概要

従来の音声認識では，MFCC(Mel frequency cepstral coefficients) が一般的な特徴量として使用されている．この特徴量は，音声に含まれている位相情報を使用していない．しかし，位相情報と併用することで認識精度の向上が報告されている [1]．また，話者認識の分野でも精度向上が報告されている [2]．そこで本研究では，位相情報を含む特徴量を用いて単語音声認識の実験を行い，認識精度の向上を目指した．提案した特徴量は，離散フーリェ変換の出力に対して，実数成分と虚数成分の値を独立した情報として扱う．通常，離散フーリェ変換の出力は，実数成分と虚数成分について絶対値をとったパワースペクトルが使用されるが，これは位相情報を含んでいないためである．実験の結果，提案した位相情報を含む特徴量の単語認識率は従来の特徴量に比べて認識精度が減少した．以上のことから，本研究で提案した特徴量は位相情報を含んでいるが，認識精度には影響しないという結果となった．

目 次

図 目 次

表 目 次

1　はじめに

文字入力インタフェースの一つとして，音声認識という手法が研究されている．これは音声による入力が，キーボードなどから手を使用して入力するよりも簡易で早いためである．また，同時に別の作業を行う場合など，手を使用できない状況においても，音声によって入力が可能となる．主な使用例として，カーナビゲーション (音声による操作，目的地の入力)，携帯端末に対する入力 (情報検索時，文章入力)，音声対話受付案内システムなどがある．

音声認識システムは以下の処理を必要とする．入力となる音声に対するモデルの作成，音声から抽出された特徴量の時系列に対する尤度の計算，そして計算された尤度を最大にするモデル (文字列) の出力である．特に特徴量の抽出に関して，これまで音声認識の分野では人の聴覚原理に基づいて研究されている [4]．そして人の聴覚は位相の変化に鈍感であるため，位相情報は必要ないとされていた．従来の音声認識では，MFCC(Mel frequency cepstral coefficients) が一般的な特徴量として使用されている．この特徴量は，音声に含まれている位相情報を使用していない．しかし，位相情報と併用することで認識精度の向上が報告されている [1]．また，話者認識の分野でも精度向上が報告されている [2]．そこで本研究では，位相情報を含む特徴量を用いて単語音声認識の実験を行い，認識精度の向上を目指す．

2　音声認識

音声認識とは，音声波に含まれる情報を計算機によって抽出し，判定することである．音声認識の処理は，以下のような利点がある．

- 音声の入力は，キーボードや押しボタンなどに比べて，操作に慣れる必要がないため使い易い．
- 情報の入力速度がタイピングの約3〜4倍，手書き文字と比べると約8〜10倍と速い．
- 手足，眼，耳などの器官で同時に別の作業を行う場合に並列的に、あるいは動きながらでも情報の入力が可能である．

しかし，その実現は容易ではない．以下に音声認識の難しさとして，4点について説明する．音響モデルに関連して，一つは調音結合の効果として，単語や文章を発声したときの各音素スペクトルが，前後の音素の影響を受けて変化する点である．また区分化の難しさとして，音声は草書で書かれた文字列のようなもので，音素や単語の境界を決定することが難しい．これに関しては，音声の開始と終了の検出も同様に困難である．三点目として，個人差による変動がある．同じ言葉でも，話し方や発声器官の違いから人によって音声は異なる．また，同じ人の音声でも，雑音や伝送歪みなどによって変動する．最後に，言語モデルに関連して，そのモデルのあいまい性が挙げられる．話し言葉には書き言葉とは違う多くの文法的ゆらぎがあり，モデル化が非常に難しい．

2.1　音声認識の構成

一般に人が発声した音声を計算機などで認識する過程は，図1のように通信理論 (情報処理論) の問題として，確率モデルを用いて定式化できる．話者が文を考える過程が文発生部で，これを通信理論の情報源に対応させる．音声認識システムを音響処理部と言語復号部に分ける．話者による発声部と音響処理部を合わせて，一つの音響チャンネルとしてモデル化し，これを歪み (雑音) のある通信路に対応させる．音声認識システムの主な部分である言語復号部を復号部に対応させる．話者はまず，情報源に対応する文 ω を頭の中で組み立て，それに基づいて，その話者の発話習慣に従って音声波形 s を生成する．s には通常，話者の個人差，付加雑音，伝送歪みなどが重畳している．音響処理部は音声波形データの分析・変換を行って，例えば短時間スペクトルなどの時系列データ (ベクトル系列)y を出力する．言語復号部は y から送信文の推定値として $\hat{\omega}$ を出力する．

