平成14年度　修士論文

クロストーク孤立単語音声認識

英文タイトル

指導教官　池原 悟　　教授
村上 仁一　助教授
徳久 雅人　助手

鳥取大学大学院 工学研究科 知能情報工学専攻
M01T2021M　中野晃

内容梗概

会議など，様々な人々が同時に会話などをする場面における認識，システムの実現が望まれる．本研究ではクロストークの音声認識評価実験を行う．
本研究においてクロストーク音声とは，同一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．この条件はクロストークの中でも最も厳しい条件の 1 つと考えられる，この条件に対応できればこれよりも緩い条件に対しても可能と考えられる．

音声信号には主に，フォルマントとピッチの 2 つの成分が含まれている．一般的に音声認識ではフォルマントが利用される．

本研究ではピッチを利用することを考え，特徴パラメータに FBANK を用いる．重ね合わせた音声において，それぞれの音声のピッチが分離されて抽出される可能性があるからである．

FBANK はモーラ数，モーラ位置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，混合ガウス分布に Full-covariance を用いた場合，M FCC より音声認識率が高いことが報告されている．

そこで本研究では，クロストーク音声の認識において，特徴パラメータに FBANK を用いさらにモーラ情報を加えて使用することで，有効性の検証を行った．

検証の結果，一般的に使用されている MFCC と比較して FBANK パラメータは認識率が向上した．また，モーラ情報は女性話者において改善が見られた．クロストーク音声の人手による聴取実験と比較した場合，認識率は低い結果となった．

目 次

図版目次

表目次

第1章　はじめに

会議などさまざまな場面において人々が同時に会話などをする．このような場面において同時に違う声の大きさで話された時，計算機を用いて両方の音声を認識，もしくはその一歩手前のレベルとして，片方の音声例えば声の大きい側などを認識できるシステムの実現が望まれる．しかし，そのシステムの実現はなかなか難しい．

そこで本研究ではその手始めとしてクロストーク音声認識評価実験を行う．本研究においてクロストーク音声とは，現実にはありえないが，同一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．クロストーク音声としてデータベースからランダムに抽出した 2 つの音声を重ね合わせて作成する．そして従来手法を用いてどの程度の認識率が得られるか評価する．

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の項にフォルマント，高次にピッチが抽出される．一般的に音声認識において低次のフォルマントがパラメータとして用いられている．しかしケプストラムの低次の項は高次のピッチの影響を受けることが知られている．クロストーク音声は重ね合わせた音声であるため，パラメータとしてケプストラムを用いると不安定になると考えられる．

一方，ピッチ周波数と単語のモーラ数，モーラ位置に依存関係が存在していることが報告されている [3]．FBANK はモーラ数，モーラ位置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，混合ガウス分布に Full-covariance を用いた場合，MFCC より音声認識率が高いことが報告されている [2]．

重ね合わせた音声において，それぞれの音声のピッチが分離されて抽出される可能性からピッチを利用することが有効ではと考え本研究では，特徴パラメータに FBANK を利用する．

クロストーク音声の認識において，特徴パラメータに FBANK を用いさらにモーラ情報を加えて使用することで，パラメータの有効性の検証を行う．

第2章　HMMによる音声認識

2.1　音声認識における歴史的流れ

我々は，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/情報処理技術にかかっている．とくに，計算機による音声認識技術を有効に利用することが，鍵となる．

研究の歴史的流れとして，1970 年前後に，日本で動的計画法に基づく時間軸非線形伸縮アルゴリズム (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) の研究が米国ではじめられた．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 [5]．

2.2　HMM

2.2.1　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は，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繊維の様子は外から見られないことから呼ばれている．音声パターンは時系列の形で表され，さまざま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．

