修 士 論 文

韻律情報を利用した
雑音重畳音声の認識に関する研究

指導教官　下平 博 助教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

川崎 真護

1998 年 2 月 13 日

要 旨

雑音環境下において、音韻情報のみを用いた音声認識では認識精度は低下する。一方、韻律情報の一つであるピッチパターン情報は、それのみでは音声認識を行うことができないが、白色雑音などの影響を受けにくい特徴があり、雑音環境下でも認識の検証に有用であることが考えられる。そこで本研究では、白色雑音環境下において音韻情報と韻律情報の両者を利用することによる認識精度の改善について検討を行う。

目 次

図 目 次

表 目 次

第 1 章

序論

1.1 背景

音声認識の基本的な技術が確立され、近年ではその実用化へと研究課題が移行しつつある。実環境下という点では人混み雑踏などの雑音環境下が想定され、一つの課題となっている。この雑音環境下での音声認識研究としては、周波数重み付け HMM を用いた学習による音声認識 [1]、スペクトルサブトラクション法を用いた雑音除去 [2]、人間の聴覚系のマスキング効果を利用したディジタル蝸牛モデルによる音声パラメータ抽出 [3] などがある。しかし、雑音環境下の音声認識において音声の韻律情報を考慮したものは未だ検討されていない。

韻律情報の一つであるピッチパターンは、音声波形の周期性に影響を及さばない白色雑音下では、ある SNR 以下ではその特徴が乱れないという特徴がある。このことから、白色雑音環境下の音声認識に韻律情報を考慮するのは有効的であると考えられる。雑音が重畳していない音声の認識に関しては、韻律情報を利用した高橋ら [4] の研究が存在する。この研究ではピッチパターン情報を用いることにより、音韻系列やスペクトルパタンの類似した単語 (同音異義語など) の音声認識精度の向上が報告されている。また、相川ら [5] の研究においても、パワー情報が認識率の向上や情報量の圧縮に有効であることが報告されている。

1.2 目的

本研究では、高橋ら [4] の研究をベースとして雑音重畳単語音声の認識における韻律情報の有効性を検討し、実環境下にも耐えうる認識手法を検討する。

具体的には、まず韻律情報を白色雑音環境下における単語音声から抽出する。次に、ここで抽出された韻律情報から得られる韻律認識結果と、音韻情報から得られる音韻認識結果とを本報告によって提案するシステムを用いて組み合わせることにより、雑音重畳単語音声の認識精度向上を実現する。

1.3 本論文の構成

本論文は5章で構成され、第1章は序論である。第2章では、韻律情報を考慮した単語音声認識システムの構成と、新たに認識に用いる韻律情報についての説明およびその抽出法、正規化の手法について述べ、さらに辞書の学習法、韻律尤度の計算手法について述べる。第3章では非雑音環境下、第4章では雑音環境下での韻律情報を用いた単語音声認識実験をそれぞれ行い、その有効性について検討した。第5章では、本論文で提案した韻律尤度計算手法の有効性を検討する目的で、同音異義語の認識実験を行った。第6章は結論である。

第 2 章

韻律情報を用いた単語音声認識システム

2.1 はじめに

本章では、音韻情報と韻律情報との両者を考慮した単語音声認識システムについて説明を行う。また、認識に用いる韻律情報についての説明およびその抽出法、正規化の手法について述べる。さらに、認識に必要な辞書の学習法や、韻律尤度の計算法についても説明を行う。

2.2 システム構成

音声認識の際に韻律情報を考慮して、認識精度の向上を実現させた研究として高橋ら [4] の研究が報告されている。そこで本稿では、この研究をベースとして図 2.1 のシステムを構築した。

図の左側の処理の流れは HMM に基づく音韻認識部であり、右側が韻律認識部である。韻律情報にはピッチパターン情報を用いる。音韻認識部は音韻 HMM と単語辞書を備え、入力単語が kであるという音韻尤度 S_k^{ph} を出力する。一方、韻律認識部はピッチテンプレートと韻律辞書 (各単語の各テンプレートへの出現確率テーブル) を備え、後述 (2.4.3 節) のピッチパターン整合法により韻律尤度 S_k^{pr} を出力する。総合スコア S_kは結合係数αを用いて、

$$S_k = (1.0 - \alpha)S_k^{ph} + \alpha S_k^{pr}$$

によって求める。

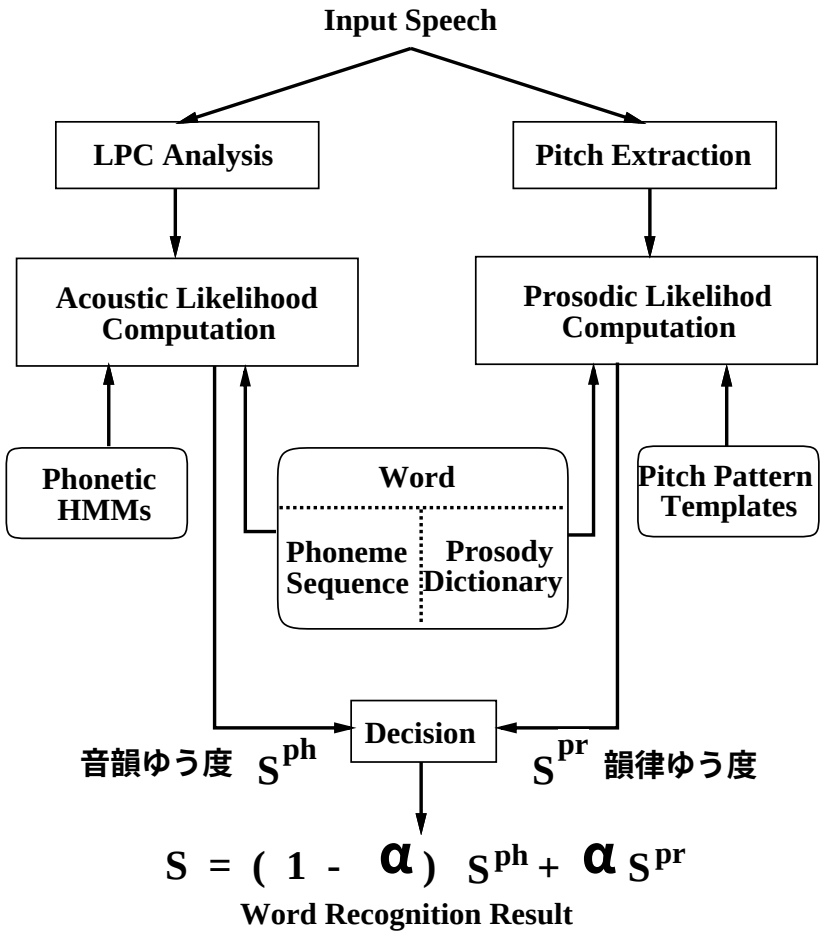

図 2.1: 韻律情報を用いた認識システムの構成

2.3 音韻認識部

音韻認識部では入力単語音声に対して HMM による認識を行い、入力単語が k であるという音韻尤度 S_k^{ph} を出力する。ここで HMM は、単語 HMM ではなく音素 HMM を用いて、大語彙単語認識が可能な単語辞書の音韻系列を参照して単語の音韻尤度を算出した。

2.4 韻律認識部

本システムにおける韻律認識部では、高橋らの研究で用いていた手法を拡張したものを用いた。拡張点は次の 2 つである。

- 先行研究では、単語のピッチパターンがどのアクセント型に属するかを視察で与えていたために、単語数が増加するとそれだけ手間も必要であったが、本システムではクラスタリング分類法を用いてテンプレートを作成することにより視察の手間を省き、単語数の増加にも対応できるようにした。

- 先行研究では、ピッチテンプレートの学習及びパターン整合の際に、時間伸縮によるピッチパターン時間長の正規化を行っていたために、アクセント核の位置が移動するという問題が生じていたが、本システムでは、ピッチパターンに対し F_0生成モデル (藤崎モデル)[9] を用いてアクセント情報とフレーズ情報を推定し、その情報を基にピッチパターンを一定の時間長に再構成することで、この問題を解決した。

2.4.1 韻律特徴の抽出

韻律情報の一つであるピッチは、声の高さを表すもので、基本周波数、F_0周波数とも呼ばれる。多言語音声の分類 [6] や連続音声のアクセント句境界検出 [7] においては、ピッチの時系列であるピッチパターンが有効的な特徴量であることが報告されている。

本論文では、ピッチパターン抽出にラグ窓法 [8] を用いる。ラグ窓法はまずフレームごとの音声波形の標本相関関数 $v(\tau)$ に対してフーリエ変換を行い、標本スペクトル $I_n(\lambda)$ を計算する。その一方で、フレームごとの音声波形の標本相関関数 $v(\tau)$ に対してラグ窓ω_τを乗じてからフーリエ変換を行い、平滑化された標本スペクトル$\tilde{I}_n(\lambda)$ を得る。ここで、$I_n(\lambda)$ を$\tilde{I}_n(\lambda)$ で割るとピッチ構造のみが抽出され、これに逆フーリエ変換を行うと基本周期に相当する時間に著しいピークが得られ、その位置からピッチが求められる。ラグ窓ω_τは以下の式によって表される (L は窓長)。図 2.2 に概略図を示す。

$$\omega_\tau = \frac{(L!)^2}{(L+\tau)!(L-\tau)!}$$

なお、ピッチをフレーム独立に計算する場合隣り合うピッチ同士の関連性がなくなり、倍ピッチ誤り、半ピッチ誤りなどのピッチ抽出誤りを回避できないという問題がある。これに対して、F_0信頼度を用いてピッチ抽出誤りを除去する。ラグ窓法では $u(\tau)$ の値が 0〜1 に正規化され、音声信号の繰り返し波形が明瞭に読み取れる区間では F_0の探索区間内での $u(\tau)$ の最大値が非常に高いという特徴がある。このため $\max u(\tau)$ の値を F_0信頼度として利用できる。

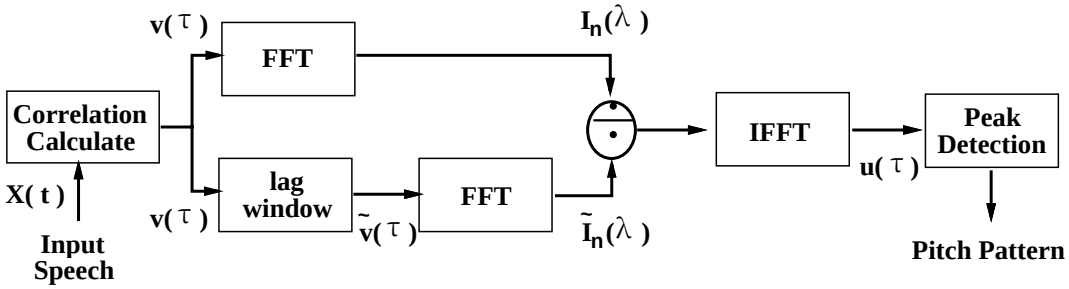

図 2.2: ラグ窓法を用いたピッチパターン抽出法

2.4.2 学習

全単語 (複数回発声) をピッチパターンの形によってクラスタ分類することにより、ピッチテンプレートと韻律辞書を作成する。ピッチテンプレートの学習には、まず、自動抽出ピッチパターンに対して F_0生成モデル [9] のパラメータ推定を行う。

