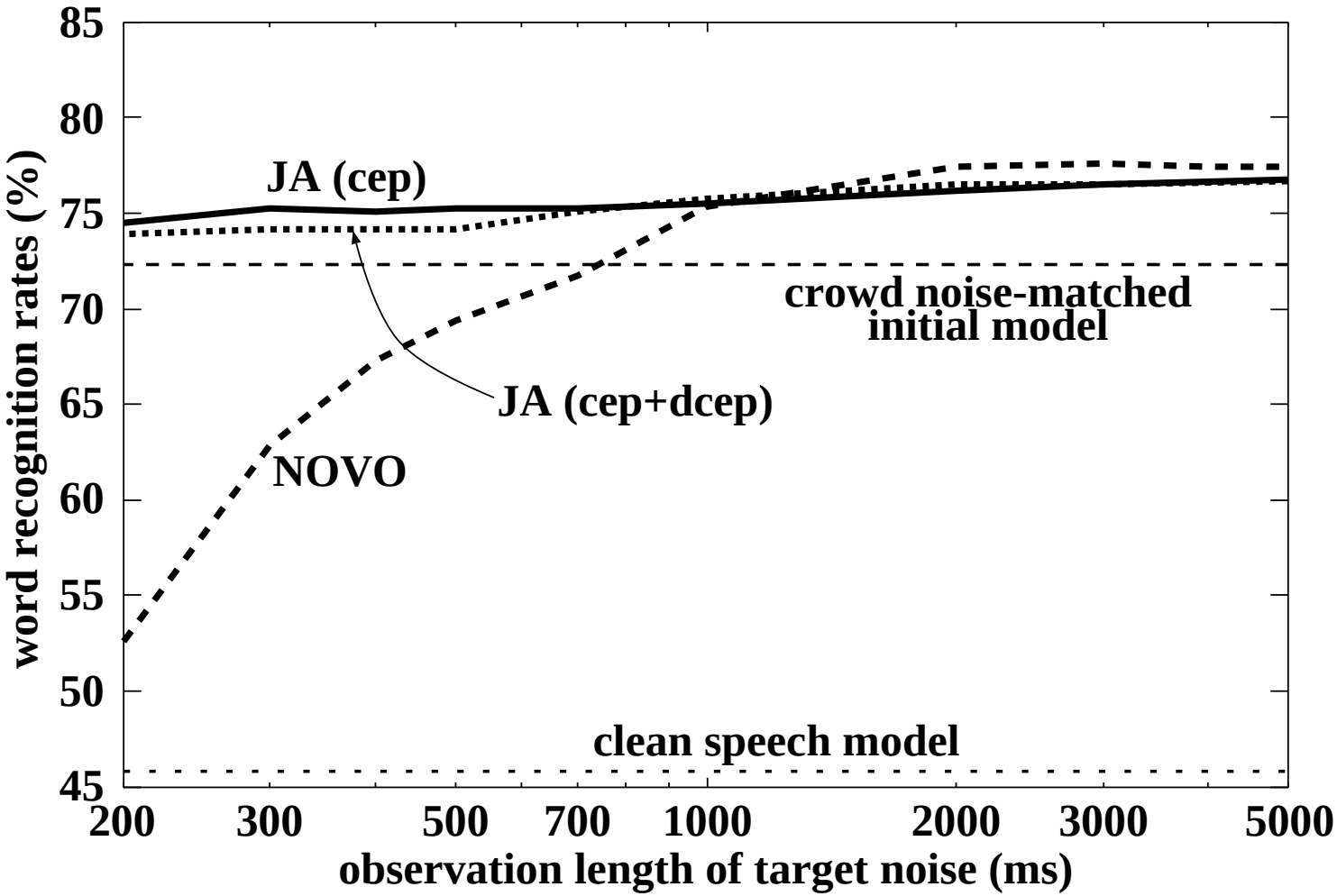


図9. Noise adaptation from the initial model created by NOVO with *Crowd* at 10 dB SNR to the *Exhibition Hall* at 10 dB

ピッチ周波数情報の利用

- ピッチ抽出法 — Lag Window 法 (1978)
  - 音声スペクトルを，それをLag Window によりスペクトルを平滑化したもので割って，フーリエ変換する．精度が高い．
- 単語音声認識とピッチパターンの組合せ (1990 高橋)
  - かんびょう vs かんぴょう，びょういん vs びよういん
- 韻律の情報量 (1991 村上)
  - アクセント/ポーズ情報 ～ かな1字分の情報
- 音素パターンのピッチ周波数依存性 (1991 Singer)
  - ピッチ周波数とスペクトルには相関
- ピッチ周波数パターンによるアクセント句境界推定 (1991 下平)
  - ピッチパターンのクラスタリング，最適セグメンテーション(One-Pass DP)

嵯峨山茂樹 (東京大学大学院情報理工学系研究科システム情報学専攻)