2.2.2　HMM と音声認識

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 2.1 に示す．

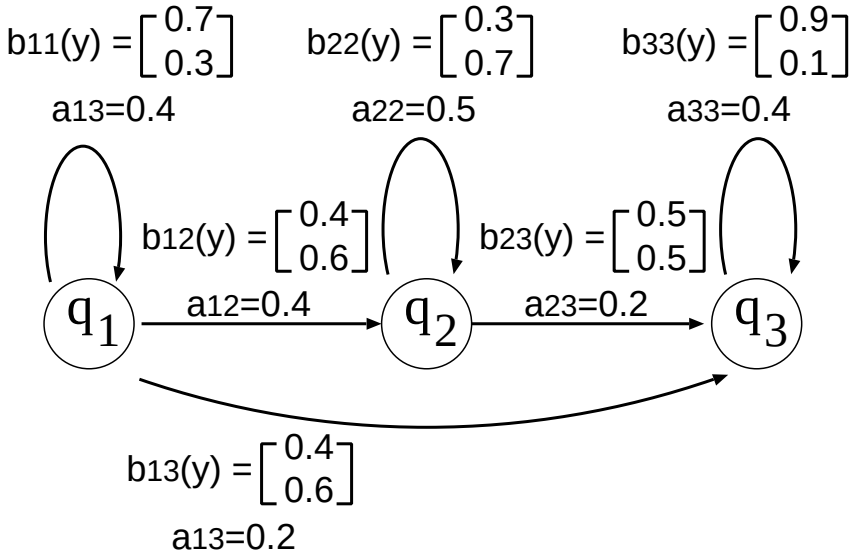

図 2.1: ベイキス型 HMM の例

図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から，$i > j$ なら $a_{ij} = 0$ である．各状態 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，連続地として表す場合 (連続分布，連続 HMM) と，有限個 (K 個) のシンボルの組み合わせで表現する場合 (離散分布モデル，離散 HMM) がある．図 2.1 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．この例では，遷移確率行列は，

$$A = (a_{ij}) = \begin{bmatrix} 0.4 & 0.4 & 0.2 \\ 0.0 & 0.5 & 0.2 \\ 0.0 & 0.0 & 0.4 \end{bmatrix} \tag{2.1}$$

となり，$\pi_1 = 1$，$\pi_i = 0 (i > 1)$，$F = \{q_3\}$ である．

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．

2.2.3　HMM 法の利点と問題点

HMM が音声認識によって有利な点は，以下のような点があげられる [6].

1. 個人差や調音結合，発声法 (強さ，早さ，明瞭さ) など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.
2. 従って，統計理論や情報理論・確率過程論による理論的展開がしやすい.
3. 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.
4. 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, 両者を統合しやすい.
5. 認識時の計算量は比較的少ない.

また，HMM による問題点は，以下のような点があげられる. [6]

1. モデルの設計法が確立されておらず試行錯誤的，ノウハウ的要素が強い.
2. HMM のパラメータ推定に多量の訓練用サンプルを要し計算量も多い.
3. 音声の過渡的パターンの表現力に乏しく，また時系列パターン中の 2 時点におけるパターンの相関が考慮できない.

2.3　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んで，これを認識結果とする．図 2.2 に単語 HMM を用いた認識の方法を示す．