F_0生成モデルとは、ピッチパターンを句頭から句末にかけて緩やかに下降するフレーズ成分と、局所的に起伏するアクセント成分との重畳形としてとらえるモデルである。フレーズ指令が I個、アクセント指令が J個ある場合、対数 F_0周波数は時刻 t の関数として次の式によって与えられる。

$$\begin{aligned}\ln F_0(t) = \ln F_b &+ \sum_{i=1}^{I} A_{p_i} G_{p_i}(t - T_{0i}) \\ &+ \sum_{j=1}^{J} A_{a_j} \{G_a(t - T_{1j}) - G_a(t - T_{2j})\}\end{aligned}$$

F_bは話者に依存した最低基本周波数、A_{p_i},A_{a_j}は i 番目のフレーズ指令および j番目のアクセント指令の大きさ、T_{0i}は i 番目のフレーズ指令の発声時刻、T_{1j},T_{2j}は j番目のアクセント指令の開始時刻および終了時刻である。また、$G_{p_i}(t)$ はフレーズ制御機構のインパルス応答関数で、$G_{a_j}(t)$ はアクセント制御機構のステップ応答関数である。α_i,β_jをそれぞれ固有角周波数とすると、$G_{p_i}(t)$,$G_{a_j}(t)$ はそれぞれ以下の式によって表される。θ_jは $G_{a_j}(t)$ の天井値 (約 0.9) である。

$$G_{p_i}(t) = \begin{cases} \alpha_i^2 t e^{-\alpha_i t}, & (t \geq 0) \\ 0, & (\text{otherwise}) \end{cases}$$

$$G_{a_j}(t) = \begin{cases} \min[1 - (1 + \beta_j t)e^{-\beta_j t}, \theta_j], & (t \geq 0) \\ 0, & (\text{otherwise}) \end{cases}$$

ただし、ここで示した F_0生成モデルは文音声のモデルである。本システムは単語音声認識システムであるので、フレーズ指令 $I = 1$ 個、アクセント指令 $J = 1$ 個としてこのモデルを用い、入力単語のピッチパターン $P = \{P_1, P_2, \cdots, P_M\}$ のフレーズ成分とアクセント成分を AbS 分析により推定する。ここで分析に使用するピッチは F_0信頼度がある閾値 (テンプレート、韻律辞書作成時は 0.5) 以上のものを使用する。推定されたフレーズ成分とアクセント成分を基にピッチパターンを再構成することで時間長の正規化$\hat{P} = \{\hat{P}_1, \hat{P}_2, \cdots, \hat{P}_L\}$ の処理を行う。これにより、アクセント核の位置を変えずに全学習パターンの長さがそろえられる。さらにこの際、話者に依存した最低基本周波数 F_bを引くことで、

$$\bar{P}_i = \hat{P}_i - F_b$$

高さの正規化$\bar{P} = \{\bar{P}_1, \bar{P}_2, \cdots, \bar{P}_L\}$ も行う。この手法により再構成したピッチパターンを学習資料として用いる。

次に、LBG 法 [10] によるクラスタリングを行ない、複数の話者の発声による全学習資料を J個のクラスに分類する。この時のクラスタ重心ベクトルをピッチテンプレートとする。図 2.3 にここまでの概略図を示す。

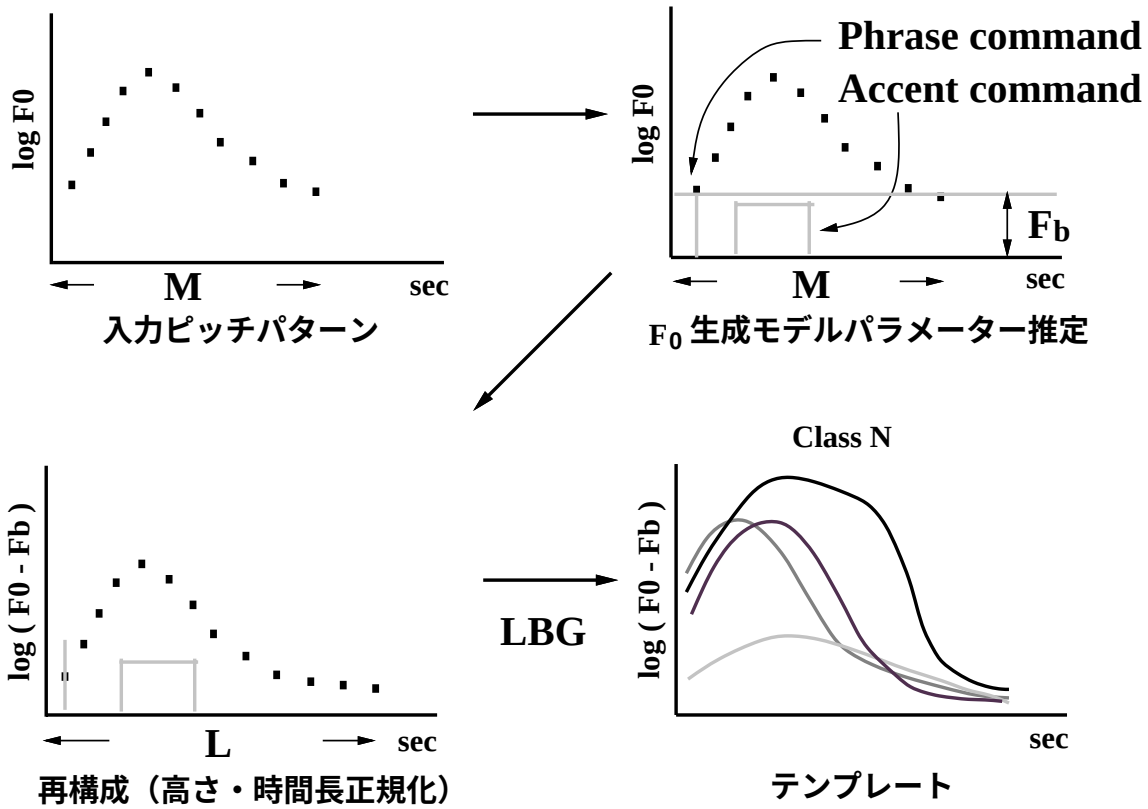

図 2.3: ピッチテンプレートの学習

韻律辞書には、複数の話者によって発声された同一単語の各クラスへの出現確率を記載する。韻律辞書は図 2.4 のようなテーブルで記述できる。例えば、10 人の話者が単語' ai ' を発音した時に、そのうちの 5 人の発声した単語' ai ' のピッチパターンがテンプレート 1 に分類された場合、$P_{11} = 0.50$ となる。したがって$\sum_{j=1}^{J} p_{ij} = 1.0$ である

Word \ Tempalte	1	2	···	N
1 : ai	p_{11}	p_{12}	···	p_{1N}
2 : aida	p_{21}	p_{22}	···	p_{2N}
⋮				
K : ···	p_{K1}	p_{K2}	···	p_{KN}

図 2.4: 韻率辞書:各単語の各クラスへの出現確率

2.4.3 認識

まず、入力単語に対して、ラグ窓法を用いてピッチパターン $P = \{P_1, P_2, \cdots, P_M\}$ を抽出する。ただし、テンプレート作成時のような F_0生成モデルパラメーター推定は行わずに、ピッチ信頼度の高い点のみ (認識実験時は 0.1 以上) をテンプレートとのマッチングに用いる。一方、テンプレートに対しては F_0生成モデルパラメーター推定を行ない、入力単語の時間長 Mにあわせてテンプレートの再構成を行なう$\bar{P} = \{\bar{P_1}, \bar{P_2}, \cdots, \bar{P_M}\}$。これにより、入力単語との時間長がそろえられるので、マッチングの際のパターン間の距離を次式のベクトルの二乗距離で定義できる。

$$D_F(P_1, \bar{P_1}) = \sum_{l=1}^{M} (P_l - \bar{P_l})^2$$

韻律尤度は次の式によって計算される。入力単語のピッチパターンとテンプレート n との距離を d_nとし、単語 kがクラス n に出現する確率を p_{kn}とする。この時、入力単語が単語 kであると仮定した場合の韻律尤度は次式で定義する。

$$S_k^{pr} = -\sum_{n=1}^{N} p_{kn} d_n$$

第 3 章

非雑音環境下における単語音声認識実験

3.1 はじめに

本章では、韻律情報を用いた非雑音環境下における単語音声認識の実験を行い、その認識精度への有効性について検討する。

3.2 実験条件

音声資料

実験には ATR の多数話者データ (男性 20 話者:m101〜m120) による最重要単語 520 語を用いた。サンプリング周波数は 20kHz で、非雑音環境下における収録音声である。この内、男性 15 話者 (m101〜m115) のデータを学習データとして用い、残りの男性 5 話者 (m116〜m120) のデータを評価用データとして用いた。また、学習データ内の 1 話者 (m101) のデータを結合係数 α 調整用に用いた。

ピッチパターン分析条件

FFT 分析	1024 ポイント
分析シフト (1 frame)	200 ポイント (10 msec)
ピッチ探索範囲	50Hz〜500Hz
ラグ窓幅	200 ポイント

表 3.1: ピッチパターン分析条件

HMM 分析条件

状態数	3
音素モデル数	26
特徴量	メルケプストラム (分析次元 24 次元)
ハミング窓	20msec

表 3.2: HMM 分析条件

ピッチテンプレート作成

学習用データから抽出されたピッチパターンに対して長さや高さの正規化を行なったものを用いて、クラスタ数 $N = 8$ でクラスタリングを行った。これによって得られたクラスタ代表パターンであるピッチテンプレートを図 3.1 に示す。この図を見ると、それぞれの代表パターンのアクセントピーク位置に差があり、おおよそアクセント型を良好に近似していることが見られる。1 モーラ平均 150msec であるので、例えば、テンプレート 3 は 1 型のアクセント型に相当すると考えられる。