$\hat{\omega}$ は，事後確率 $P(\omega|y)$ が最大になるように推定する． $P(\omega|y)$ を直接求めるのは，通常困難であるので，ベイズ則によって，次式を満たすように推定する．

$$P(\hat{\omega}|y) = \max_{\omega} \frac{p(y|\omega)P(\omega)}{P(y)} \tag{1}$$

ここで，$P(y)$ は ω に無関係であるので無視できる．尤度 $P(y|\omega)$ は音響モデルによって得られ，文 ω が発生される事前確率 $\mathrm{P}(\omega)$ は言語モデルによって得られる．従って音声認識では，音響モデルと言語モデルをいかに作り，$P(y|\omega)$ と $P(\omega)$ を計算するがが重要である．

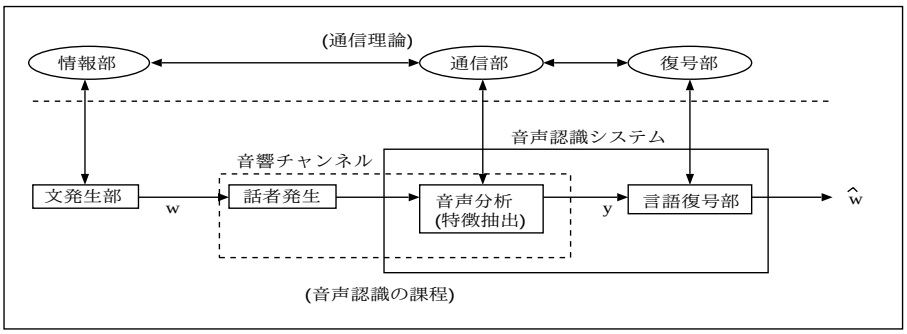

図 1: 音声認識過程の確率モデル

2.2 音声認識の分類

音声認識の形態は，以下の通り分類できる．

(a) 対象となる音声による分類

単語音声認識:区切って発音した単語の認識.

連続音声認識:単語を連続して発声した音声の認識.

連続音声認識は，さらに以下の通り分類される．

連続単語音声認識:連続数字のように，比較的少数の語彙を対象とし，言語的知識は用いない認識.

文音声認識，会話音声認識，音声理解:比較的多数の語彙を対象とし，言語的知識を用いて，その意味内容の理解を目的とする．

(b) **対象となる話者による分類**

特定話者音声認識:学習を行った特定の話者の音声のみを認識.

不特定話者音声認識:話者を限定せず，不特定の音声を認識.

話者適応:新しい話者の音声を用いて認識装置を自動的にその話者の声に適応させ，その後あるいは適応と同時に認識.

2.3 HMM とは

音声認識は，パターン認識の一分野であるため，音声波形から特徴量を抽出した後の処理は，通常のパターン認識技術と本質的には同じである．その違いは，音声が時系列パターンであること，そして言語情報 (音素，単語，構文，意味等) の制約を受けることである．近年，音声の時系列パターンに対して統計的・確率的なパターン認識の手法として HMM(Hidden Markov Model) が一般的に使用されている [3].

HMM は，出力シンボルによって一意に状態遷移先が決まらないという意味での非決定有限状態オートマトンとして定義される．一般に，マルコフモデルは最終状態の概念がない．しかし音声認識に用いる場合は初期状態，最終状態を設定する．音声認識で用いられる HMM は，left-to-right モデルと呼ばれる．このモデルでは，状態と出力シンボルの二過程を考え, 状態が確率的に遷移するときに対応して確率的にシンボルを出力する．このとき観測できるのはシンボル系列だけであることから隠れマルコフモデルと呼ばれる.

HMM による音声認識では，各カテゴリの HMM に対して入力パターンの特徴量の時系列に対する尤度を求め, それを最大にするモデルに対応するカテゴリを認識結果とするのが基本手法である.

HMM の音声認識における利点を以下に示す.

- 個人差や調音結合，発声法 (強さ，速さ，明瞭さ) 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.
- 従って，統計理論や情報理論・確率課程論による論理的展開がしやすい.
- 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，両者を統合しやすい.

- 認識時の計算量が比較的少ない.

しかし，次のような問題点もある.

- モデルの設計法が確立されていなく，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.(学習に時間を要するが，認識には時間がかからない点が人の聴覚システムに似ている)
- 音声の過渡的パターンの表現に乏しく，時系列パターンの中の 2 時点におけるパターンの相関が考慮できない.

2.4 HMM の種類

HMM にはスペクトルパターンの表現方法によって二種類に分類される．また両方の中間的な性質を持った HMM がある．以下にそれぞれの特徴を示す.

離散分布モデル (離散 HMM)

出現されるスペクトルパターンは，有限個のシンボルの組合せで表現される．出現確率は，スペクトルパターンのクラスタ化 (ベクトル量子化) によって代表スペクトルパターン (符号ベクトル) を生成し，各符号ベクトルの出現確率の組合せによって表す.

連続分布モデル (連続 HMM)

出現するスペクトルパターンは，連続値で表現される．出現確率は，単一ガウス分布 (正規分布)，または混合ガウス分布が用いられ，パラメータの自由度を減らすために無相関ガウス分布 (Diagonal) が用いられることが多い.