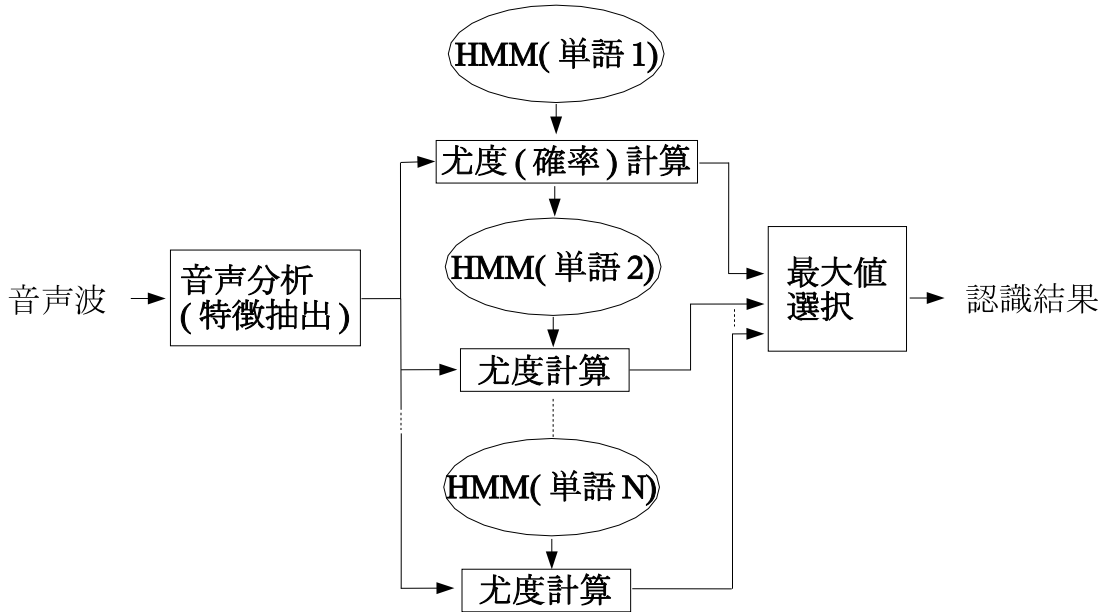

図 2.2: 単語 HMM を用いた単語音声認識の方法

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

$$P(y|M) = \sum_{i_0, i_1, \cdots, i_T} P(y|q, M) \cdot P(q|M) \tag{2.2}$$

と表すことができる．そして一般に $P(y|M)$ 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，

$$\begin{aligned} \alpha(i,t) &= \sum_j \alpha(j, t-1) \cdot a_{ji} \cdot b_{ji}(y_t)(t = 1, 2, \cdots, T) \\ &= \pi_i(t = 0) \end{aligned} \tag{2.3}$$

であるからこれを計算して，最後に

$$P(y|M) = \sum_{i, q_i \in F} \alpha(i,t) \tag{2.4}$$

を求めればよい．

2.3.1　Viterbi アルゴリズム

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

$P(y|M)$ 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って，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て，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，トレリス法と音声認識性能はほとんど変わらないことが，実験的に確認されている [7]．

Viterbi アルゴリズムは，今回の実験では，HMM の初期モデルの作成，認識に使用されている．

2.3.2　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．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．

2.4　音響分析

2.4.1　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 $P(y|w)$ を計算するには, 音声区間の時系列データ y の表現形式を決める必要がある．音声波形そのものを用いたのでは情報量が多すぎ，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ので，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ごとに，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，帯域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 [7]．

2.4.2　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，音の高さに関して，メル (mel) 尺度と呼ばれる対数に近い非線形の特性を示し，低い周波数では，細かく，高い周波数では荒い周波数分解能をもつ．

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 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 24 次のメルスケールフィルタを用いる．

2.4.3　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ル包絡を用いることが多い．この平滑化の手法としてよ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，通常は低次のフォルマント成分が使用される．これは言い換えると，声道特性のみを用いていることになる．本研究において 24 次の FBANK で分析した後，12 次の MFCC を用いる．

第3章　モーラ情報

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．

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 [3)]．

図 3-1 は単一話者のナレータが発話した地名，2800 件のピッチ周波数の平均と分散をあらわしている．この図において，時間軸はモーラ数で正規化した後計算した．図中の■はピッチ周波数の平均値を，縦線の長さは分散を示している．またピッチの周波数の算出には，Xwaves ＋を用いた．図より，ピッチ周波数は単語のモーラ数および単語のモーラ位置が決まればほぼ一定であることがわかる．

