概要

本研究におけるクロストークとは, 男女 2 話者が別々の孤立単語を同時に発声する状況を想定している.

従来の研究において, クロストーク音声認識を扱う研究は困難が予想されるため, あまり行われていなかった. そこで, 本研究では従来手法によるクロストーク音声認識の精度を調査し, 今後の研究のための基礎データを作成する.

また, モーラ情報を考慮することでピッチ成分を効率的に用いることができ、単語音声認識率が向上すると報告されている. このことから, 本研究ではクロストーク音声認識にもモーラ情報が有効であると仮定して実験を行う.

モーラとは仮名文字単位に相当するもので、単語の仮名文字の個数をモーラ数, 単語の仮名文字の位置をモーラ位置とし, それらをまとめてモーラ情報と定義する. また、アクセントとは発音の高低を指し, 本研究ではモーラ情報にアクセント情報を付加して使用する.

音声認識で用いられているケプストラムは低次にフォルマント情報, 高次にピッチ情報を各々含み, 相互作用があると知られている. そのため, フォルマント情報のみを使用していた従来の音声認識にはピッチ成分の影響がある. しかし, 最近の研究で単語のモーラ位置, モーラ数, アクセント型が決まればピッチ周波数がほぼ決まり, この関係を利用することでピッチの影響を分離できることが知られている.

そこで本研究ではモーラ情報とアクセント情報を使用し, クロストーク音声音声認識における認識精度を調査する.

加えて, 計算機による認識率と比較するため, 人手による聴取実験を行う.

その結果, FBANK よりも MFCC を用いた方が認識率が高く, MFCC においてモーラ情報を使用することで認識率に改善が見られた. しかし, 人手による聴覚実験と比較すると認識率が非常に低いという結果が得られた.

目 次

図 目 次

表 目 次

1　はじめに

会議や緊急時の災害現場など, 様々な場面において, 複数の話者が同時に, 違う声の大きさで発話したとき, 計算機を用いて全ての話者の音声を認識できるシステムの実現が望まれる. しかし, そのようなシステムの実現は困難である. このようなシステムの初歩として, クロストーク音声認識があげられる. このクロストーク音声とは, 2 話者が同時に発声する状況を想定している. しかし, 過去の研究において, クロストーク音声認識は困難さが予想されるためにあまり行われていなかった [1]. そこで本研究では, まず, クロストーク音声を男女 2 話者が別々の孤立単語を同時に発話する状況と想定する. そして, 手始めとして従来手法によるクロストーク音声認識の評価実験を行い, 今後の研究のための基礎データを作成する.

音声信号には主にフォルマントとピッチの 2 つの情報が含まれている. これらを分離するためにケプストラム分析が用いられる. ケプストラム分析において低次にフォルマント成分, 高次にピッチ成分が抽出される. 音声認識において, 一般的にケプストラムがパラメータとして用いられ, これには低次のフォルマント成分のみが含まれている. しかし, ケプストラムは高次のピッチ成分の影響を受けることが知られている.

その一方で, ピッチ周波数とモーラ数, モーラ位置に依存関係が存在することが知られている. また, 最近の研究で単語のモーラ位置, モーラ数, アクセント型が決まればピッチ周波数がほぼ決まり, この関係を利用することでピッチ成分を効率的に用いることができると報告されている.

そこで本研究では, モーラ情報とアクセント情報を使用し, クロストーク音声認識における認識精度を調査する. また, モーラ情報の有効性を確認するため, 音素の隣接情報を考慮したトライフォンモデルでの認識も行う. その後, 計算機の認識率と比較するため人手による聴取実験を行う.

結果として, 特徴パラメータには FBANK よりも MFCC を用いた方が認識率が高く, MFCC モーラモデルにおいて認識率が最も高くなった. しかし, 計算機の認識率は聴取実験と比較して非常に低いという結果が得られた.

2　HMMによる音声認識

2.1　音声認識の歴史

人間は, 日常のコミュニケーションの大半を音声を介して行う. 人と計算機のインターフェースを人にとって容易かつ自然なものにするには音声メディアの利用と情報処理技術にかかっている. 特に, 計算機による音声認識技術が中核を担うことになる.

研究の歴史的な流れとして, 1970年前後に, 日本で動的計画法に基づく時間軸非線型伸縮アルゴリズム (DTW) や線形予測分析 (LPC) などの, 現在でも音声認識技術の基礎となっている分野の先端的な研究がなされた. その後, 1980年代に, 現在の音声認識アルゴリズムの基本となっている統計的な時系列モデルの隠れマルコフモデル (HMM) の研究が米国で始められた. HMMは, その統計的アルゴリズムの高い学習能力と認識性能により, 広く使われるようになってきている.

2.2 音響分析

2.2.1 特徴抽出

音声認識を行うためには, まず, 音声区間の検出を行うことが必要である. そして尤度 $(y \mid w)$ を計算するには, 音声区間の時系列データ y の表現形式を決定する必要がある. 音声波形そのものを用いたのでは情報量が多過ぎ, 波形の位相情報は伝送系や録音系によって変わりやすい上, 人間による音声の知覚にはほとんど寄与しない. このことから, 位相情報はむしろ取り除いたほうがよい. このため, 音声波から一定周期毎に短時間スペクトル (密度) を抽出して用いることが多い. 現在短時間スペクトル分析の手法としては, 手結行フィルタ群を用いる方法,FFT を用いて直接的にスペクトルを計算する方法, 相関関数を用いる方法, および LPC 分析を基礎とする方法の 4 種類にわけることができる.[5]

2.2.2 FBANK

帯域フィルタは, ハードウェアによる実時間分析の実現が容易なため, 古くから用いられている. 人の聴覚は, 悪との高さに関して, メル (mel) 尺度と呼ばれる対数に近い非線型の特性を示し, 低い周波数では細かく, 高い周波数では新井周波数分解能力を持つ.