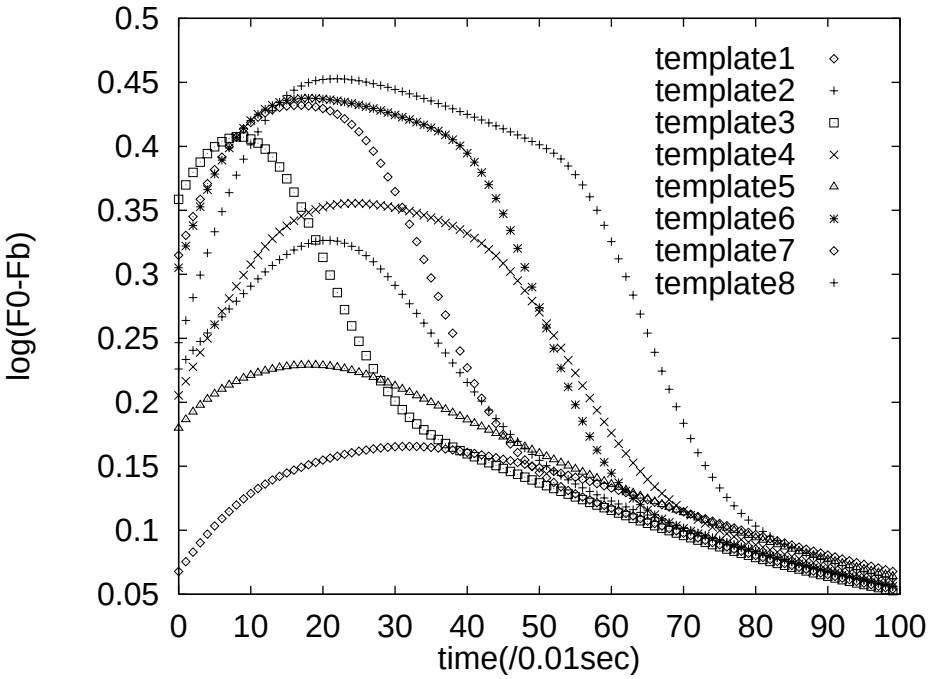

図 3.1: ピッチテンプレート

3.3 音韻尤度のみ (従来法) による認識実験

学習データ 5 話者 (m111〜m115)、評価データ 5 話者 (m116〜m120) に対して、HMM のみによる認識実験を行った際の誤認識単語数と単語認識率を表 3.3 に示す。この表をみると、学習データ、評価データともに単語認識率にあまり差がでないことが確認された。

	学習データ			評価データ	
話者	誤認識単語数	単語認識率 (%)	話者	誤認識単語数	単語認識率 (%)
m111	38	92.7	m116	67	87.1
m112	51	90.2	m117	57	89.0
m113	85	83.7	m118	57	89.0
m114	52	90.0	m119	55	89.4
m115	81	84.4	m120	36	93.1
平均	61.4	88.2	平均	54.4	89.5

表 3.3: 音韻尤度のみによる単語音声認識実験

3.4 音韻尤度と韻律尤度の総合評価による認識実験

3.4.1 特定話者による結合係数の決定と不特定話者実験

学習データ内の m101 によって発声された単語音声に対して、韻律情報を考慮した認識実験を行い、結合係数 α の変化による認識率の変化を調べた。その結果を図 3.2 に示す。この図を見ると、 $\alpha = 0.00$ の時 (HMM のみによる認識) に誤認識単語数が 51 あったのに対し、 $\alpha = 0.87$ の時には誤認識単語数が 47 まで減少することが分かった。このことから、 $\alpha = 0.87$ が最も誤認識単語数が少なくなる最適な結合係数であることが分った。

この時変化のあった単語を表 3.4 に示した。この表の中で改善された単語というのは、韻律情報によって誤認識から正解へと変化した単語であり、改悪された単語というのはその反対に正解から誤認識へと変化してしまった単語である。この表を見ると、改善された単語は 10 単語あるが、改悪されてしまった単語も 6 単語存在するため、結果的には 4 単語の改善となっていることが分った。

改善された単語 (koutai:交替 → kyoudai:兄弟) のピッチパターンを図 3.3 に示した。この図を見ると、両単語のピッチパターンの形が大きく異なっていることが分る。この傾向は、他の改善された単語においても確認された。このことから正解単語と誤認識単語とのピッチパターン形状が大きく異なり、両単語の韻律尤度 S^{pr} に大きな差が生じた場合に誤認識から正解へと改善されることが確認できた。

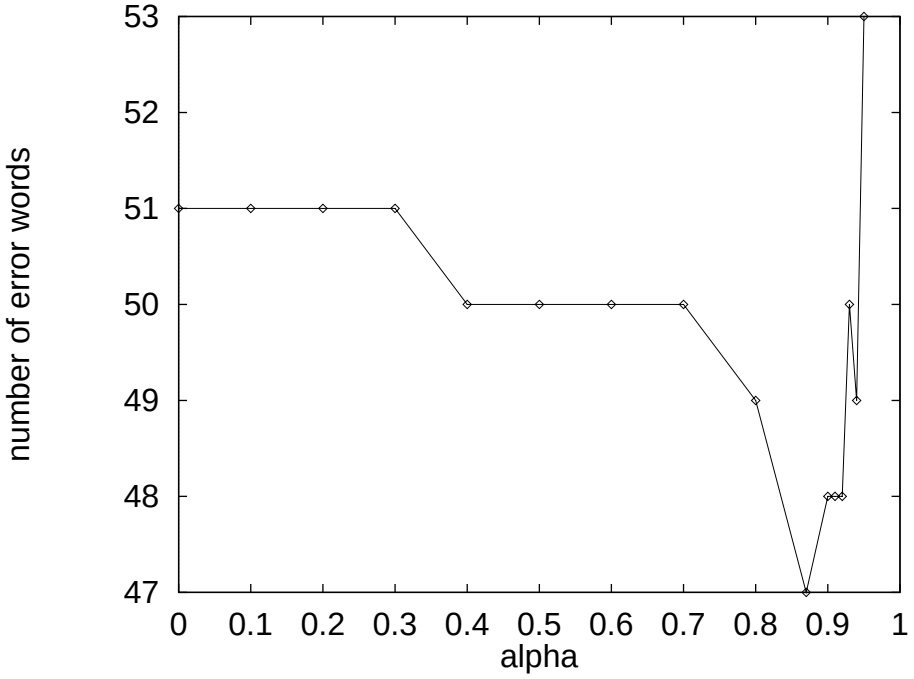

図 3.2: 特定話者による結合係数 α の決定

改善された単語			改悪された単語		
誤認識	→	正解	正解	→	誤認識
kaigi:会議	→	ai:愛	undou:運動	→	bangou:番号
shiru:知る	→	kieru:消える	shitsumon:質問	→	senmon:専門
koutai:交替	→	kyoudai:兄弟	jissai:実際	→	chiisai:小さい
otto:夫	→	kooto:コート	tomeru:止める	→	homeru:誉める
sakai:境	→	shakai:社会	buji:無事	→	michi:道
shougyou:商業	→	shinyou:親友	muchuu:夢中	→	nichiyou:日曜
kikan:期間	→	jikan:時間			
kayou:通う	→	tayoru:頼る			
youso:要素	→	tekitou:適当			
koutai:交替	→	botan:ボタン			

表 3.4: $\alpha = 0.87$ の時に変化のあった単語

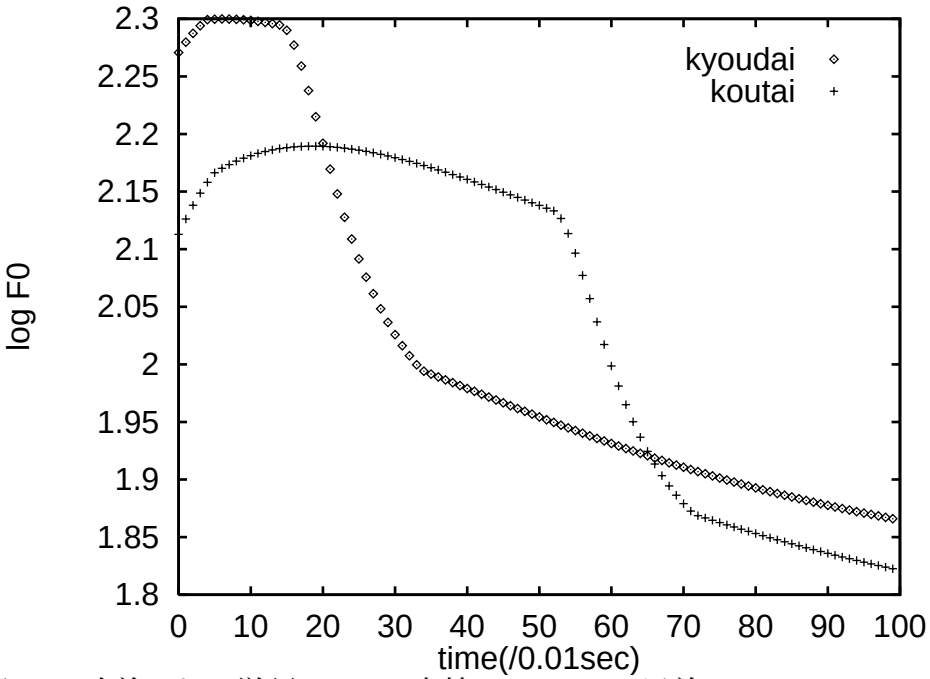

図 3.3: 改善された単語 (koutai:交替 → kyoudai:兄弟) のピッチパターン

次に、上の実験から得られた最適な結合係数を利用して、評価データ 5 話者 (m116〜m120)、および α について open な学習データ 5 話者 (m111〜m115) についても韻律情報を考慮した認識実験を行った。その認識結果を表 3.5 に示す。表には HMM のみの認識からの改善数、単語認識率をそれぞれの話者において示した。改善数の+ は改善、−は改悪されたことを示している。この表を見ると、誤認識単語数が減少している話者と増加している話者が存在することが分るが、平均的には若干の単語認識率の改善結果が得られることが確認された。

	学習データ			評価データ	
話者	改善数	認識率 (%)	話者	改善数	認識率 (%)
m111	+4	93.5	m116	+3	87.7
m112	+4	91.0	m117	+3	89.6
m113	−4	82.9	m118	+5	90.0
m114	+2	90.4	m119	+4	90.2
m115	−1	84.2	m120	0	93.1
平均	+1.0	88.4	平均	+3.0	90.1

表 3.5: 特定話者によって決定した結合係数 $\alpha = 0.87$ の不特定話者への適用

さらに、学習データ 5 話者 (m111〜m115)、評価データ 5 話者 (m116〜m120) の結合係数 α の変化による平均誤認識単語数の変化を調べた。その結果をそれぞれ図 3.4 、図 3.5 に示す。これらの結果を見ると、$\alpha = 0.87$ という値が学習データ、評価データのどちらにおいても最も平均誤認識単語数を改善する値であることが確認された。そこで結合係数 α の話者依存性について 3.4.2 節で調べる。