上述した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ると期待できる．

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，モーラ情報を使って分類する．具体的には，母音，促音，撥音の前方に単語のモーラ数，後方にモーラ位置情報を付け加えて分類する．分類例を表 3-1 に示す．

音声ラベルが，”kurosu”の場合、単語のモーラ数は 3 なので，母音の前方に 3 をつけ，後方に各々のモーラ位置をつける．2 番目と 6 番目の音素 u は，分類後 3u1，3u3 という音素に置き換え，モーラ位置が異なるため，異なった音素として扱う．

表 3.1: 母音・促音・撥音の分類例

分類前	k	u	r	o	s	u
分類後	k	3u1	r	3o2	s	3u3

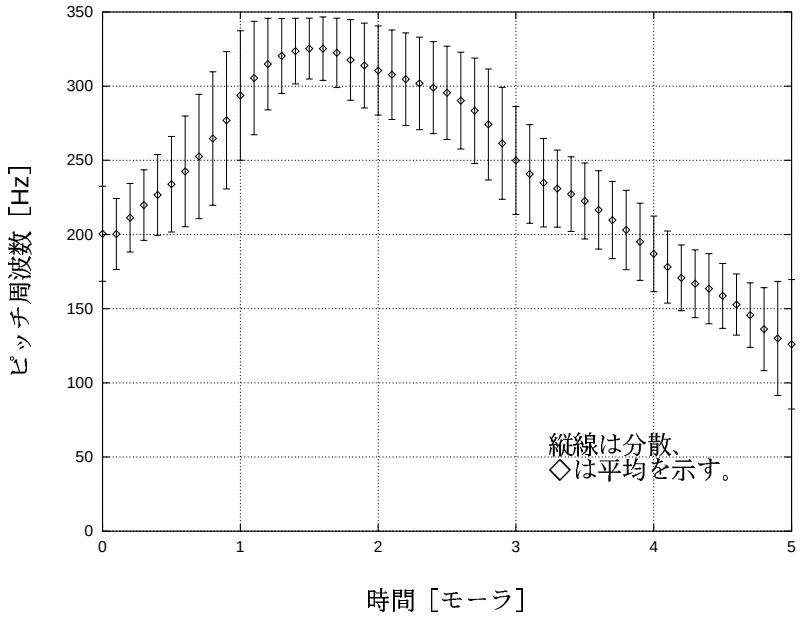

図 3.1: 5 モーラ語 2,800 件のピッチ周波数の平均値と分散

第4章　評価実験

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，現実にはありえないが，同一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 3 名 (男性:mau，mmy，mtk，女性:faf，ftk，fyn) 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．偶数番音声 2620 単語においてランダムに 2 単語抽出し，2 つの音声を重ね合わせて 1310 単語を作成する．重ね合わせは，波形ファイルの 0ms から重ね合わせをする．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．詳細を図 4.1 に示す.

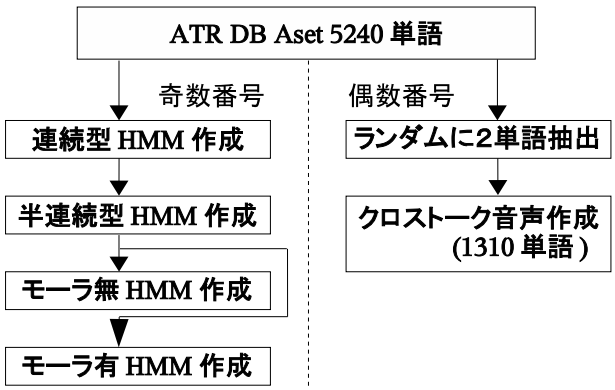

図 4.1: 音素 HMM，クロストーク音声作成手順

4.2　実験条件

本実験では HTK を使用して実験を行う．特徴パラメータに FBANK と MFCC を使用する．また，音素HMMの混合ガウス分布には Full-covariance を使用する．MFCC と FBANK において，モーラ情報を使用した場合と使用しない場合の計 4 種類で実験を行う．詳細を表 4.1 に示す．