FBANK は音声周波数に対して FFT スペクトルを求め, メル分割されたフィルタに通し, その対数パワーを求めたもので, 特徴パラメータにフォルマント成分及びピッチ成分が含まれる.

FBANK は混合ガウス分布に Full-covariance を用いた場合に MFCC よりも認識率が高いことが知られている [2].

本研究において, 基本周波数 16KHz の音に対して FBANK24 次+ Δ FBANK24 次の形で用いる.

2.2.3 MFCC

FFT によって計算されたそのままのスペクトルの類似度を用いることも可能であるが, スペクトルの微細構造はピッチ等の影響を受けて不安定なため, これを平滑化したスペクトラル崩落を用いることが多い. この平滑化の手法として良く知られているものにケプストラムによる方法がある.

MFCC は, まず音声周波数に対して FFT スペクトルを求め, メルスケール上に等間隔に配置された帯域フィルタバンクの出力を抽出する. そして, 対数変換を行い, 逆フーリエ変換することにより得られるケプストラム係数である.

高次においてピッチ成分, 低次においてフォルマント成分が見られ, 通常は扱いやすさの観点から低次のフォルマント成分が使用される. これは, 言い換えれば声道特性のみを用いていることになる.

本研究では MFCC12 次+ Δ MFCC12 次の計 24 次の形で用いる.

2.3 HMM

2.3.1 HMM とは

HMM(隠れマルコフモデル) とは, 外から観測できるものがモデルによって生成された出力データ系列だけであって, 一般にモデルの内部の状態とその遷移の様子は外から見られないことから付けられた呼称である.

音声パターンは時系列の形で表され, 様々な原因により変動がある. 音響パラメータの時系列は変動分を含み, このようなパターンの確率的な性質は HMM によって精密に表現できると考えられている. HMM は非定常信号源を定常信号源の連結で表す.

2.3.2 HMM と音声認識の問題

入力音声パターンを I フレームの時系列として, $X = x_1, x_2, \cdots, x_I$ と表す.

音声認識の問題は, X を観測して最もよくマッチする単語列 $W = w_1, w_2, \cdots, w_N$ を見つけ出す問題として単純化できる. N は単語列における単語数を表す.

このように問題を設定すると, 音声認識は P(W | X) を最大にする単語列 W を見つけ出す問題となる.

音声認識に用いられる HMM は, left-to-right 型で 1 つの初期状態と 1 つの最終状態がある構造が多い. ベイキス (Bakis) モデルと呼ばれる型の例を図 1 に示す.

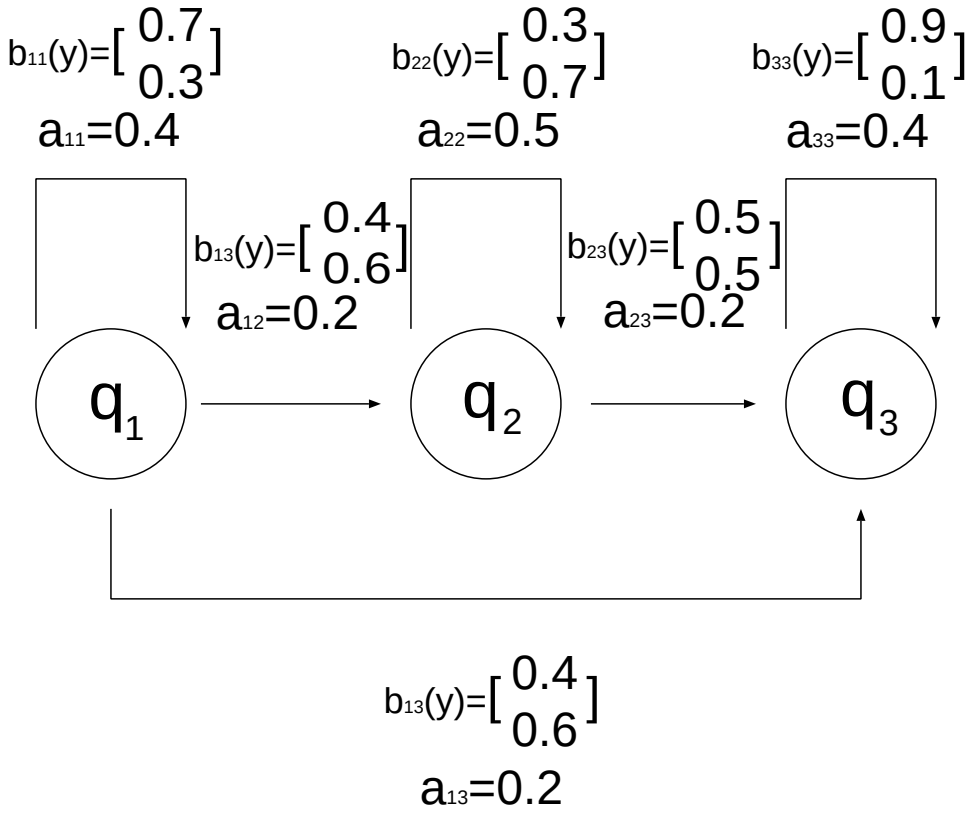

図 1: ベイキス型 HMM の例

図 1 の状態遷移のアークに付けられた数値 a_{ij} は, 状態 q_i から状態 q_j への遷移確率を表し, 状態数を S とすると $S \times S$ の行列で表現できる. 音声パターンには, 時間的に非可逆性の性質があるので, $i > j$ なら $a_{ij} = 0$ である. 角状態 q_i の初期確率を π_i で表し, 最終状態の集合を F で表す.