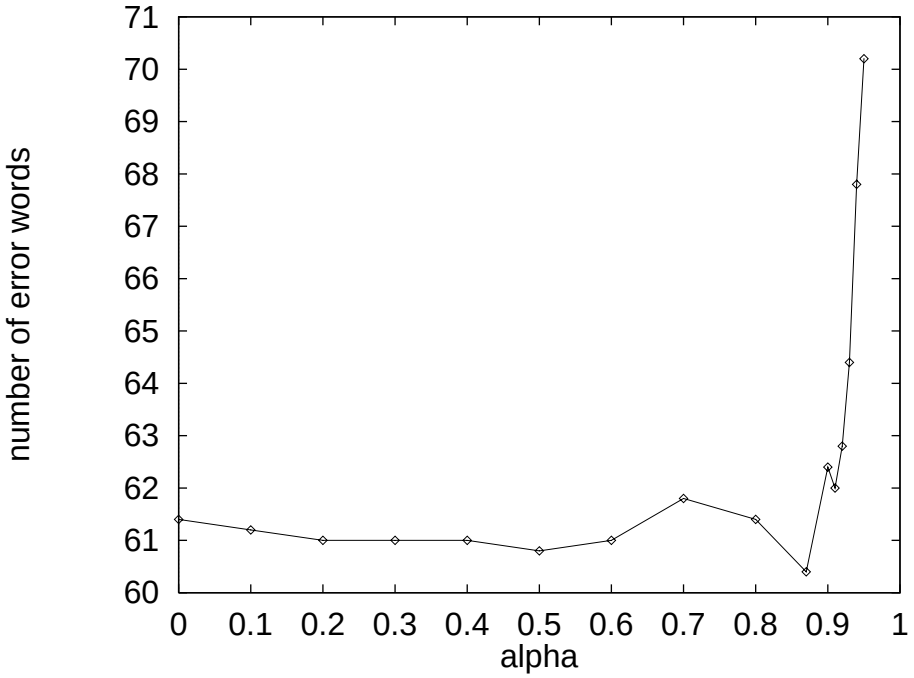

図 3.4: 評価データ 5 話者 (m111〜m115) の平均誤認識単語数の変化

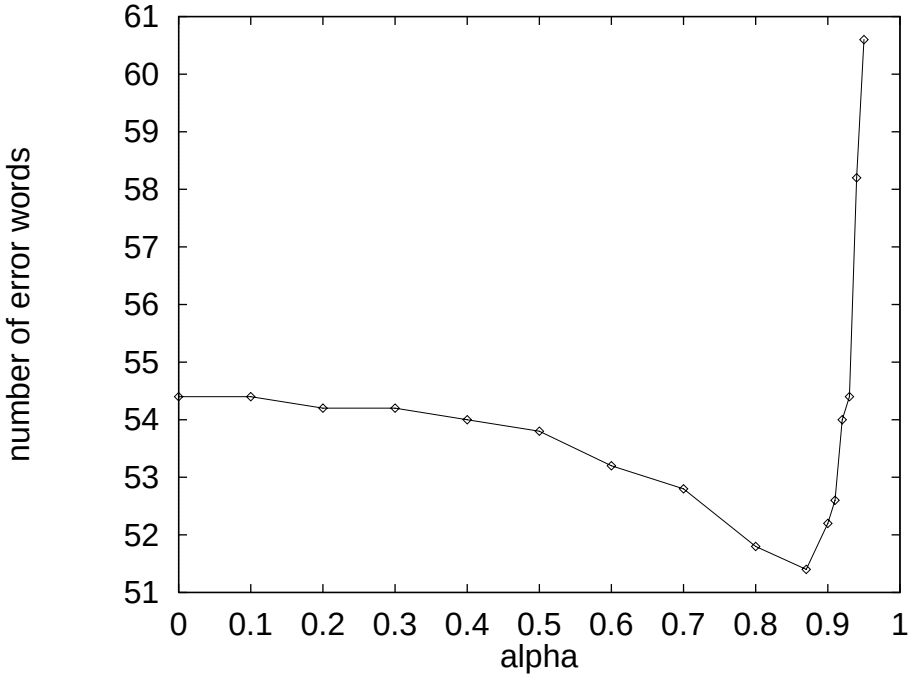

図 3.5: 評価データ 5 話者 (m116〜m120) の平均誤認識単語数の変化

3.4.2 結合係数の話者依存性に関する検討

3.4.1 節の実験では、m101 に対して最適であった $\alpha = 0.87$ という結合係数を他の話者にも適用したが、誤認識単語数が改善される話者と改悪される話者が存在してしまった。そこで、話者ごとに最も誤認識単語数が減少する最適な結合係数 α の値を調べた。その結果を表 3.6 に示す。表には最適な結合係数 α の値、HMM のみの認識からの改善数、単語認識率をそれぞれの話者において示した。改善数の+は改善、−は改悪されたことを示している。この表を見ると、各話者とも最適な結合係数 α の値にばらつきはあるものの、ほぼ $\alpha = 0.87$ に近い値にあることが確認できた。

	学習データ				評価データ		
話者	α	改善数	認識率 (%)	話者	α	改善数	認識率 (%)
m111	0.91	+5	93.7	m116	0.80	+5	88.1
m112	0.87	+4	91.0	m117	0.90	+4	89.8
m113	0.50	+1	83.8	m118	0.87	+5	90.0
m114	0.87	+2	90.4	m119	0.87	+4	90.2
m115	0.92	0	84.4	m120	0.87	0	93.1
平均	0.81	+2.4	88.7	平均	0.86	+3.6	90.2

表 3.6: 結合係数 α の話者依存性に関する検討

3.4.3 改善率の上限値に関する考察

これまでの実験では、各話者の全単語において結合係数 α を固定して韻律情報を考慮した場合の認識実験を行っていた。ここではピッチパターンによって改善できる単語の上限数を調べる実験として、結合係数 α を入力毎に $0 < \alpha < 1$ の範囲で振らした場合に誤認識から正解へと改善する可能性がある単語数を調べた。その結果を表 3.7 に示す。表には HMM のみの認識からの改善数、単語認識率をそれぞれの話者において示した。改善数の+は改善、−は改悪されたことを示している。この表を見ると、どの話者においても HMM のみの認識時おける誤認識単語数の約 50%強を韻律情報によって改善できることが分る。これは何らかの基準で結合係数 α を入力毎に自動制御することができれば、韻律情報が単語音声認識に有効な特徴量となることを示している。

	学習データ			評価データ	
話者	改善数	認識率 (%)	話者	改善数	認識率 (%)
m111	+30	98.5	m116	+35	93.8
m112	+23	94.6	m117	+26	94.0
m113	+31	90.0	m118	+32	95.2
m114	+25	94.8	m119	+31	95.4
m115	+34	91.0	m120	+21	97.1
平均	+28.6	93.8	平均	+29	95.1

表 3.7: 韻律情報による改善率の上限値

3.5 音韻尤度の上位候補に韻律尤度を加えた認識実験

これまでの実験ではすべての HMM の認識結果に対して韻律情報を考慮していたために、3.4.1 節で示したような HMM では正解であった単語が誤認識へと改悪されてしまうという問題が存在した。そこで、この問題を解決するために N-best 法を用い、HMM の認識結果の内上位 2 位までの結果にのみ韻律情報を考慮する認識実験を行った。図 3.6 に m101 によって発声された単語音声に、2-best を適用した場合と適用しなかった場合の結合係数 α の変化に対する誤認識単語数の変化を示した。この図を見ると、$\alpha = 0.90$ の時に誤認識単語数が 46 まで減少していることが分り、2-best を適用しなかった場合と比べて若干の単語認識率の向上が確認された。

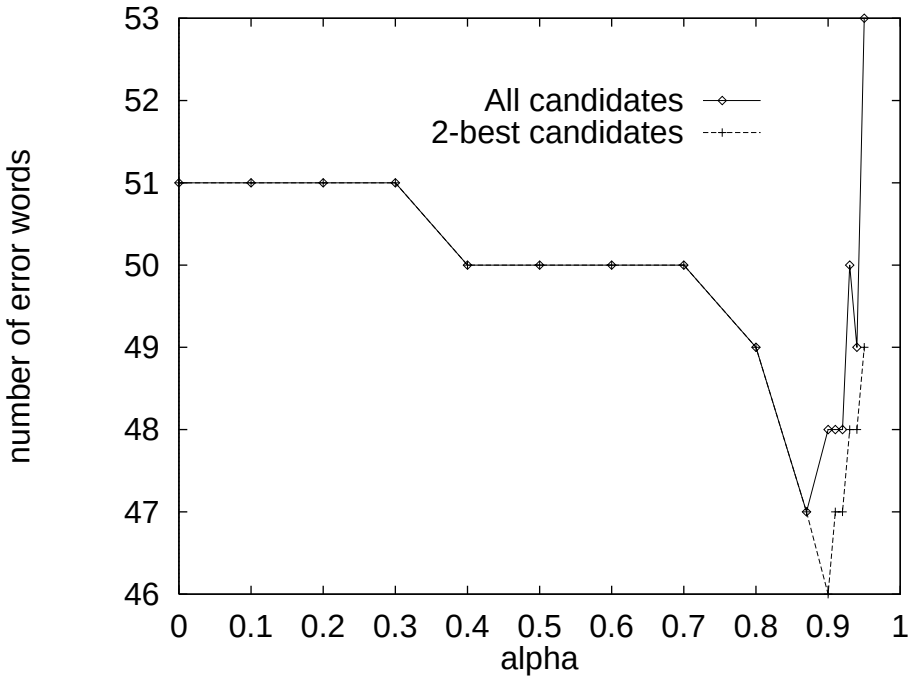

図 3.6: 音韻尤度で候補を絞った認識実験 (2-best)

さらに、上記の手法に加え、HMM による音韻尤度の値に上限と下限の範囲を設定して、その範囲内に 1 位の HMM 認識結果が存在する場合にのみ韻律情報を考慮するという閾値 (Limits) を追加した認識実験を行った。つまり、

- 下限値<1 位<上限値

の場合にのみ韻律情報を考慮し、これ以外の場合

- 1 位<下限値<上限値
- 下限値<上限値<1 位

には韻律情報は考慮しないで HMM のみによる判定を行った。その結果を図 3.7 に示す。HMM の上限と下限の値は予備実験によって上限 $= -61.16$, 下限 $= -68.30$ と決定した。図には全候補に韻律情報を適用した場合と 2-best を適用した場合、2-best+Limits の閾値を適用した場合の結合係数 α の変化に対する誤認識単語数の変化を示した。この図を見ると、$\alpha = 0.94$ の時に誤認識単語数が 42 まで減少していることが分り、この 2 つの条件によって大きな単語認識率の改善が得られることが確認された。