半連続分布モデル (半連続 HMM)

連続分布モデルと離散分布モデルの中間の性質を持つ．これは連続分布モデルにおける混合ガウス分布を，全てのモデルの全ての状態で共通にし，各分布の重みだけを変えるようにしたものである．結び混合 (tied-mixture) 分布モデルとも呼ばれる．離散分布モデルにおける各符号化ベクトルに確率分布を持たせたものということもできる.

実際の音声認識では，対象に応じて適切に状態数やモデルを決定する必要がある．モデルの自由度を大きくすれば，きめ細かい変動が表現できるが，推定すべきモデルパラメータが多くなり，推定精度が悪くなる．次章では簡略化した HMM の例を示す.

2.5 HMM の例

音声認識に用いられる HMM は，left-to-right モデルである．left-to-right モデルの例を図 2 に示す.

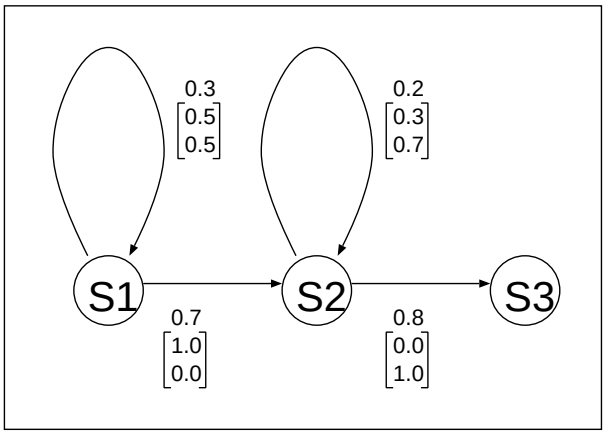

図 2: left-to-right モデルの例

例の HMM は 3 状態で構成され，出力は有限個のシンボル a と b の 2 種類である．最終状態は S_3 とし，図 2 のような遷移のみを行うものとする.
a_{ij} は，状態 S_i から状態 S_j への遷移確率を示しており，a_{12} では遷移確率が 0.7 となる．また，[ ] 内の数字は上段がラベル a の出力確率，下段がラベル b の出力確率を示す．状態 S_1 を例にとると，状態 S_1 から S_2 への遷移は 0.7 の確率で行われ，遷移の際に a を出力する確率は 1.0，b を出力する確率は 0.0 であることを意味する．出力シンボルが “aab” である場合の状態遷移系列と尤度 (確率) の計算を以下に示す.

状態遷移系列 $S_1 - S_1 - S_2 - S_3$

$$0.3 * 0.8 * 0.7 * 1.0 * 0.8 * 1.0 = 0.1344$$

状態遷移系列 $S_1 - S_2 - S_2 - S_3$

$$0.7 * 1.0 * 0.2 * 0.4 * 0.8 * 1.0 = 0.0448$$

この HMM が “aab” を出力する確率は合計値で決定する.

0.1344+0.0448=0.1792

2.6 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ベクトルの時系列である．このとき，各 HMM モデルによって y が生起する確率 (尤度)$P(y/M)$(ここで 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．単語音声認識システムの流れを図 3 に示す．本研究で扱う処理は破線で囲われた範囲である．

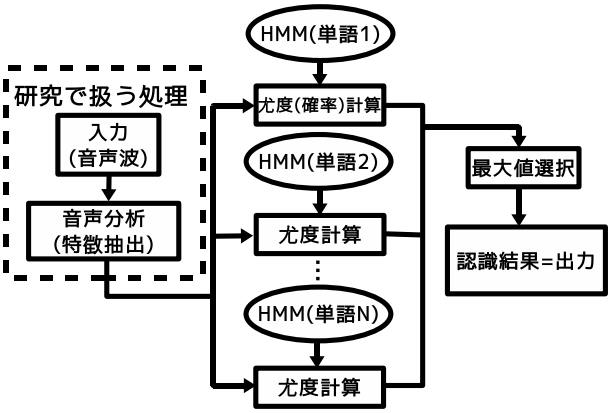

図 3: 単語音声認識システムの流れ

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

$$P(y \mid M) = \sum_{i_0, i_1, \cdots, i_T} P(y \mid q, M) \cdot P(q \mid M) \tag{2}$$

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 α (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．

$$\alpha(i,t) = \begin{cases} \pi_i & (t = 0) \\ \sum_j \alpha(j, t-1) \cdot \alpha_{ji} \cdot b_{ji}(y_t) & (t = 1, 2, \cdots, T) \end{cases} \tag{3}$$