$b_{ij}(y)$ は状態 q_i から状態 q_j への遷移で, スペクトルパターン y が観測 (出力) される観測 (出現) 確率を表し, $b_{ij}(y)$ を出現確率行列と呼ぶ. 出現するスペクトルパターンに関しては, 連続値として表す場合 (連続分布, 連続 HMM) と, 有限個 (K 個) のシンボルの組合せで表現する場合 (離散分布モデル, 離散 HMM) がある. 図 1 における数値例は, 離散分布モデルで出力符号ベクトルを a, b の 2 つに限り, 図の [ ] 内にそれぞれの出現確率を示している. この例の場合, 遷移確率行列は以下のようになり, $\pi_1 = 1, \pi_i = 0 (i > 1), F = q_3$ である.

$$\mathrm{A} = (\mathrm{a}_{ij}) = \begin{bmatrix} 0.4 & 0.4 & 0.2 \\ 0.0 & 0.5 & 0.2 \\ 0.0 & 0.0 & 0.4 \end{bmatrix} \quad (2.1)$$

実際の音声認識に用いる HMM においては, 対象に応じて適切に状態数やモデル構造 (遷移構造) を決定し, スペクトルパターンの表現法 (離散分布モデルの場合はその種類 K, 連続分布モデルの場合はそのモデル化の方法) を決定する必要がある.

2.3.3 HMM 法の利点と問題点

HMM が音声認識において有利な点を以下に示す.

- 個人差や調音結合, 発声法 (強さ, 速さ, 明瞭さ) 等による音声パターンの変動を確率モデルで捉え, 統計的処理で対処できる.
- 従って, 統計理論や情報理論/確率仮定論による理論展開がしやすい.
- 比較的簡単なモデルのパラメータ推定法が知られている.
- 言語レベルの処理も音響処理部と同様に確率モデルで表現できるため, 両者を統合しやすい.
- 認識時の計算量は比較的少ない.

HMM が音声認識における問題点を以下に示す.

- モデルの設計法が確立されていないため, 試行錯誤的/ノウハウ的要素が強い.
- HMM のパラメータ推定に多量の学習用サンプルを必要とし, 計算量も多い.
- 音声の過渡的パターンの表現力に乏しい.
- 時系列パターンの２時点におけるパターンの壮観が考慮できない.

2.4 認識アルゴリズム

$y = y_1, y_2, \cdots, y_T$ を観測 (出力) 系列とする. 具体的には, スペクトルやケプストラムの時系列である. このとき, 各 HMM モデルによって y が生起する確率 (尤度)P(y | M)(M は HMM によって表現される単語や音素に対応) を求め, 最大確率 (最大尤度) を与えるモデルを選出しこれを認識結果とする. 図 2 に単語 HMM を用いた認識方法を示す.