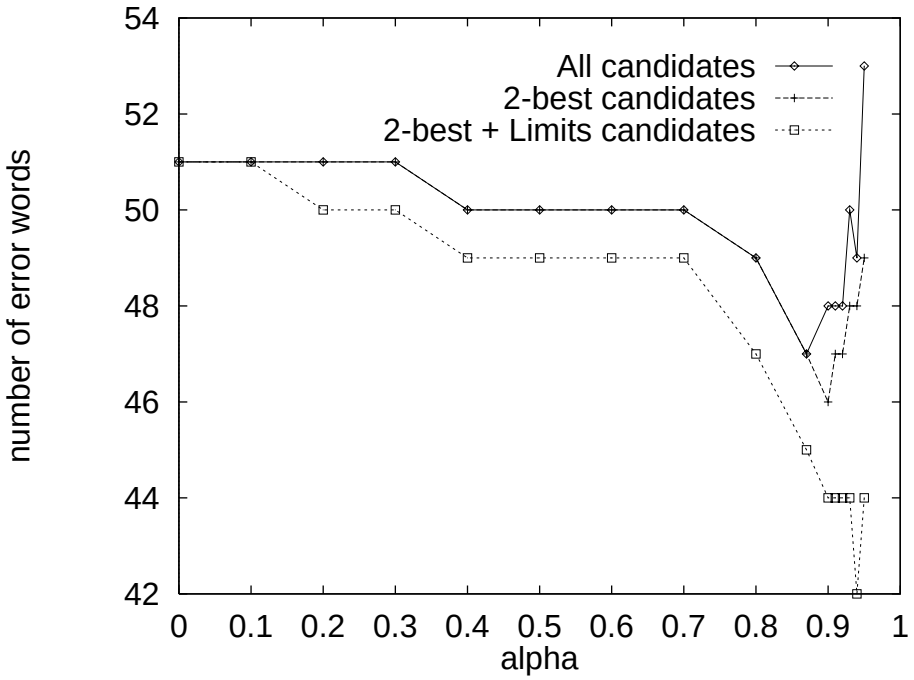

図 3.7: 音韻尤度で候補を絞った認識実験 (N-best + Limits)

3.6 まとめ

本章の実験により、韻律情報が非雑音環境下における単語音声認識の認識率向上に有効性のあることを示した。また、特定話者 1 名によって求めた最適な結合係数 α の値が不特定話者の平均認識率向上に貢献することを確認した。さらに、特定 1 話者の認識において、HMM の認識結果に閾値 (N-best + Limits) を考慮にいれることによって、さらなる認識率の改善が得られることを示した。

今回の実験では、正解単語と誤認識単語のピッチパターンの形が異なっている場合に認識改善結果が得られることを確認したが、今後はアクセント型との関係についても検討して行く予定である。また、認識率を向上させる HMM の閾値 (N-best + Limits) を 1 話者に対してのみ検討したが、今後は各話者すべてに対し HMM の閾値について検討する予定である。さらに、入力毎に結合係数 α を可変とした場合の制御法などについても検討を行う予定である。

第 4 章

雑音環境下における韻律情報を用いた単語音声認識実験

4.1 はじめに

本章では、韻律情報を用いた雑音環境下における単語音声認識の実験を行い、その認識率への有効性について検討する。

4.2 実験条件

音声資料

実験に用いた音声資料は第 4.2.1 節と同様である。男性 20 話者 (m101〜m120) のデータの内、雑音の重畳していない男性 15 話者 (m101〜m115) のデータを学習データとして用い、雑音を重畳した残りの男性 5 話者 (m116〜m120) のデータを評価用データとして用いた。雑音としては、SNR が 45dB,25dB,5dB となる白色雑音を評価用データの単語音声に重畳した。話者 m120 によって発声された単語' gaikoku ' の SNR $= \infty$ の波形を図 4.1 に、SNR = 5dB 波形を図 4.2 に示した。

ピッチパターン分析条件、HMM 分析条件、ピッチテンプレート作成は 3.2 節と同様である。

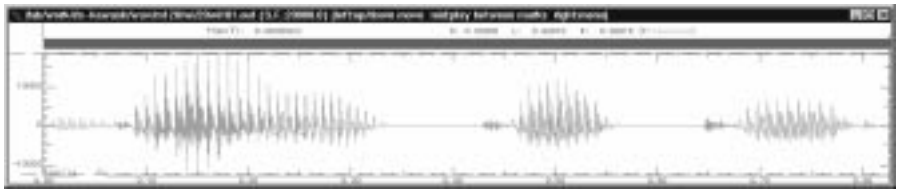

図 4.1: 単語’ gaikoku ’(SNR = ∞)

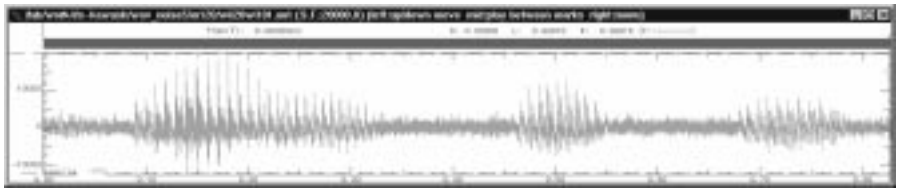

図 4.2: 単語’ gaikoku ’(SNR = 5dB)

雑音重畳単語音声からのピッチパターン抽出

話者 m120 によって発声された単語’ ai ’ の各 SNR におけるピッチパターン波形を図 4.3 に示す。この図を見ると、どの SNR においてもピッチパターンの抽出精度にほとんど差がなく点が重なっており、白色雑音の影響を受けていないことが分る。このことからピッチパターン情報が耐雑音性にすぐれた特徴量であることが確認できる。

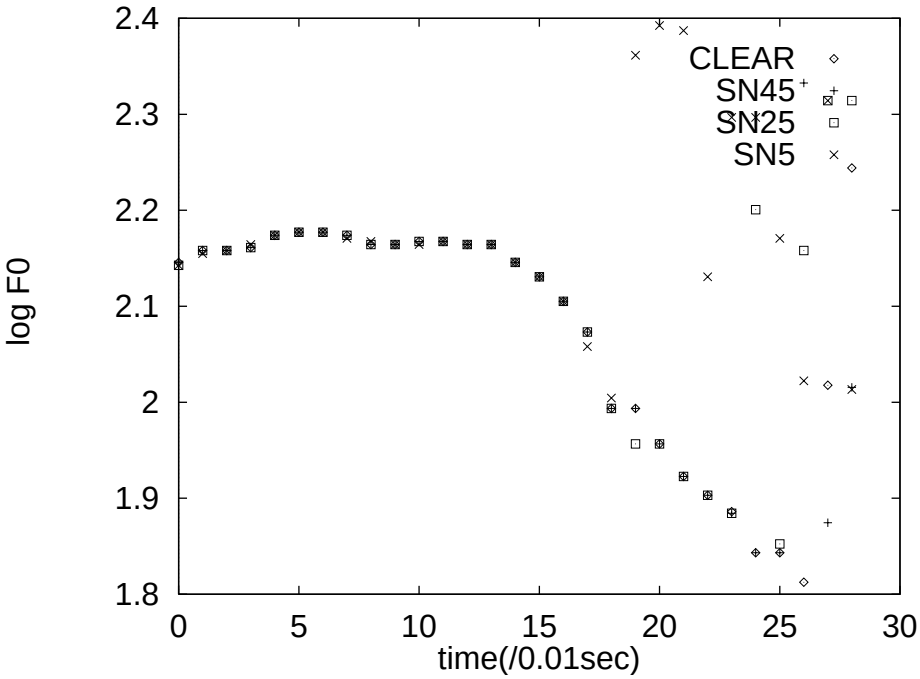

図 4.3: 各 SNR における単語’ ai ’ のピッチパターン波形

4.3 音韻尤度のみ (従来法) による認識実験

各 SNR における雑音重畳単語音声 (評価データ 5 話者 (m116〜m120)) を HMM のみを用いて認識した時の誤認識単語数と単語認識率を、それぞれ表 4.1 に示した。この表を見ると、雑音が大きくなるほどに誤認識単語数が多くなり、単語認識率も低下していることが確認できた。

また、特定話者において SNR45dB 時の誤認識単語を調べてみると、SNR∞ 時の誤認識単語と比較して表 4.2 の単語が誤認識単語として増加していることが確認できた。

	SNR∞		SNR45dB		SNR25dB		SNR5dB	
話者	誤認識単語数	認識率 (%)	誤認識単語数	認識率 (%)	誤認識単語数	認識率 (%)	誤認識単語数	認識率 (%)
m116	67	87.1	67	87.1	134	74.2	439	15.6
m117	57	89.0	59	88.7	96	81.5	403	22.5
m118	57	89.0	70	86.5	167	67.9	493	5.2
m119	55	89.4	69	86.7	155	70.2	481	7.5
m120	36	93.1	47	91.0	104	80.0	446	14.2
平均	54.4	89.5	62.4	88.0	131.2	74.8	452.4	13.0

表 4.1: HMM のみによる雑音重畳単語認識

増加した誤認識単語		
正解	→	誤認識
○ kitai:期待	→	sukkari:すっかり
○ dairi:代理	→	kaigi:会議
○ bamen:場面	→	ageru:上げる
hidari:左	→	kitai:期待
buzi:無事	→	kushin:苦心

表 4.2: SNR45dB 時に SNR∞ の誤認識単語から増加した誤認識単語

4.4 音韻尤度と韻律尤度の総合評価による認識実験

4.4.1 不特定話者の平均改善率による結合係数の決定

本章の実験では、4.4.3 節と同様に、評価データ 5 話者 (m116〜m120) の平均誤認識単語数を用いて、誤認識単語数が最も減少する最適な結合係数 α の値を調べた。各 SNR における誤認識単語数変化の結果をそれぞれ図 4.4 , 図 4.5 , 図 4.6 , 図 4.7 に示す。これらの図を見ると、SNR∞ は $\alpha = 0.87$ 時に 51.4、SNR45dB は $\alpha = 0.90$ 時に 56.6、SNR25dB は $\alpha = 0.90$ 時に 123、SNR5dB は $\alpha = 0.91$ 時に 443 まで平均誤認識単語数が減少していることが分る。さらに、どの SNR においても最適な結合係数 α の値がほぼ $\alpha = 0.90$ で一致していることが確認された。

各 SNR における最適な結合係数 α での話者ごとの単語認識率の変化を表 4.3 に示した。表には HMM のみの認識からの改善数、単語認識率を示している。改善数の+は改善、–は改悪されたことを示している。

また、特定話者において SNR45dB 時に変化のあった単語を調べてみると、表 4.2 に示した誤認識単語のうち〇で示した 3 単語が改善されることが確認された。この時改善された単語 (bamen:場面 → ageru:上げる) のピッチパターンを図 4.8 に示す。この図をみると 3.4.1 節と同様に、正解単語と誤認識単語とのピッチパターン形状が大きく異なっている場合に誤認識から正解へと改善されることが確認できる。