これを計算し，最後に以下を求めれば良い．

$$P(y \mid M) = \sum_{i, q \in F} \alpha(i, T) \tag{4}$$

$P(y|M)$ を厳密に求めないで，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 Viterbi アルゴリズムと呼ばれる方法もある．尤度は，各遷移での確率値を対数変換しておくことで高速に求めることができる．このアルゴリズムを以下に示す．$i = 1, 2, \cdots, S$ において，

$$f'(i,t) = \begin{cases} \log \pi_i & (t = 0) \\ \max_j \{f'(i, t-1) + \log a_{ji} b_{ji}(y_t)\} & (t = 1, 2, \cdots, T) \end{cases} \tag{5}$$

を計算し，対数尤度

$$L = \max_{i, s_i \in F} f'(i,t) \tag{6}$$

を求める．この Viterbi アルゴリズムによる利点は以下のようなものである．

- 計算値のダイナミックレンジが小さく，アンダーフロー問題を解消できる．
- 計算量が少ない．
- 音声認識性能がほとんど変わらない．

このため Viterbi アルゴリズムは広く用いられている．

3　音声分析

3.1　音声の特徴量抽出

音声波形は，そのもの全てを用いたのでは情報量が多すぎる．そのため，音声から切り出された短時間の信号が定常確率過程に従うと仮定して，スペクトル解析を行う．すなわち，与えられた信号 $s(n)$ に長さ N の分析窓を掛けることで以下のように信号系列 $s_w(m;l)$ を取り出す．

$$s_w(m;l) = \sum_{m=0}^{N-1} w(m)s(l+m)(l = 0, T, 2T, \cdots) \tag{7}$$

ここで，添え字 l は，信号の切出し位置に対応している．すなわち，l を一定間隔 T で増加させることで，定常とみなされる長さ N の音声信号系列 $s_w(n)(n = 0, \cdots, N-1)$ が間隔 T で得られる．この処理はフレーム化処理と呼ばれ，N をフレーム長，T をシフト幅と呼ぶ．また，フレーム化処理を行う窓関数 $w(n)$ としては，ハミング窓やハニング窓がしばしば用いられるが，本研究では使用しない (直接切り出した波形について直接処理を行う) ため，説明は割愛する．フレーム化処理によって得られた音声信号系列の短時間フーリエスペクトルは，離散フーリエ変換 (DFT) により以下で与えられる．

$$S(e^{j\omega}) = \sum_{n=0}^{N-1} s_w(n)e^{-j\omega n} \tag{8}$$

実際の信号処理過程では，離散フーリエ変換 (DFT) をその高速算法である FFT を用いて実行し，当該音声区間のスペクトル表現とすることが一般的である．すなわち，

$$S'(k) = S(e^{j\frac{2\pi}{N}k}) = \sum_{n=0}^{N-1} s_w(n)e^{-j\frac{2\pi}{N}kn}(k = 0, \cdots, N-1) \tag{9}$$

なる複素数系列 $S'(k)$ が音声のスペクトル表現として最も一般的に用いられる．音声信号の音素的特徴は主として調音フィルタの振幅伝達特性に含まれている．従って，音声認識においては，離散フーリェ変換の出力の絶対値であるパワースペクトルが注目すべきスペクトル表現である．このパワースペクトル (FFTpower) についても特徴量として扱う．

3.2 ケプストラム分析

ケプストラム (cepstrum) は，波形のパワースペクトル $|S(e^{j\omega})|$ の対数の逆フーリエ変換として定義される．つまり，式 9 の出力を逆フーリェ変換すると，

$$c_n = \frac{1}{N}\sum_{k=0}^{N-1} log|S(k)|e^{j2\pi kn/N}(0 \leq n \leq N-1) \tag{10}$$

となる．この出力は LFCC(Linear Frequency Cepstrum Coefficient) という特徴量になる．ケプストラムという言葉は，スペクトルを逆変換するという意味から，spectrum をもじって作った造語であり，その変数は frequency をもじってケフレンシー (quefrency) と呼ばれる．従来の音声認識では, 特徴パラメータとしてケプストラムが使われてきた. ケプストラムは低次にフォルマント情報を高次にピッチ情報を含んでいる. しかしピッチ情報は正確なピッチ周波数の抽出が困難であるため, 音声認識ではフォルント情報のみを使用する．

3.3 MFCC

最も一般的に使用されるケプストラムは MFCCC(Mel Frequency Cepstrum Coefficient) である．この特徴量は，パワースペクトルを少ない次数で効率的に表現するために，メル分割されたフィルタバンクの対数パワーを使用する．つまり，人間の聴覚の特性にあわせて低周波部分は細かく，高周波部分は粗く調べるためメルスケールに沿って等間隔に配置された三角関数のフィルタをかける．この三角関数の個数がフィルタバンクのチャンネルのチャンネル数 (特徴量のベクトル数) を表している．周波数メル分割の式は，

$$Mel(f) = 2592\log_{10}(1 + \frac{f}{700}) \tag{11}$$

となる．このフィルタバンクの出力に対数をとったものを FBANK と呼ぶ．最終的に，フィルタバンク分析により得られた出力を離散コサイン変換 (逆フーリェ変換) することで，MFCC が求められる．

$$c_i = \sqrt{\frac{2}{N}}\sum_{j=1}^{N} m_j \cos(\frac{\pi i}{N}(j-0.5)) \tag{12}$$

N はフィルタバンクチャンネル数を表し,m_j は対数フィルタバンクの振幅を表す.