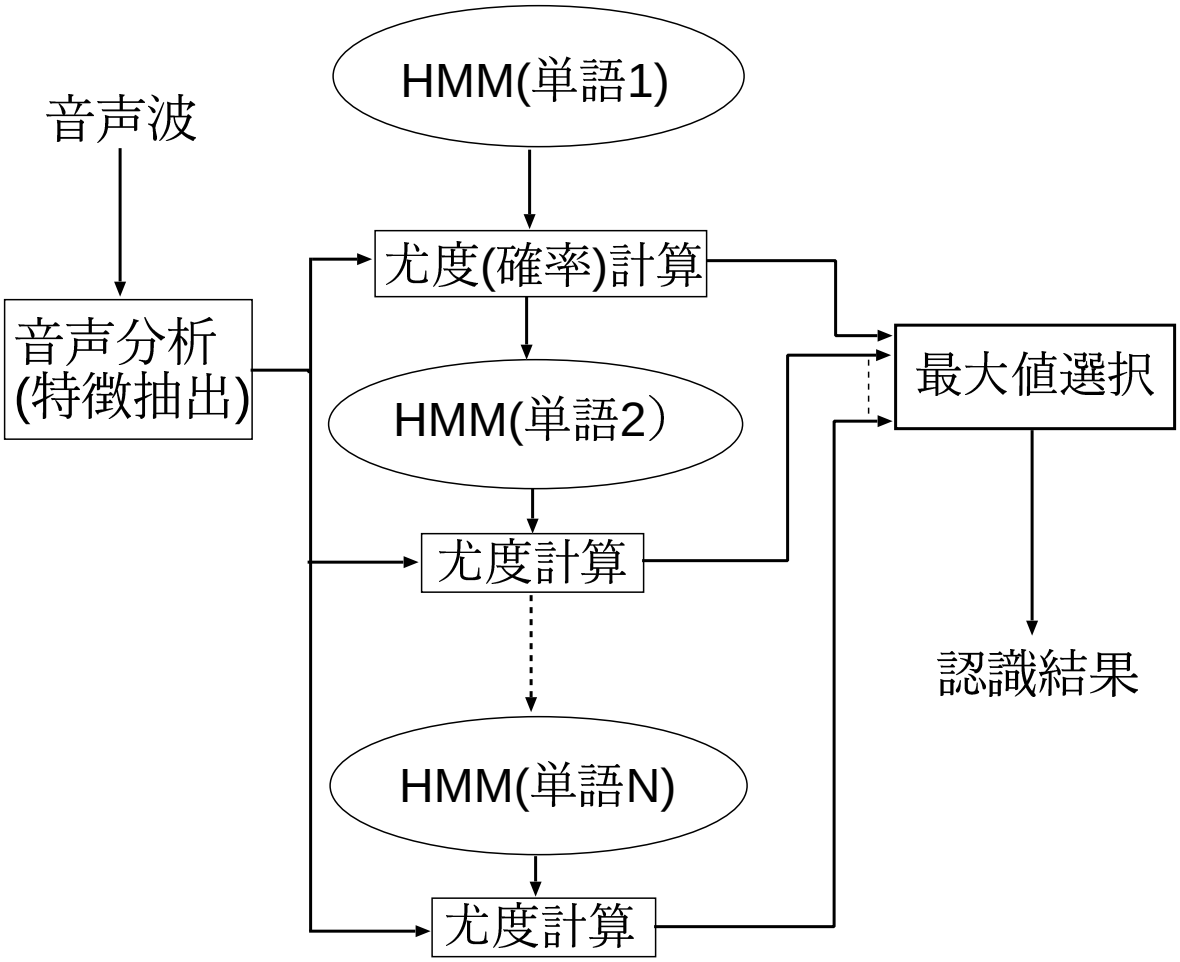

図 2: 単語 HMM を用いた単語音声認識の方法

$q = q_{i0}, q_{i1}, \cdots, q_{iT}$ を状態遷移行列 (ただし $q_{iT} \in F$) とすれば,

$$\mathrm{P(y \mid M)} = \sum_{i_0, i_1, \cdots, i_T} P(y \mid q, M) \cdot P(q \mid M) \quad (2.2)$$

と表すことができる. そして一般的に $P(y \mid M)$ の値は, トレリスアルゴリズムで求められる.

フォワード変数 $\alpha\ (i,t)$ を定義し, 符号ベクトル y_t を出力して状態 q_t にある確率とすれば, $i = 1, 2, \cdots, S$ とおいて, 以下の式を得る.

$$\alpha\ (\mathrm{i,\ t}) = \sum_j \alpha\ (j, t-1) \cdot \alpha_{ji} \cdot b_{ji}(y_t)(t-1, 2, \cdots, T) \cdot \pi_i(t=0) \quad (2.3)$$

これを計算し, 最後に以下を求めれば良い.

$$\mathrm{P(y \mid M)} = \sum_{i, q \in F} \alpha\ (i, T) \quad (2.4)$$

2.4.1 Viterbi アルゴリズム

Viterbi アルゴリズムは, モデルの最適な状態遷移行列 (際的経路) と, この経路上での確率を求めるアルゴリズムである.

$P(y \mid M)$ を厳密に求めないで, 近似的に, モデル M が符号ベクトル系列 y を出力するときの, 最も可能性の高い状態系列上での出現確率を用いることを考える. この出現確率 (尤度) は, 各遷移での確率値を対数変換しておくことにより, 加算と大小判定のみからなる DP 演算によって高速に求めることができる. 対数を用いた計算なので, トレリス法を用いる場合に比べ, 計算値のダイナミックレンジが小さくてすみ, アンダーフローの問題が解消できる. また, 計算量が少ないにもかかわらず, 音声認識性能はトレリス法とほとんど変わらないことが実験的に確認されている [5].

Viterbi アルゴリズムは, 本実験において HMM の初期モデル作成と認識に使用されている.

2.4.2 Baum-Welch アルゴリズム

Baum-Welch アルゴリズムとは, 学習データの尤度を最大にするようにパラメータを学習する方法で, 基本的には gradient 学習によってパラメータを収束させる方法である.

学習用音声として, N 個の観測シンボル系列 $y_1^T(n) = y_1, y_2, \cdots, y_T(n)_{n=1}^N$ が与えられたとして, 以下の式を考える.

$$\Pi\, {}_{n=1}^{N} P(y_1^T(n) \mid \pi_{\,i}, a_{ij}, b_{ij}(k)) \quad (2.5)$$

これを最大化するパラメータセット $\pi_{\,i}, a_{ij}, b_{ij}(k)$ は, 次のような Baum_Welch アルゴリズムによって推定することができる.