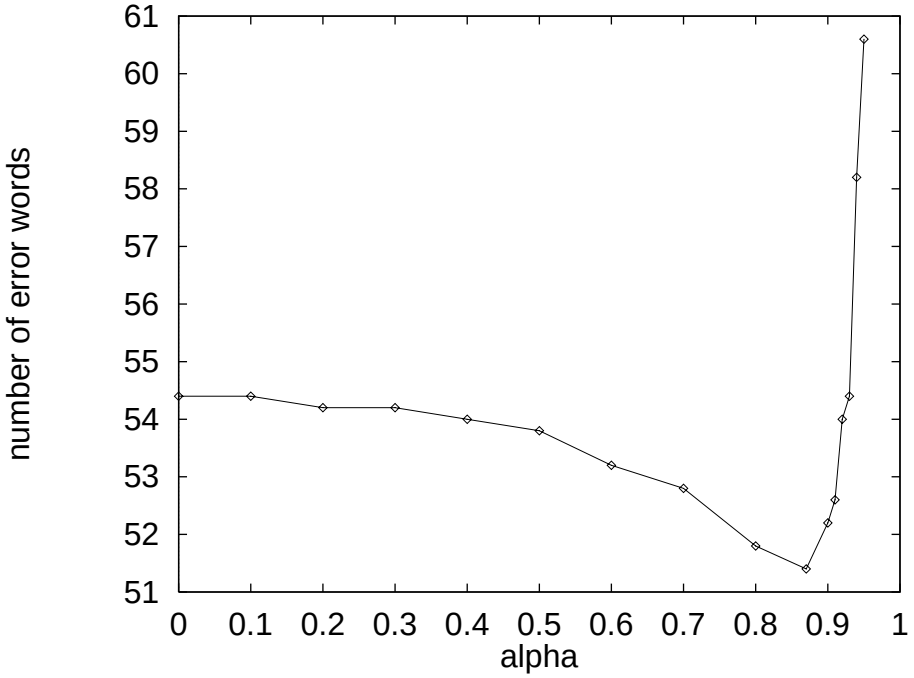

図 4.4: 評価データ 5 話者 (m116〜m120) の平均誤認識単語数の変化による結合係数 α の決定 (SNR∞)

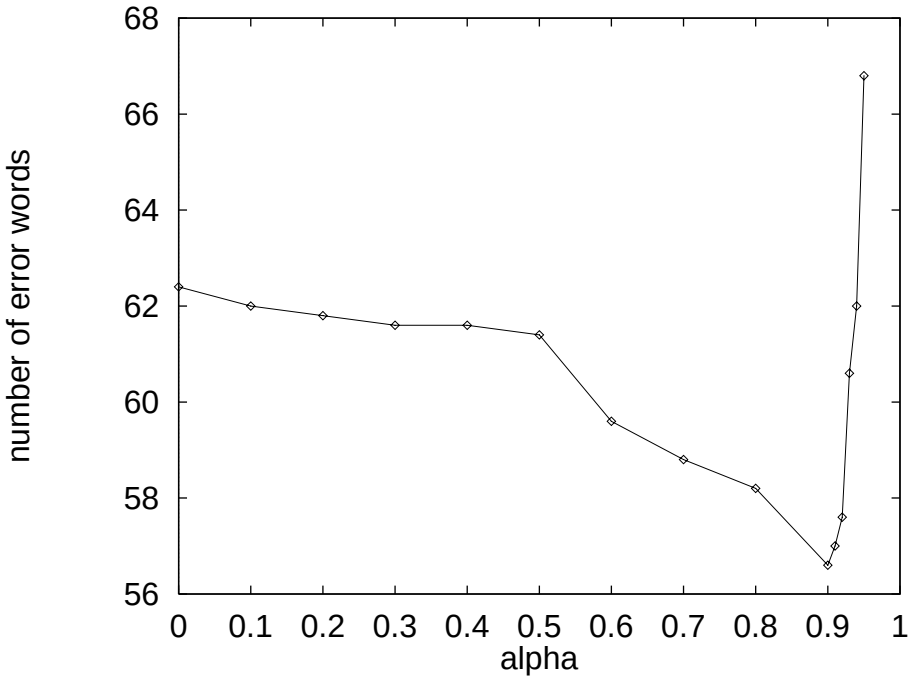

図 4.5: 評価データ 5 話者 (m116〜m120) の平均誤認識単語数の変化による結合係数 α の決定 (SNR45dB)

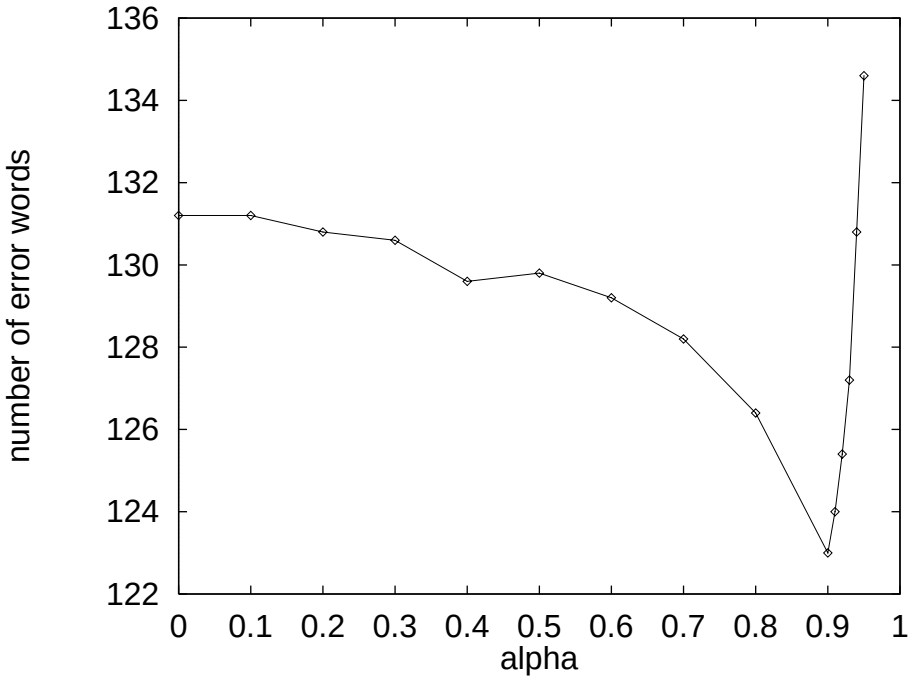

図 4.6: 評価データ 5 話者 (m116〜m120) の平均誤認識単語数の変化による結合係数 α の決定 (SNR25dB)

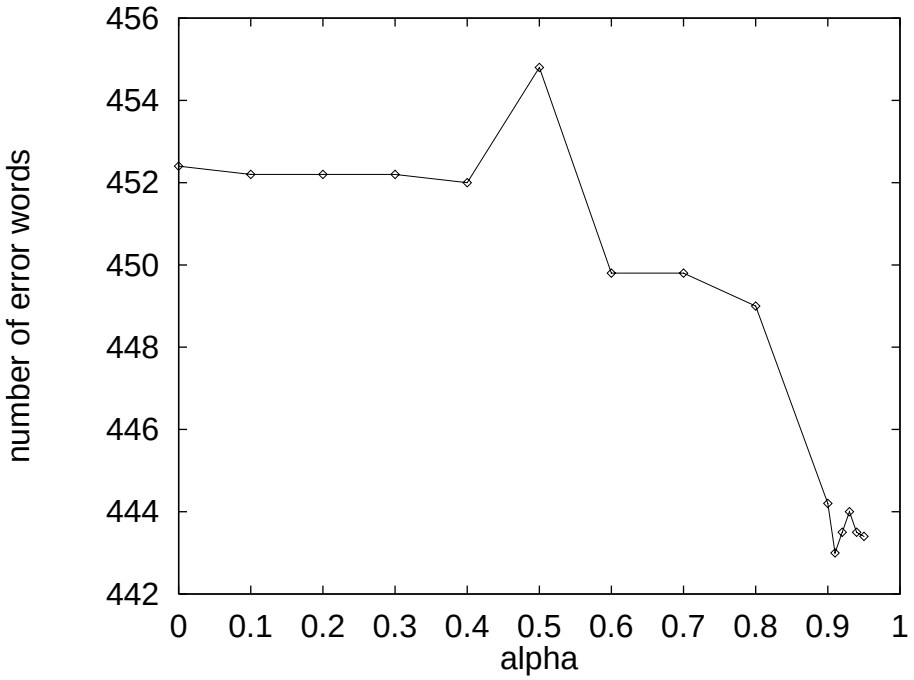

図 4.7: 評価データ 5 話者 (m116〜m120) の平均誤認識単語数の変化による結合係数 α の決定 (SNR5dB)

	SNR∞ $\alpha = 0.87$		SNR45dB $\alpha = 0.90$		SNR25dB $\alpha = 0.90$		SNR5dB $\alpha = 0.91$	
話者	改善数	認識率 (%)	改善数	認識率 (%)	改善数	認識率 (%)	改善数	認識率 (%)
m116	+3	87.7	+1	86.9	+3	74.8	+18	19.0
m117	+3	89.6	0	88.7	+3	82.1	+5	23.5
m118	+5	90.0	+14	89.2	+8	69.4	+5	6.2
m119	+4	90.2	+9	88.5	+17	73.5	+8	9.0
m120	0	93.1	+7	92.3	+10	81.9	+11	16.3
平均	+3	90.1	+5.8	89.1	+8.2	76.3	+9.4	14.8