3.4 研究の着目点

3.4.1 位相情報とは

従来の音声認識ではMFCCを主に使用しており，音声に含まれている位相情報は無視されている．しかし近年，位相情報と併用することで認識精度の向上が報告されている[1]．また，話者認識の分野でも精度向上が報告されている[2]．このため，位相情報を利用した特徴量に着目する．

位相情報は，音源波形の特徴によって大きく影響を受け，声道の形によっても影響される．一般に使用される特徴抽出では離散フーリェ変換は以下の式で表される(式8と同様)．

$$S(e^{j\theta(\omega,t)}) = \sum_{n=0}^{N-1} Input(n)e^{-j\theta(\omega,t)} \tag{13}$$

ここで，同じ角周波数ωでも切り出す位置によって位相情報$\theta(\omega,t)$が異なってしまう問題が生じる．また，位相パラメータθは0〜2πの範囲を超える場合がありえ，$\pi-\theta1$と$\theta2=-\pi+\theta1$では$\theta1$が小さい場合に，本来位相差が小さいにも関わらず$|\pi-\theta1-\theta2|=2\pi-2\theta1$と，大きな差として比較されてしまう(本来は0に近い値である)．これは位相が連続値でないために生じる問題である．このため本研究では，$\theta1$に対して$cos\theta1$と$sin\theta1$という変換を行い，$\theta1$に対応する座標値として位相情報を用いる．つまり，離散フーリェ変換は複素数として出力されるため，振幅の情報と位相の情報がある．図4に複素数平面における情報を示す．

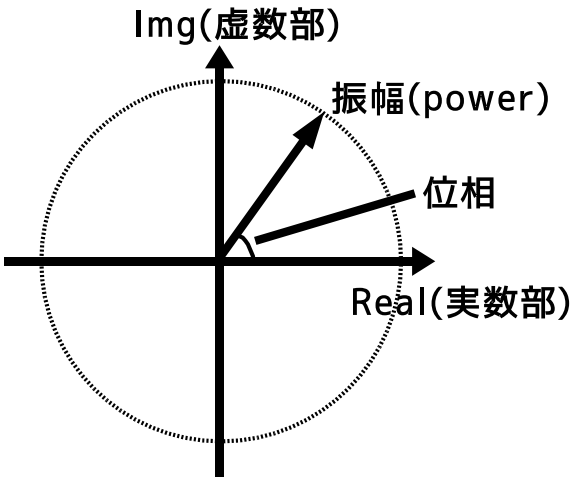

図 4: 複素数平面における情報

従来の特徴量抽出では，絶対値をとったパワースペクトル (振幅情報) のみが使用されているため，位相の情報を除外している．

3.4.2 本研究で提案する特徴量

本研究で提案する特徴量は，入力となる音声波形に対して離散フーリェ変換を行い，実数成分 (Real) と虚数成分 (Img) の値を，独立した情報として用いる．具体的には，式 13 に対してオイラーの公式 ($e^{j\theta} = cos\theta + jsin\theta$) を用いて，次のように扱う．