バックワード変数 $\beta\ (i < t)$ を, 時刻 t に状態 q_i にあって, 以後観測シンボル y_{t+1}^T を出力する確率, $\gamma\ (i, j, t)$ を, モデル M が y_i^T を出力する場合において, 時刻 t に状態 q_i から q_j へ遷移し, シンボル y_t を出力する確率と定義する. このとき, 以下の関係が得られる.

$$\beta\ (\mathrm{i,\ T}) = \begin{cases} 1 & \mathrm{q}_i \in F \\ 0 & \mathrm{q}_i \notin F \end{cases} \quad (2.6)$$

$$\beta\ (\mathrm{i,\ t}) = \textstyle\sum_j a_i j \cdot b_i j(y_t) \cdot\ \beta\ (j, t+1) \quad (2.7)$$

$$(t = T, T-1, \cdots, 1; i = 1, 2, \cdots, S)$$

$$\gamma\ (\mathrm{i,\ j,\ t}) +\ \alpha\ (\mathrm{i,\ t\text{-}1}) \cdot \mathrm{a}_{ij} \cdot b_{ij}(y_t) \cdot\ \beta\ (j, t) \frac{}{P(y_1^T \mid M)} \quad (2.8)$$

これらを用いて, パラメータ $\pi_{\,i}, a_i j, b_{ij}(k)$ を, 次の再推定のくり返しによって求める.

$$\pi_{\ i} = \frac{\sum_j \gamma\ (i,j,1)}{\sum_i \sum_j \gamma\ (i,j,1)} \quad (2.9)$$

$$\mathrm{a}_i j = \frac{\sum_{t=1}^{T} \alpha\ (i,t-1) \cdot a_{ij} \cdot b_{ij}(y_t) \cdot \beta\ (j,t)}{\sum_t \alpha\ (i,t) \cdot \beta\ (i,t)} \qquad = \frac{\sum_t \gamma\ (i,j,t)}{\sum_t \sum_j \gamma\ (i,j,t)} \quad (2.10)$$

$$\mathrm{b}_{ij}(k) = \frac{\sum_{t,y_t=k} \gamma\ (i,j,t)}{\sum_t \gamma\ (i,j,t)} \quad (2.11)$$

実際には, すべての学習用サンプルに関してこの計算を行ってから, パラメータを 1 回更新する, というサイクルを値が収束するまでくり返して, パラメータの値を決定する. Baum-Welch アルゴリズムは, HMM 初期モデルの再推定に使用されている.

3　評価モデル

特定話者の単語の発声において, 単語のモーラ数, モーラ位置, アクセント型が決まれば単語によらず, ピッチ周波数はほぼ一定であることが知られている. モーラ情報とピッチ周波数の依存関係を利用することで, フォルマントを示すケプストラムの低次の項に対するピッチの影響を分離でき, これにより認識率の改善が期待できる.

それをふまえ, 基本モデル, 基本モデルにモーラ情報を考慮したモーラモデル, モーラモデルにアクセントを考慮したモーラアクセントモデル, また, モーラ情報の有効性と比較するため, 従来の音声認識で有効とされるトライフォンモデルの計 4 種類の HMM モデルを用意する.

3.1　モーラモデル

モーラとは仮名文字単位に相当し, 単語の仮名文字の個数をモーラ数, 仮名文字の位置をモーラ位置とし, それらをまとめてモーラ情報と定義する.

本研究ではデータベースの音声ラベルファイルに含まれる母音, 促音, 撥音をモーラ情報使って分類する. 具体的には母音, 促音, 撥音の後に 2 桁の数字でモーラ位置, さらに 2 桁の数字でモーラ数を付け加えて分類する. 比較のため, 分類例を他のモデルの分類例とともに表 1 に示す.

表 1: ラベル付けの例

基本ラベル	t	e	i	sh	i
モーラ	t	e0104	i0204	sh	i0303
モーラアクセント	t	e0401000	i0402001	sh	i0303001
トライフォン	pau-t+e	t-e+i	e-i+sh	i-sh+i	sh-i+pau

音声ラベルが’”teishi”の場合, 単語のモーラ数は 3 なので, 母音, 促音, 撥音の後方にそれぞれのモーラ位置, モーラ数 03 の順につける. また, 3 番目と 5 番目の i はモーラ位置が異なるため異なった音素として扱う.

3.2　モーラアクセントモデル

本研究において, アクセントとは発音の高低を指し, モーラ情報にアクセント情報を付加する形で使用する. 具体的には, モーラモデルの後に 2 桁の数字で高低の推移を示すアクセント型, その次に,1 桁の数字で高低情報を追加し, 分類する. 尚, 高低情報は低ければ 0, 高ければ 1 として表す.

表 1 の例では, 低い状態で始まり, 一度高くなった後は下がらない, という推移形式の 0 型であり, 最後の 1 桁の数字でその音素の高低を示している.

3.3　トライフォンモデル

トライフォンモデルとは前後の音素との隣接情報を考慮したモデルである. このモデルは従来の音声認識において有効な手法とされるため, モーラ情報との比較対象として使用する.

具体的には, ある音素に対し, その前に前状態の音素, 後に次状態の音素を付加する形で分類する.

尚, 表 1 の例にある pau は無音を表す.

4　評価実験

2 話者分の単語データベース (以下 DB) から, 偶数番をランダムに 1 単語ずつ抽出し, 合成することでクロストーク音声を作成する. 次に, 同単語 DB の偶数番を学習データとして話者毎に HMM を作成し, 話者毎に音声認識を行うことでクロストーク音声を認識する. クロストーク音声認識の手順を図 3 に示す.