表 4.3: 各 SNR において最適な結合係数 α での各話者の認識率

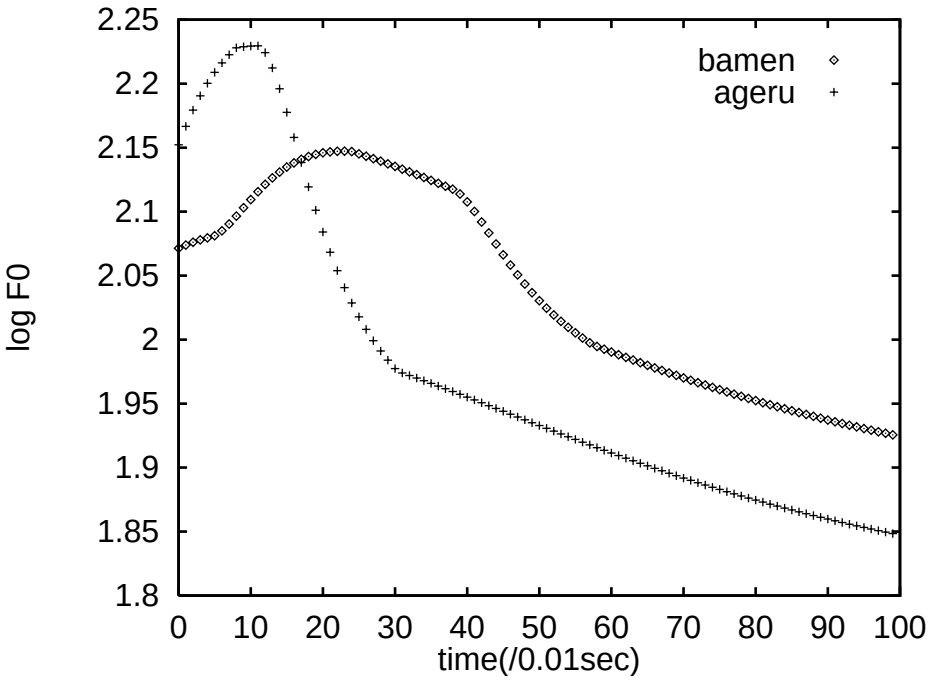

図 4.8: 改善された単語の (bamen:場面 → ageru:上げる) のピッチパターン

4.4.2 改善率の上限値に関する考察

ここでは 4.5 節の実験と同様に、ピッチパターンによって改善できる単語の上限数を調べる実験として、結合係数 α を入力毎に $0 < \alpha < 1$ の範囲で振らした場合に、誤認識から正解へと改善される可能性がある単語数を調べた。その結果を表 4.4に示す。表にはHMM のみの認識からの改善数、単語認識率をそれぞれの話者、SNR において示した。改善数の+は改善、−は改悪されたことを示している。この表を見ると、どの話者のどのSNR においても大きな誤認識単語数の改善結果が得られることが分る。4.5 節と同様に、何らかの基準で結合係数 α を入力毎に自動制御することができれば、韻律情報が雑音重畳単語認識に有効な特徴量となることを示している。

	SNR∞ $\alpha = 0.87$		SNR45dB $\alpha = 0.90$		SNR25dB $\alpha = 0.90$		SNR5dB $\alpha = 0.91$	
話者	改善数	認識率 (%)	改善数	認識率 (%)	改善数	認識率 (%)	改善数	認識率 (%)
m116	+35	93.8	+30	92.9	+64	86.5	+68	28.7
m117	+26	94.0	+26	93.7	+44	90.0	+104	42.5
m118	+32	95.2	+42	94.6	+86	84.4	+62	17.1
m119	+31	95.4	+40	94.4	+83	86.2	+87	24.2
m120	+21	97.1	+29	96.5	+61	91.7	+98	33.1
平均	+29	95.1	+33.4	94.4	+67.6	87.8	+83.8	29.1

表 4.4: 韻律情報による改善率の上限値

4.4.3 SNR と改善率の関係に関する検討

5.5 節、5.6 節から得られた結果を利用して、以下の式により各 SNR における改善率の検討を行った。韻律情報による改善数、HMM のみの認識時の誤認識単語数は各話者の平均値を用いた。

$$\text{改善率}\,(\%) = \frac{\text{韻律情報による改善数}}{HMM\text{のみの認識時の誤認識単語数}} \times 100$$

その結果を図 4.9 に示す。この図を見ると、結合係数 α 固定値の場合は SNR∞ から 25dB までの区間で 8.0%弱の改善率しか得られなかった。しかし、結合係数 α 可変値の場合は SNR∞ から 25dB までの区間で 50.0%強の改善率が得られ、この区間において韻律情報が認識率向上に有効である可能性のあることが分った。

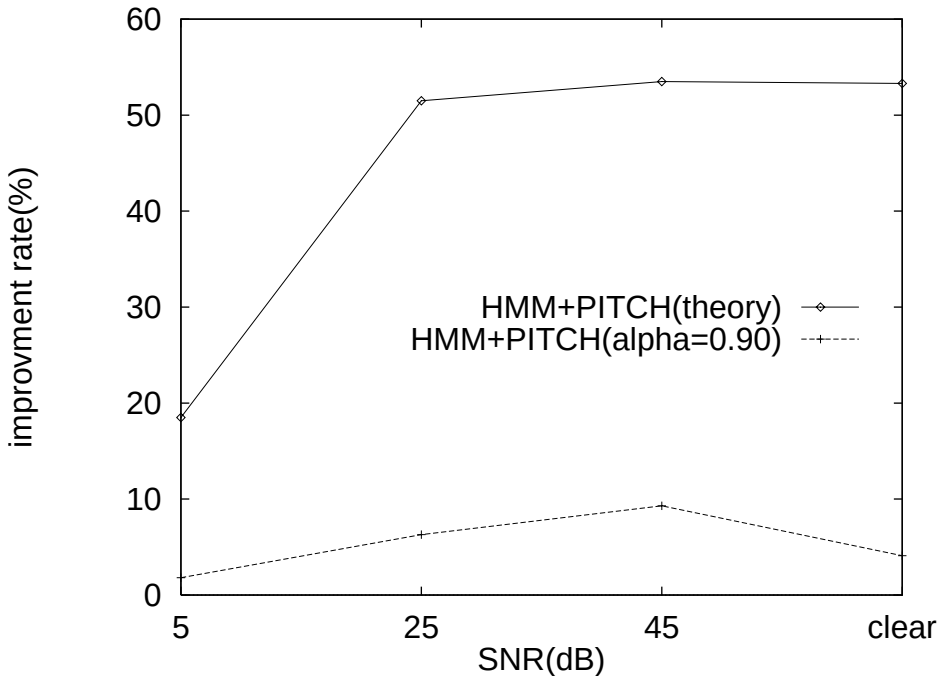

図 4.9: 各 SNR における改善率

4.5 音韻尤度の上位候補に韻律尤度を加えた認識実験

4.7 節の実験から、韻律情報を考慮する際に HMM による認識結果の順位およびスコアに閾値を適用すると認識率の向上が得られることが確認されている。そこで本章では、HMM の上位 2 位までの結果に韻律情報を考慮するという閾値を、結合係数 α 固定値 ($\alpha = 0.90$) 、結合係数 α 可変値 ($0 < \alpha < 1$) の場合に適用することによって評価データ (m116〜m120) の誤認識単語数の変化を調べた。その結果を表 4.5 、表 4.6 にそれぞれ示す。表には HMM のみの認識からの改善数、単語認識率をそれぞれの話者、SNR において示した。改善数の+は改善、−は改悪されたことを示している。また、5.6 節の式を利用したそれぞれの場合における各 SNR の改善率を図 4.10 に示した。

これらの結果をみると、HMM の認識結果の順位およびスコアに閾値を適用したことによって結合係数 α 固定値の場合も可変値の場合も、すべての SNR において改善率が大きく減少していることが分った。とくに雑音レベルが大きくなるほど改善率の割合が減少することが分かった。これは、韻律情報によって正解となるべき単語が、雑音レベルが大きい音声では HMM の認識結果の上位 2 位までに含まれていないことが原因であると考えられる。

このことから雑音重畳単語音声認識に韻律情報を考慮する際に、HMM の上位 2 位までの結果に韻律情報を考慮するという手法は適切でないことが分った。

	SNR∞		SNR45dB		SNR25dB		SNR5dB	
話者	改善数	認識率 (%)	改善数	認識率 (%)	改善数	認識率 (%)	改善数	認識率 (%)
m116	+2	87.5	−1	86.9	−1	74.0	+17	18.8
m117	+3	89.6	0	88.7	+3	82.1	+3	23.1
m118	+4	89.8	+13	89.0	+6	69.0	+3	5.8
m119	+2	89.8	+8	88.3	+17	73.5	+6	8.7
m120	−3	92.5	+7	92.3	+9	81.7	+5	15.2
平均	+1.6	89.8	+5.4	89.0	+6.8	76.1	+6.8	14.3

表 4.5: 音韻尤度で候補を絞った認識実験 (α 固定値)

	SNR∞		SNR45dB		SNR25dB		SNR5dB	
話者	改善数	認識率 (%)	改善数	認識率 (%)	改善数	認識率 (%)	改善数	認識率 (%)
m116	+28	92.5	+23	91.5	+41	82.1	+29	21.2
m117	+20	92.9	+19	92.3	+27	86.7	+31	28.5
m118	+29	94.6	+36	93.5	+46	76.7	+8	6.7
m119	+22	93.7	+29	92.3	+49	69.2	+13	10.0
m120	+16	96.2	+22	95.2	+38	87.3	+19	17.9
平均	+23	94.0	+25.8	93.0	+40.2	82.5	+20	16.8

表 4.6: 音韻尤度で候補を絞った認識実験 (α 可変値)

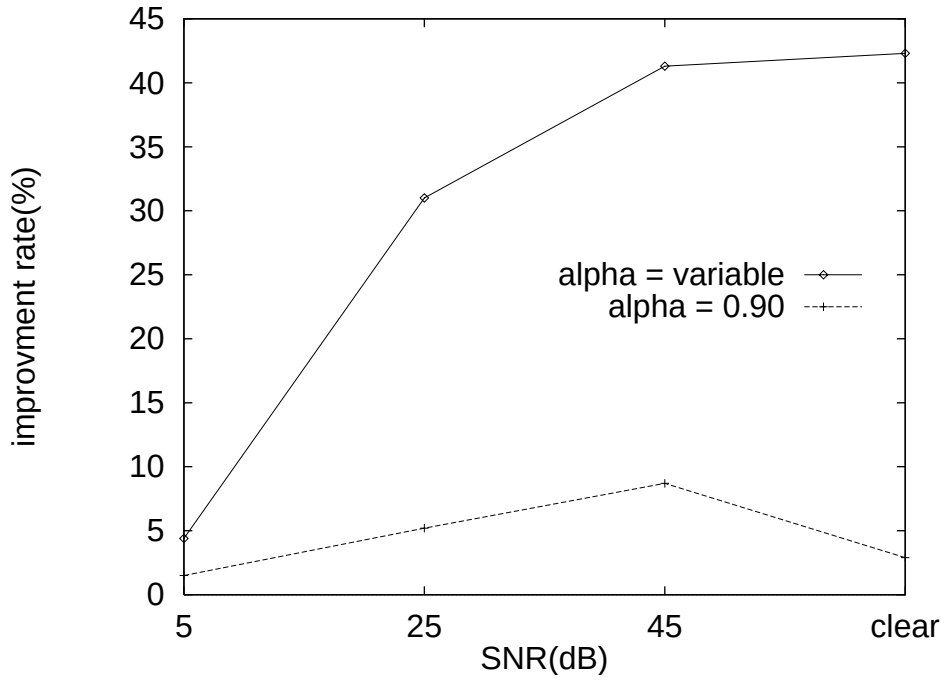

図 4.10: 各 SNR における改善率

4.6 まとめ

本章の実験により、韻律情報が雑音重畳単語音声の認識率向上に有効性のあることを示した。また、どの雑音領域においても結合係数 $\alpha = 0.90$ という値が認識率改善に最適な値であることを確認した。