$$\text{実数成分}\,(Real) = \sum_{n=0}^{N-1} Input(n) \times \cos\theta \tag{14}$$

$$\text{虚数成分}\,(Img) = \sum_{n=0}^{N-1} Input(n) \times (-\sin\theta) \tag{15}$$

次に出力波形の例を示す．入力として音声「a」の 1 フレームを図 5 に示す．

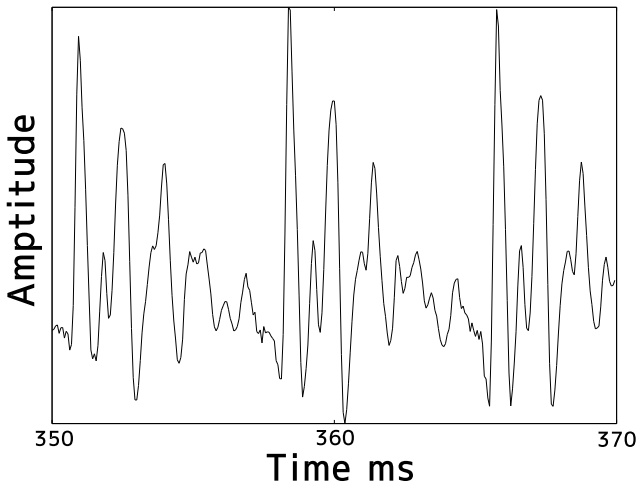

図 5: 音声波形「a」の 1 フレーム

絶対値をとったパワースペクトルを図 6 に，提案する特徴量として，実数成分を図 7 に，虚数成分を図 8 に示す．

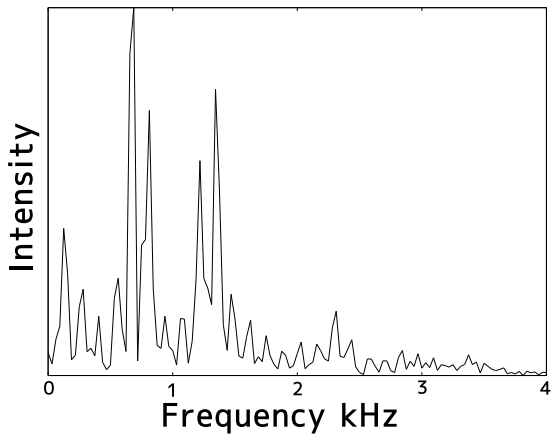

図 6: パワースペクトル (FFTpower)

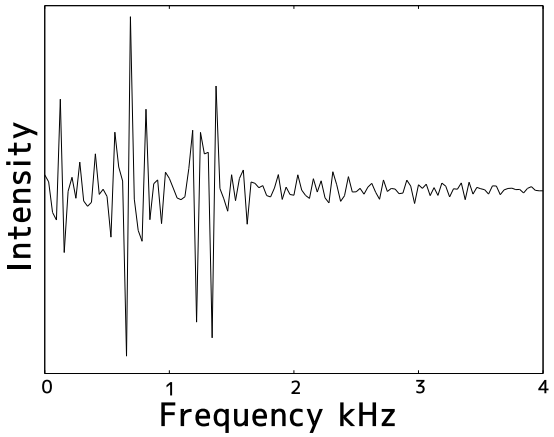

図 7: フーリェ変換後の実数成分

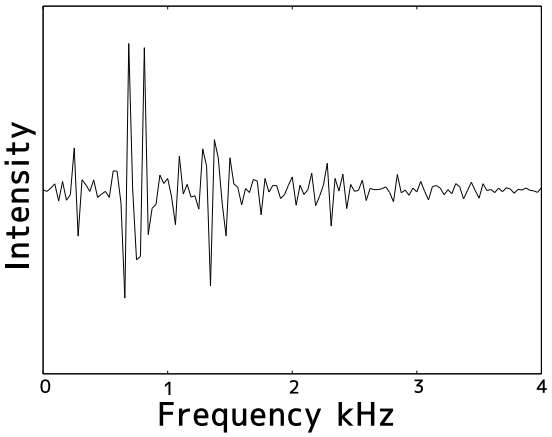

図 8: フーリェ変換後の虚数成分

パワースペクトルは実数成分と虚数成分の値の絶対値である．提案する特徴量は，二つの情報を独立して用いることで，位相情報を含めた値となる．位相情報を含んでいる点がパワースペクトルとは異なる．本研究ではこの特徴量を FFT_Real_Img と呼ぶ.

4 実験

4.1 実験データ

本研究では，実験のためにATR単語発話データベースAsetの男性話者3名(mau, mms, mxm)，女性話者3名(faf, ftk, fyn)のデータを使用する．データベースには話者毎に5,240単語の音声データがある．また，各音声データには，人手によって付与された音素境界位置情報が与えられる．本研究では，学習データとして，各話者の奇数番号データ(2,620単語)，評価データとして，各話者の偶数番号データ(2,620単語)を使用する．

4.2 実験条件

評価実験は，男性話者3名と女性話者3名で行う．実験には音声認識ツールのHTK [7]を使用する．HMMの共分散行列にはDiagonal-covarianceを使用する．その他の実験条件は表1に示す．実験条件は話者ごとに統一している．特徴量のベクトル数は同一にするのが困難であるため同じではない．

表 1: 実験環境

基本周波数	16kHz
フレームの長さ	20ms
シフト幅	10ms
音響モデル	状態数 3 混合数 話者に依存
特徴量 (特徴ベクトル数)	MFCC(14次元) LFCC(64次元) パワースペクトル(FFTpower)(256次元) 提案特徴量(FFT_Real_Img)(256+256次元) FFTpower+FFT_Real_Img(二手法併用)(256+256+256次元)
共分散行列	Diagonal-covariance

4.3 実験結果

表 2 に特定話者 (学習データと評価データが同一話者) における単語音声認識の実験結果を示す．表中の行は話者，列は特徴量である．

表 2: 実験結果 1(単語認識率 [%])