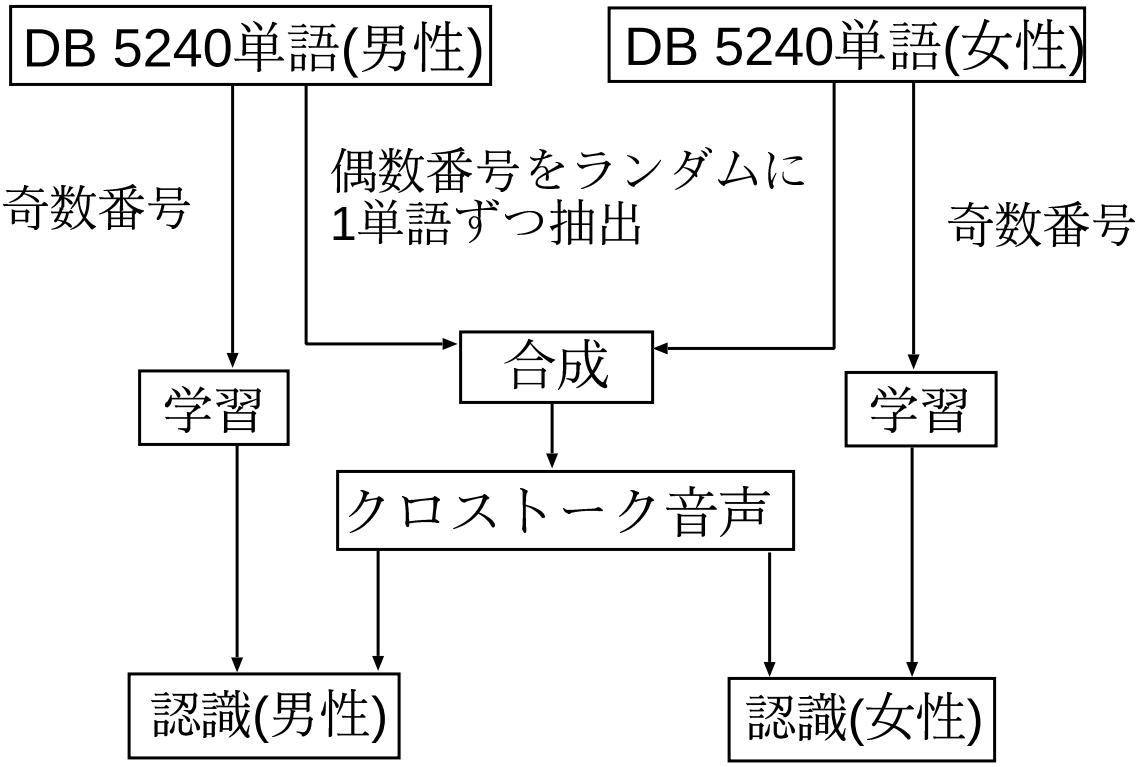

図 3: クロストーク音声認識の手順

4.1 クロストーク音声の作成

本研究において, クロストーク音声とは男女 2 話者が別々の孤立単語を同時に話す状況を想定している. 音声データベース (以下 DB) として, ATR の単語発話 DB Aset(1 話者につき 5240 単語) 男女各 2 名 (男性:mau, mms , 女性:ftk, faf) を使用する. 5240 単語を奇数番, 偶数番に分け, 奇数番を学習データとする. 2 話者分の偶数番 2620 単語からランダムに 1 単語ずつ抽出し, 2 つの音声を重ね合わせてクロストーク音声 2620 単語を作成する. 発話時間の長さの違いは考慮せず, 発話時間の長い方にあわせる.

4.2 実験条件

本実験では認識に HTK を使用し, 実験環境は表 2 にまとめる. 特徴パラメータには MFCC と FBANK を使用する. また, 音素 HMM の混合ガウス分布には Diagonal-covariance を使用する. MFCC と FBANK において, 基本モデル, モーラモデル, モーラアクセントモデル, トライフォンモデルの 4 種の HMM モデルを用意し, 計 8 種類で実験を行う. 本研究では, パラメータを一定とするため, 半連続型で基本モデルを作成する. そして, 初期モデルが重要なため, 基本モデルからモーラモデル, モーラアクセントモデル, トライフォンモデルを作成する.

表 2: 実験条件

	MFCC	FBANK
基本周波数	16kHz	
分析窓	Hamming 窓	
分析窓長	25ms	
フレーム周期	10ms	
音響モデル	3 ループ 4 状態・半連続分布型	
stream 数	1	
特徴パラメータ	MFCC 12 次 +Δ MFCC 12 次	FBANK 24 次 +Δ FBANK 24 次

4.3 評価方法

通常, クロストーク音声の認識では重ね合わせる前の音声両方をそれぞれ認識する必要がある.

本研究では組み合わる前の 2 つの音声それぞれの認識率の平均を求め, さらにモデル毎の平均を比較する.

4.4 実験結果

4.4.1 MFCC と FBANK の比較

MFCC と FBANK の実験結果の比較を表 3 に示す.

表 3: 実験結果 (MFCC と FBANK の比較)

モデル	MFCC	FBANK
基本モデル	23.3%	12.7%
モーラモデル	24.0%	12.6%
モーラアクセントモデル	18.0%	9.7%
トライフォンモデル	21.5%	6.7%

この表から, 特徴パラメータには FBANK よりも MFCC を用いた方が認識率が高いことがわかる. また, モーラモデルを用いた場合, MFCC では認識率が改善され, 本実験において最も認識率が高くなっているが, FBANK では基本モデルよりも認識率が低下している. そして, モーラアクセントモデル, トライフォンモデルでは MFCC, FBANK ともに認識率の低下が見られた.