表 4.1: 実験条件

基本周波数	16kHz	
分析窓	Hamming 窓	
分析窓長	20ms	
フレーム周期	5ms	
音響モデル	3 ループ 4 状態 半連続分布型	
stream 数	3	
特徴パラメータ	MFCC 12 次 Δ MFCC 12 次 対数パワー 1 次 Δ 対数パワー 1 次	FBANK 24 次 Δ FBANK 24 次 対数パワー 1 次 Δ 対数パワー 1 次
学習ＤＢ	2620 単語	
評価ＤＢ	1310 単語	

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.

本研究では重ね合わせる前の 2 つの音声のどちらか一方が正しく認識できているか評価する.

4.4　実験結果

4.4.1　FBANK と MFCC の比較

FBANK と MFCC の実験結果の比較を表 4.2 に示す.

表 4.2: 実験結果 (MFCC と FBANK の比較)

話者	MFCC	FBANK
mau	28.2%(369/1310)	29.8%(390/1310)
mmy	28.7%(376/1310)	30.4%(398/1310)
mtk	30.4%(398/1310)	29.5%(387/1310)
faf	26.4%(346/1310)	28.4%(372/1310)
ftk	30.1%(394/1310)	31.4%(411/1310)
fyn	28.2%(370/1310)	30.3%(397/1310)
平均	28.7%(2253/7860)	30.0%(2355/7860)

この表から特徴パラメータには MFCC よりも FBANK を用いた場合の方が認識率が高いことが分かる.

4.4.2　モーラ情報を使用した場合の比較

FBANK においてモーラ情報を使用した場合と使用しない場合の比較を表 4.3 に，また，MFCC においてモーラ情報を使用した場合と使用しない場合の比較を表 4.4 に示す．

表 4.3: 実験結果 (FBANK モーラ有無)

話者	モーラ無し	モーラ有り
mau	29.8%(390/1310)	27.0%(354/1310)
mmy	30.4%(398/1310)	29.8%(391/1310)
mtk	29.5%(387/1310)	29.8%(391/1310)
faf	28.4%(372/1310)	28.9%(379/1310)
ftk	31.4%(411/1310)	32.4%(425/1310)
fyn	30.3%(397/1310)	30.6%(401/1310)
平均	30.0%(2355/7860)	29.8%(2341/7860)

表 4.4: 実験結果 (MFCC モーラ有無)

話者	モーラ無し	モーラ有り
mau	28.2%(369/1310)	26.0%(340/1310)
mmy	28.7%(376/1310)	30.1%(394/1310)
mtk	30.4%(398/1310)	30.2%(396/1310)
faf	26.4%(346/1310)	28.5%(373/1310)
ftk	30.1%(394/1310)	32.1%(421/1310)
fyn	28.2%(370/1310)	29.4%(385/1310)
平均	28.7%(2253/7860)	29.4%(2309/7860)

この結果から,FBANK ではモーラ情報を使用することによる全体での認識率改善は得られず，わずかに低下がみられた．MFCC においては逆に改善が見られた．

4.4.3　男性女性の差

FBANK においてモーラ情報使用した場合と使用しなかった場合の男女平均での比較を表 4.5 に示す．また MFCC でモーラ情報使用した場合と使用しなかった場合の男女平均での比較を表 4.6 に示す．

表 4.5: 実験結果 (男女平均 FBANK モーラ有無)

話者	モーラ無し	モーラ有り
男性平均	29.9%(1175/3930)	28.9%(1136/3930)
女性平均	30.0%(1180/3930)	30.7%(1205/3930)

表 4.6: 実験結果 (男女平均 MFCC モーラ有無)

話者	モーラ無し	モーラ有り
男性平均	29.0%(1141/3930)	28.8%(1130/3930)
女性平均	28.2%(1110/3930)	30.0%(1179/3930)