本章では、4.7 章のような HMM の認識結果に閾値を考慮にいれることによる認識率の向上結果は確認されなかった。しかし、4.6 章と同様に、入力毎に結合係数 α を可変とした場合には大きな認識率向上結果が得られることが確認された。今後はこの制御法などについて検討を行う予定である。また、3.4.1 節と同様に、正解単語と誤認識単語とのピッチパターン形状が大きく異なっている場合に誤認識から正解へと改善されることが確認できたことから、アクセント型との関係についても検討を行う予定である。

第 5 章

韻律情報を用いた同音異義語認識実験

5.1 はじめに

3.4.1 節や 4.4.1 節から、韻律情報を考慮した際に、正解単語と誤認識単語とのピッチパターン形状が大きく異なっている場合に誤認識から正解へと改善されることが確認できた。しかし、実際にピッチパターンの形が異なっていれば正解単語の方を正しく認識できているのかということを確認するために、本章では、本論文で提案した韻律尤度計算法を用いた同音異義語の認識実験を行った。

5.2 実験条件

音声資料

実験には ATR の単語音声データ 5240 単語 (男性 10 話者) から抽出した同音異義語 697 単語を用いた。サンプリング周波数は 20kHz で、非雑音環境下における収録音声である。この内、男性 9 話者のデータを学習データとして用い、残りの男性 1 話者のデータを評価用データとして用いた。

ピッチパターン分析条件

FFT 分析	1024 ポイント
分析シフト (1 frame)	200 ポイント (10 msec)
ピッチ探索範囲	50Hz〜500Hz
ラグ窓幅	200 ポイント

表 5.1: ピッチパターン分析条件

5.3 提案した韻律尤度計算法を用いた同音異義語認識実験

本論文で提案した韻律尤度計算法を用いた同音異義語の認識実験結果を図 5.1 に示した。横軸は入力単語の韻律尤度、縦軸は入力単語に対する同音異義語の韻律尤度を表している。$Y = X$の直線は、入力単語とそれに対する同音異義語とのピッチパターンの形が全く一致しているために、同じ値の韻律尤度が計算されてしまい、認識が不可能な線を表している。この直線より上側は誤認識の領域で、下側が正解の領域である。

これらのことから、この線付近に存在する点というのは、入力単語とそれに対する同音異義語とのピッチパターンの形が類似していて、認識が困難な単語であると考えられる。しかし、この線から離れている点というのは、入力単語と同音異義語とのピッチパターンの形が異なっているもので、このような点が$Y = X$の直線より下側にある正解の領域に多く存在していることが分る。このことは、本論文で提案した韻律尤度計算法が、ピッチパターンの形が異なっている場合には、入力単語の韻律尤度の値を大きく計算し、正しく認識を行っていることを示している。

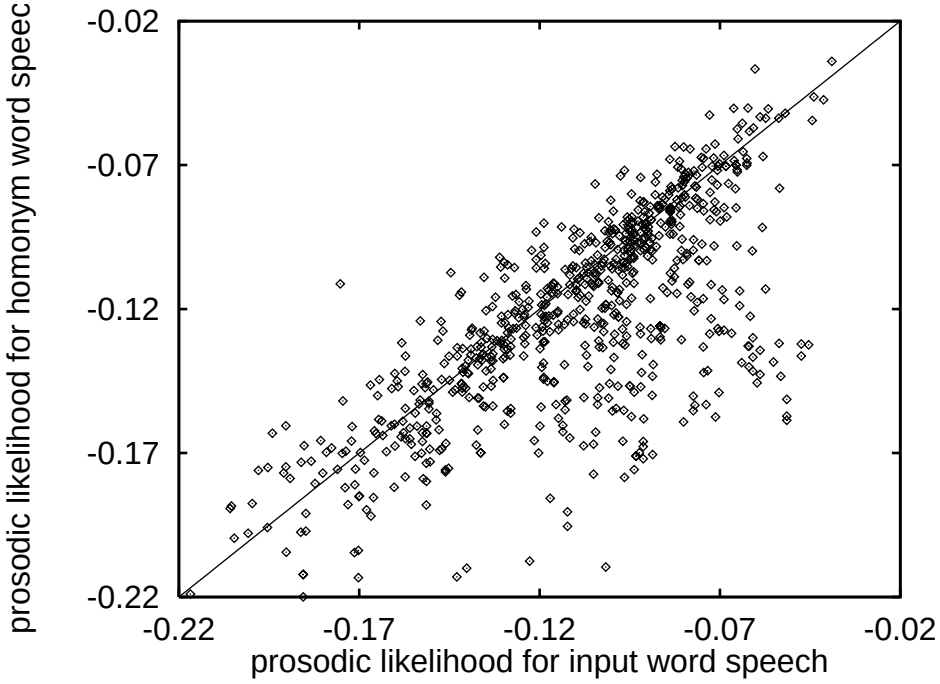

図 5.1: 韻律尤度計算法を用いた同音異義語認識実験

5.4　まとめ

本論文で提案した韻律尤度計算法が、ピッチパターンの形の違いを正しく認識し、同音異義語の認識において有効性のあることが確認できた。今後は、韻律尤度が大きく計算されるピッチパターンのアクセント型との関連について検討して行く予定である。

第 6 章

結論

6.1 研究結果

本研究により、韻律情報の一つであるピッチパターンが非雑音環境下および雑音環境下の単語音声認識の認識率向上に有効性のあることを示した。

非雑音環境下における単語音声認識においては、話者ごとに誤認識単語数が最も減少する最適な結合係数 α の値に多少のばらつきはあったが、すべての話者に共通な結合係数 α の値を適用しても平均誤認識単語数は HMM のみの認識時の平均誤認識単語数よりも減少することが確認された。また、1 話者に対して最適な結合係数 α の値を用いることに加えて、音韻情報の認識結果に閾値を適用することでさらなる認識率の向上が実現できることが確認された。さらに、結合係数 α の値を可変値にした場合には、すべての話者において誤認識単語数が HMM のみの認識時の誤認識単語数とくらべて約半分にまで減少できることが確認された。

雑音環境下における単語音声認識においては、ピッチパターンがどの SNR においてもほとんど白色雑音の影響を受けずに抽出され、耐雑音性にすぐれた特徴量であることが確認された。また、どの SNR においても各話者の平均誤認識単語数が最も減少する最適な結合係数 α の値がほぼ一致したものとなるとこが確認された。さらに、結合係数 α の値を可変値にした場合には非雑音環境下における単語音声認識と同様に、各話者の平均誤認識単語数が SNR∞〜25dB の区間では HMM のみの認識時の誤認識単語数とくらべて約半分にまで減少できることが確認された。

また、本論文で提案した韻律尤度計算法が、ピッチパターンの形の違いを正しく認識し、同音異義語の認識において有効性のあることが確認できた。

6.2 今後の課題

本研究の非雑音環境下における単語音声認識では、1 話者に対して最適な結合係数 α の値と音韻情報の閾値を適用することで大きな認識率の向上を実現したが、今後は他の話者すべての認識率を向上させる結合係数 α の値と音韻情報への閾値について検討する予定である。また、単語毎に結合係数 α を可変とした場合の制御法などについても検討を行う予定である。

また、雑音環境下の単語音声認識においては、非雑音環境下の時のような音韻情報の認識結果に閾値を考慮にいれることによる認識率の向上結果は確認されなかったが、入力毎に結合係数 α を可変とした場合には大きな認識率向上結果が得られることが確認されることから、今後はこの制御法などについて検討を行う予定である。

さらに、どちらの環境下の場合もクラスタ数 8 でクラスタリングを行った結果からピッチテンプレートや韻律辞書を作成したが、今後はクラスタ数を変化させた場合の認識率への有効性や、韻律尤度の計算法についても検討を行う予定である。

そして、韻律情報によって正解となる単語のアクセント型との関連についても検討して行く予定である。

謝辞

本研究を行うにあたり、全般的な御指導・御助言を賜わった、北陸先端科学技術大学院大学情報科学研究科 木村 正行教授に深く感謝致します。

また、本研究を進めていく上で必要不可欠である音声認識に関する知識の御指導・御意見を賜わった、同研究科 下平 博助教授に深く感謝致します。

北陸先端科学技術大学院大学情報科学研究科 中井 満助手には、本研究の進行や問題点に関する適切な御意見・御助言を賜わり深く感謝致します。

同研究科木村・下平研究室の高倉健次氏には、本研究に関する御意見・御助言を賜わり深く感謝致します。さらに、日頃から御討論、御協力を頂いた同研究科木村・下平研究室の皆様に心から感謝致します。

研究発表一覧

[1] 川崎真護, 中井満, 下平博:「アクセントピッチパターンを利用した情報を単語音声認識」, 電気関係学会北陸支部連合大会,B-56, 平成 9 年.

[2] 川崎真護, 中井満, 下平博:「F_0生成モデルに基づくピッチパターン整合を用いた雑音重畳単語音声の認識」, 日本音響学会 平成 10 年度春季研究発表会,3-6-14.

参考文献

[1] 妹背, 松本:「雑音環境下における周波数重み付け HMM の改良」, 平 5 秋音講論,3-8-9,(1993.10).

[2] 高橋, 山口, 嵯峨山:「スペクトルサブトラクションと NOVO 合成を用いた雑音下音声認識」, 平 8 秋音講論,2-Q-8,(1996.9)

[3] 岩田, 宇佐川, 江端:「ディジタル蝸牛モデルを用いた騒音下での音声パラメータ抽出」, 信学技報,SP94-33,1994-07.

[4] 高橋, 松永, 嵯峨山:「ピッチパタン情報を考慮した単語音声認識」, 信学技報,SP90-17,1990.

[5] 相川, 鹿野:「パワーとスペクトルの一括ベクトル量子化による単語音声認識」, 信学論,J68-D.3.pp.316-322,(1985.3)

[6] 笹沼, 板橋:「韻律情報を用いた多言語音声の分類」, 平 9 春音講論,3-6-14,(1997.3)

[7] 中井, 下平, 嵯峨山:「ピッチパタンのクラスタリングに基づく不特定話者連続音声の句境界検出」, 信学論,Vol.J77-A,pp.206-214,(1994.2)

[8] 嵯峨山, 古井:「ラグ窓法を用いたピッチの抽出の一方法」, 信学総全大,1235,(1978,3)

[9] 藤崎, 広瀬, 高橋, 横尾:「連続音声中におけるアクセント成分の実現」, 音声研資,S84-36,1984.

[10] Y.Linde,A.Buzo and R.M.Gray:「An Algorithm for Vector Quantizer Design」,IEEE $Trans.Commu.$,Vol.COM-28,No.1,pp.85-95,(1980-01).