4.4.2 MFCC の詳細

MFCC の結果の詳細を表 4 に示す.

表 4: MFCC の認識結果

mau+ftk				
話者	基本	モーラ	モーラアクセント	トライフォン
mau	22.91%	25.39%	20.62%	25.89%
ftk	19.17%	18.59%	13.55%	16.15%
平均	21.04%	21.99%	17.08%	21.07%
mau+faf				
話者	基本	モーラ	モーラアクセント	トライフォン
mau	8.82%	9.51%	7.29%	10.00%
faf	36.27%	35.47%	28.29%	32.99%
平均	22.54%	22.49%	17.79%	21.49%
mms+ftk				
話者	基本	モーラ	モーラアクセント	トライフォン
mms	20.08%	20.47%	15.20%	18.71%
ftk	20.73%	22.41%	16.49%	22.34%
平均	20.4%	21.44%	15.84%	20.52%
mms+faf				
話者	基本	モーラ	モーラアクセント	トライフォン
mms	3.89%	4.05%	3.13%	4.85%
faf	45.32%	45.90%	38.60%	40.9%
平均	29.10%	29.97%	21.36%	22.87%
全体の平均				
	基本	モーラ	モーラアクセント	トライフォン
平均	23.27%	23.97%	18.01%	21.48%

表より, 組み合わせにより認識率にばらつきが見られるが, 全体の平均を見た場合, モーラ情報を使用することで, 若干ながら認識率の改善が見られる. しかし, アクセントを考慮した場合は全パターンで認識率が低下した. トライフォンもでるでは, 音量差が小さいパターンでは認識率が改善されたが, 逆に音量差が大きいパターンでは認識率が低下し, 全体として基本モデルよりも認識率が低下している.

尚, MFCC モーラモデルの認識結果の一部を付録として付ける.

4.4.3 FBANK の詳細

FBANK の実験結果の詳細を表 5 に示す.

表 5: FBANK の認識結果

mau+ftk				
話者	基本	モーラ	モーラアクセント	トライフォン
mau	19.55%	20.35%	17.22%	5.12%
ftk	6.68%	6.26%	3.86%	3.86%
平均	13.11%	13.30%	10.54%	4.49%
mau+faf				
話者	基本	モーラ	モーラアクセント	トライフォン
mau	6.99%	6.38%	5.23%	3.97%
faf	14.59%	13.78%	10.27%	8.17%
平均	10.79%	10.08%	7.75%	6.07%
mms+ftk				
話者	基本	モーラ	モーラアクセント	トライフォン
mms	15.04%	15.69%	12.29%	7.87%
ftk	7.41%	7.52%	6.07%	3.97%
平均	11.22%	11.6%	9.18%	5.92%
mms+faf				
話者	基本	モーラ	モーラアクセント	トライフォン
mms	3.63%	3.17%	2.02%	1.95%
faf	27.29%	27.33%	21.83%	18.35%
平均	15.46%	15.25%	11.42%	10.15%
全体平均				
	基本	モーラ	モーラアクセント	トライフォン
平均	12.64%	12.55%	9.72%	6.65%

表より, モーラモデルで若干認識率が改善されたパターンがあるものの, どのモデルでも基本モデルと比較して認識率の低下が見られた. 下が見られた.

尚, FBANK 基本モデルの認識結果の一部を付録として付ける.

5　考察

5.1　特徴パラメータによる不具合

クロストーク音声認識では, 片方の話者を認識する際, もう片方の話者の音声が雑音となってしまう. FBANK はパラメータにピッチ成分を含むため, MFCC よりも影響が大きくなってしまい, 認識率に差が現れたと考えられる.

しかし, 本研究では特徴パラメータに Diagonal を使用したが, FBANK は Full-covaqriance を用いた方が認識率が高いことが知られている.

このことから, Full-covariance を用いた再実験を行うことで FBANK の認識率改善が見込まれる.

5.2 モーラ情報による改善

MFCC モーラモデルを用いた場合に, 話者 mau と話者 ftk の組合せの内, 話者 mau で改善が見られた単語の一部を表 6, 話者 mms と話者 faf の組合せの内, 話者 faf で改善が見られた単語の一部を表 7 に示す. 逆に認識できなくなった単語例を表 8, 表 9 に示す.

表 6: 改善された単語例 (mau+ftk/mau)

組み合わせた音声	基本	モーラモデル
質問 (shitsumoN) + (shimeN)	注文 (choumeN)	質問 (shitsumoN)
辞表 (jihyou) + 活動 (katsudou)	事故 (jiko)	辞表 (jihyou)
災難 (sainaN) + 実演 (jitsuen)	海岸 (kaigaN)	災難 (sainaN)

表 7: 改善された単語例 (mms+faf/faf)

組み合わせた音声	基本	モーラモデル
落とす (otosu) + 植物 (shokubutsu)	国立 (kokuritsu)	植物 (shokubutsu)
新年 (shiNneN) + 投票 (touhyou)	根拠 (koNkyo)	投票 (touhyou)
誘う (sasou) + 拾う (hirou)	潜む (hisomu)	拾う (hirou)

表 8: 誤認識するようになった単語例 (mau+ftk/mau)

組み合わせた音声	基本	モーラモデル
探す (sagasu) + 祭日 (saijitsu)	探す (sagasu)	剥す (hagasu)
潜る (moguru) + のこぎり (nokogiri)	潜る (moguru)	申し込む (moushikomu)
駆けつける (kaketsukeru) + 率 (ritsu)	駆けつける (kaketsukeru)	受け付ける (uketsukeru)