自動翻訳電話

Keyword: 音声翻訳

音声認識・言語翻訳・音声合成をつないで3ヶ国で通信

(+ 大勢 > 50 人)

背景
- **ATR 自動翻訳電話プロジェクト**
- **ハードウェア進歩，安価**
- **NTT 通信業**

自動翻訳電話の概念

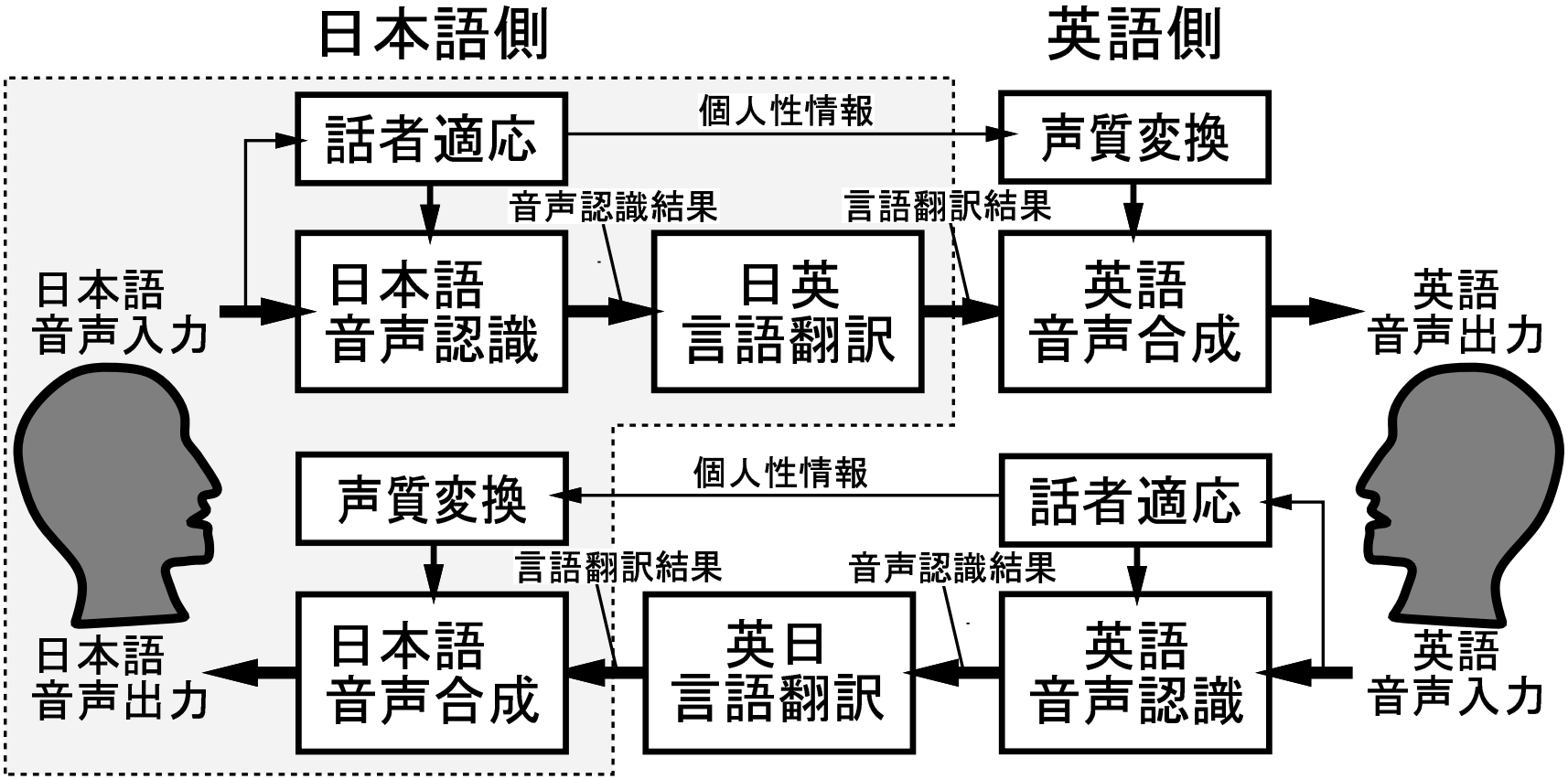


図10. 自動翻訳システムの全体の概念図。日本語音声認識、日本語音声合成、言語翻訳、話者適応、声質変換、に重点を置いて研究した

嵯峨山茂樹 (東京大学大学院情報理工学系研究科システム情報学専攻)

自動翻訳電話の構成要素

- **音声認識** ..... 音声をテキストに変換
  + 話者適応機能
- **言語翻訳** ..... テキストの言語を変換
  3 段階: 解析・変換・生成
- **音声合成** ..... テキストを音声に変換
  + 話者変換機能

嵯峨山茂樹 (東京大学大学院情報理工学系研究科システム情報学専攻)