男女別にモーラ情報の有効性を検証する．男性話者では認識率に低下が見られるのに対して，女性話者では認識率に改善が見られる．

第5章　考察

5.1　モーラ情報による改善

話者は ftk でパラメータは FBANK を用いた場合において，モーラ情報を用いることで認識率が改善された単語について一部を表 5.1 に示す．また，モーラ情報を用いることで逆に認識できなくなった単語例を表 5.2 に示す．

表 5.1: 改善された単語 (ftk)

重ね合わせた音声	モーラ無し (誤)	モーラ有り (正)
要する＋大層	よそ	要する
後ろ＋馬	今ごろ	後ろ

表 5.2: 誤認単語 (ftk)

重ね合わせた音声	モーラ無し (正)	モーラ有り (誤)
見合わせる＋平ら	見合わせる	知らせる
快活＋厳密	厳密	管轄

モーラ情報は母音特に連続母音，音素欠落に対して改善が見られた．逆にモーラ情報を用いた場合,「厳密」と認識できていたものが,「快活」と認識できず,「管轄」と認識されるような現象が見られた．

5.2　人手による聴取実験

話者 mau，faf を用いて聴取実験を行い計算機による認識率と比較した．結果を表 5.3 に示す．なお，被験者は男性 1 名である．人手による聴取実験結果と比較すると計算機による認識率は低い結果となった．

なお，人手でも認識できない音は，声の大きさ，テンポが同じような単語の組み合わせであった．逆に異なっていれば両方または片方はなんとか聞き取れた．モーラ情報を使用した FBANK において人手によって聞き取れたが，計算機による認識ができなかった例を表 5.4 に示す．

逆に計算機では認識できたが，人手によって聞き取れなかった例を表 5.5 に示す．

表 5.3: 人手による聴取実験結果

話者	認識率 (%)
mau	73%(961/1310)
faf	69%(907/1310)

表 5.4: 計算機で認識できなかった例 (faf)

重ね合わせた音声	人手	計算機
住まい＋切り	住まい (○)	機械 (×)
残業＋態と	残業 (○)	頑丈 (×)

表 5.5: 人手で聞き取れなかった例 (faf)

重ね合わせた音声	人手	計算機
看板＋七つ	かなばん (×)	看板 (○)
落ち着き＋知れる	おしれつき (×)	落ち着き (○)

第6章　まとめ

本研究では，1 話者が別々の単語を同時に話したと仮定して孤立単語を作成し，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．

特徴パラメータに FBANK を用いることで MFCC より良い結果が得られた．また女性話者においては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低い．

今後の課題としてモーラ情報の扱い方や，学習方法について新たな手法を取り入れるなど改善の余地がある．

謝 辞

最後に，本研究においてご指導を賜りました池原 悟 教授，村上 仁一 助教授，に深く感謝致します．
また，有益な御助言を頂いた徳久 雅人 助手に感謝致します．
最後にに本学大学院生の妹尾 貴宏氏，谷口 勝則氏，石田 隆浩氏に厚く御礼を申し上げると共に，計算機C研究室のみなさまの多大なるご協力に感謝の意を表します．

平成15年2月 中野 晃

参考文献

[1] 妹尾貴宏，村上仁一，池原悟: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)

[2] 谷口勝則，村上仁一，池原悟: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] 水澤紀子，村上仁一，東田正信:音節波形接続による単語音声合成，信学技報，sp99-2(1999)

[4] HTK Ver2.2 reference manual, 1997 Cambridge University

[5] 鹿野清宏他，音声・音情報のディジタル信号処理，昭晃堂

[6] 中川聖一，確率モデルによる音声認識，電子情報通信学会

[7] 古井　貞熙，音声情報処理，森北出版株式会社

[8] Introducing ESPS/waves+ with EnSigTM Entrotic Research Laboratory,Inc.