表 9: 誤認識するようになった単語例 (mms+faf/faf)

組み合わせた音声	基本	モーラモデル
追い抜く (oinuku) + 行方 (yukue)	行方 (yukue)	湯気 (yuge)
遺言 (yuigoN) + 伝える (tsutaeru)	伝える (tsutaeru)	使える (tsukaeru)
真似 (mane) + いつも (itsumo)	いつも (itsumo)	以後 (igo)

モーラ情報を使用することで, 母音, 促音, 撥音が連続する場合に対して改善された例が多く見られた. 逆に, 基本モデルで認識できていた単語が, 同一の音素を持つ別の単語として認識される現象が見られた.

5.3 人手による聴取実験

計算機による認識率と比較するため, 話者 mau と話者 ftk, 話者 mms と話者 faf の 2 種類のクロストーク音声に対して人手による聴取実験を行った. 結果を表 10 に示す. 尚, 被験者は男性 1 名である.

表 10: 聴取実験の認識率

組合せ	男性話者のみ認識	女性話者のみ認識	男女両方を認識
mau+faf	84.23%	79.0%	67.09%
mms+faf	76.1%	92.63%	70.41%
平均	80.12%	85.81%	68.75%

人手による聴取実験と比較すると, 計算機による認識率は非常に低い結果となった.

また, 人手でも認識できない音は, 2 話者の子音が重なる音の組合せであった. 逆に, 音が重ならなければ概ね認識できた. 話者 mau と話者 ftk のクロストーク音声に関して, 人手と計算機 (MFCC モーラモデル) での認識結果の比較を表 11, 12 に示す.

表 11: 例：計算機で認識, 人手で誤認識

正解 (男性/女性)	潜む (hisomu)	くすぶる (kusuburu)
計算機の結果	潜む (○)	苦痛 (kutsuu)
聴取実験結果	競う (kisou)	薬 (kusuri)
正解 (男性/女性)	訴える (uqtaeru)	快活 (kaikatsu)
計算機の結果	訴える (○)	快活 (○)
聴取実験結果	訴える (○)	解決 (kaiketsu)
正解 (男性/女性)	性質 (seishitsu)	対決 (taiketsu)
計算機の結果	平日 (heijitsu)	対決 (○)
聴取実験結果	性質 (○)	締結 (teiketsu)

計算機で認識できたが, 人手で聞き取れなかった例には, 音素が重なる組合せであった. この場合, 人手では子音を混同してしまうことが多くなり, 結果として誤認識してしまっていた. また, このとき音量差が大きいと, 片方の音が全く聞こえないという場合も見られた.

表 12: 例：人手で認識, 計算機で誤認識

正解 (男性/女性)	似合う (niau)	作物 (sakumotsu)
計算機の結果	付き合う (tsukiau)	三月 (saNgatsu)
聴取実験結果	似合う (○)	作物 (○)
正解 (男性/女性)	窮屈 (kyuukutsu)	全長 (zenchou)
計算機の結果	流暢 (ryuuchou)	現象 (geNshou)
聴取実験結果	窮屈 (○)	現象 (geNshou)
正解 (男性/女性)	適当 (tekitou)	頭 (kasira)
計算機の結果	太鼓 (taiko)	頭 (○)
聴取実験結果	適当 (○)	欠片 (kakera)

人手で認識できたが, 計算機で認識できなかったものには, 音が重なっていないものが多く見られた. この場合, 計算機では最初と最後の音素はある程度認識しているが, その途中の音素を認識できず, 結果として誤認識となっている. 逆に, 人手では概ね認識することができた.

6　結言

本研究では, 男女 2 話者が別々の単語を同時に話したと仮定し, 従来手法を用いることでどの程度認識できるか調査した.

その結果, 特徴パラメータに MFCC を用いることで FBANK より良い結果が得られた. また, MFCC においてモーラ情報を使用することで認識率に改善が見られた. しかし, 人手による聴取実験と比較すると認識率が非常に低い.

今後の課題として, Full-covariance での認識, 音量差を調整した状態での認識率調査があげられる. また, 今後の方針として, 今回得られた結果を基礎資料としてスペクトルサブストラクション等の雑音抑制に使用されている手法の適用や, 合成音声による学習, 3 次元 Viterbi による認識等を行っていきたい.

7　謝辞

最後に, 本研究において御指導を賜わりました 池原 悟 教授, 村上 仁一 助教授, 徳久 雅人 助手, ならびに本学大学院生の石田 隆浩氏, 加藤 琢也氏に厚く御礼を申し上げるとともに, 計算機 C 研究室のみなさまの多大なる御協力に感謝の意を表します.

参考文献

[1] 中野 晃:クロストーク孤立単語音声認識，鳥取大学大学院工学研究科修士論文，(2002)

[2] 谷口 勝則，他:モーラ情報を用いたフィルタバンクによる孤立単語認識，信学技報，SP2002-131，pp.63-68(2002-12)

[3] 妹尾 貴宏，他:モーラ情報を用いた単語音声認識の検討，信学技報，SP2002-130，pp.55-61(2002-12)

[4] Steve Young，etc:HTK Ver3.2 reference manual,2002 Cambridge University

[5] 古井:音声情報処理, 森北出版株式会社