特徴量＼話者	mau	mms	mxm	faf	ftk	fyn
MFCC(14 次元)	94.50	92.44	91.56	93.02	93.66	92.37
FFTpower(256 次元)	88.89	86.03	85.99	89.01	81.91	86.68
LFCC(64 次元)	92.06	89.16	88.82	89.05	86.49	88.89
FFT_Real_Img (256+256 次元)	85.42	82.21	80.84	83.05	79.20	84.73
FFTpower +FFT_Real_Img (256+256+256 次元)	86.91	84.62	84.47	86.37	80.15	86.11

実験結果から，本研究で提案した位相情報を利用した特徴量 (FFT_Real_Img) を用いた場合，従来の特徴量に比べて認識精度の向上は見られない．特に，パワースペクトル (FFTpower) と FFT_Real_Img を比べると認識率が減少している点が問題である．この二手法に着目した結果を図 9 に示す．

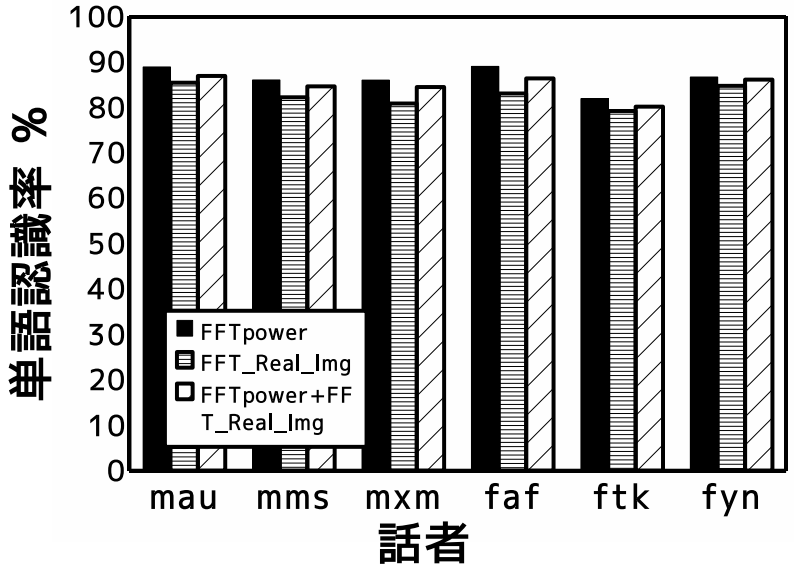

図 9: 実験結果 2(提案手法とパワースペクトルを比較した単語認識率 [%])

この図から，FFTpower の結果に対して提案した特徴量，そして両者を併用した場合，いずれにしても精度が低下していることが分かる．この二つの特徴量については，本質的な情報量の差がないため，同程度以上の結果が得られると考えていた．次章にて結果の分析を行う．

5　考察

実験結果の分析のため，音素認識の結果について調査した．使用している音素は母音が 6 種類 (無音 pau を除く)，子音が 27 種類である．結果は全ての単語に含まれている音素の合計 (例:音声「aka」では a を 2 回として計算) に対する認識結果である．実験条件は話者毎に統一しているため男性話者 mxm について分析を行った．表??に，パワースペクトル (FFTpower) と提案した特徴量 (FFT_Real_Img) の音素認識の結果を示す．

表 3: 男性話者 mxm の音素認識率 [%]

	FFTpower	FFT_Real_Img
母音	99.02	98.82
子音	85.96	79.49

以上の結果から，母音については特徴量による認識精度の差が少ないことが分かる．しかし，子音についてはその差が大きい．この原因として，学習に使用されたデータ数の差が考えられる．一般に発話された単語中に出現する音素としては，母音が多い．つまり，学習において十分な量が使用された母音では高い認識精度が得られ，学習データが不十分な子音において大きく認識精度が低下していると考えられる．表 4 に，実験において使用された母音と子音の学習データ数を示す．尚，話者毎に全ての実験において学習データの数と分布は同じ条件となる．

表 4: 話者 mxm の音素と学習データの分布 [%]

音素	学習データ数	音素	学習データ数	音素	学習データ数	音素	学習データ数
N	553	gy	14	n	276	t	373
a	1785	h	237	ny	9	ts	225
b	228	hy	10	o	1380	u	2385
by	4	i	1668	p	15	w	88
ch	143	j	193	q	119	y	202
d	178	k	1219	r	684	z	125
e	836	ky	58	ry	41		
f	77	m	492	s	588		
g	275	my	4	sh	403		

図 10 に学習データに対する子音の音素認識率の分布を示す.

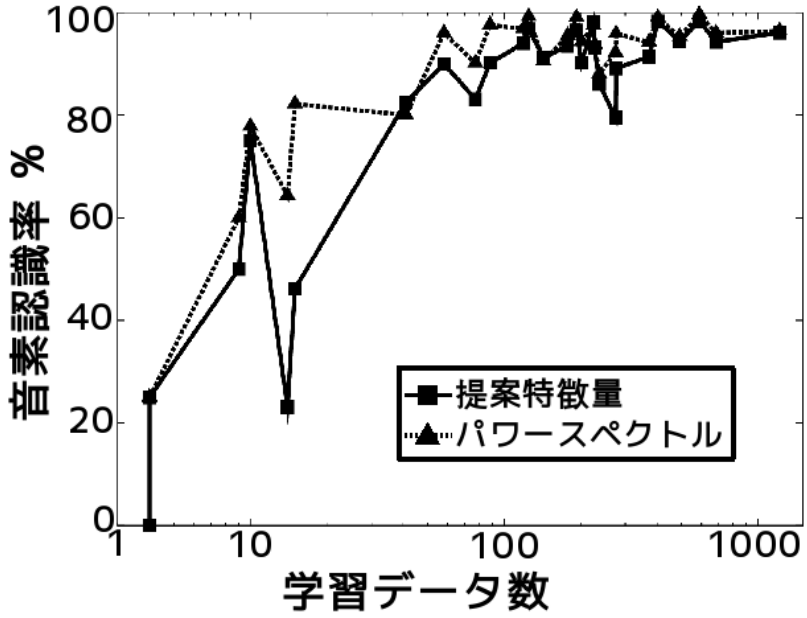

図 10: 学習データに対する子音の音素認識率 [%] の分布

このことから，特定の音素について著しい精度の減少が見られる．特に差の見られた音素について学習データと音素認識率を表 5 に示す.

表 5: 特に差が生じた同一音素に置ける認識率 [%]

音素 (学習データ数)	パワースペクトルの結果	提案手法の結果
by(4)	25.0	0.0
gy(14)	64.3	23.1
p(15)	82.2	46.2
g(275)	92.2	79.5

いくつかの音素について提案手法の精度が向上していることも確認できたが，誤差の範囲だと考えられる．上記で示した音素については，大きな差が生じている．また，音素「g」の学習データは 275 であり，一概に学習データ数が影響しているとは言えない結果となった.

原因として，今回の実験環境では学習データが足りていないため，位相情報の有無に関わらず，特徴量の情報を最大限に活用できていないと考えられる．一方，今回提案した特

徴量が，音声の特徴を表す情報として効果がないとも考えられる．そのため，学習データを増加させるだけでは，提案手法の有効性を示すことができない可能性がある．当面の課題としては，より多い学習データに対して実験を行い，提案した特徴量の有効性を検討したい．

6　おわりに

本研究では，位相情報を含む特徴量を提案し，音声認識の実験を行った．実験結果より，提案した特徴量は従来の特徴量 (MFCC，LFCC，FFTpower) の結果に及ばず，精度向上という結果は得られなかった．原因の一つとして，今回使用した実験環境では，学習のデータ量が足りないため，提案した特徴量の効果を十分に得られなかった可能性がある．今回の実験では，本研究で提案した特徴量は位相情報を含んでいるが，認識精度には影響しないという結果となった．今後の課題としては，より多いデータを使用して実験を行い，提案した特徴量の有効性を検討したい．

謝辞

最後に, 一年間に渡って御指導, 御教授して頂きました鳥取大学工学部知能情報工学科計算機 C 研究室の村田真樹教授に深くお礼申し上げます．また，本研究を遂行するにあたり日頃より暖かい御指導を賜りました村上仁一准教授に謹んで感謝の意を表します．さらに，御指導，御助言，御討論を頂きました徳久雅人講師に心から感謝致します．本研究に関して有益な御意見を頂いた計算機工学講座 C 研究室の皆様にも厚くお礼申し上げます．加えて，本稿を執筆するにあたり参考にさせて頂いた論文，本の著者の方々の皆様にもお礼申し上げます.

参考文献

[1] Ralf Schluter, Hermann Ney: “Using phase spectrum information for improved speech recognition performance”, Acoustics, Speech, and Signal Processing, 133-136, 2001.

[2] 大塚真司, 王龍標, 中川聖一: “話者認識における位相情報の改善”, 日本音響学会, 講演論文集 3-Q-2 213-214, 2007.

[3] 中川聖一: “確率モデルによる音声認識”, 社団法人 電子情報通信学会, 1988.

[4] 古井 貞熙: “音声情報処理”, 森北出版株式会社, 1998.

[5] Iosif Mporas, Todor Ganchev, Mihalis Siafarikas, Nikos Fakotakis: “Comparison of Speech Features on the Speech Recognition Task”, Journal of Computer Science 3 (8), 608-616, 2007.

[6] 西村義隆, 篠崎隆宏, 岩野公司, 古井貞熙: “重みつきスペクトル特徴量を用いた雑音に頑強な音声認識”, 日本音響学会, 講演論文集 1-6-3 5-6, 2003.

[7] Steve Young, et al.: HTK Ver3.2.1 Reference manual, Cambridge University, 2003.

付録

1. 本研究で使用したスクリプトファイル

2. 単語音声認識の実験結果
(各話者の単語認識率，誤って認識された単語一覧，音素認識率)