音声の周波数帯域と話者認識率との関係

松村　寿枝

The Relation between Frequency Band and Speech Recognition Rate

Toshie MATSUMURA

本研究では，周波数帯域を制限した音声を用いた話者認識を行い，認識率を比較することで音声の周波数帯域と話者認識率の関係を明らかにすることを目的にした．男性話者10名が10回発声した孤立発声５母音をLow Pass Filter(以下LPF)，High Pass Filter(以下HPF)，Band Pass Filter(以下BPF)を通し周波数帯域を制限した音声を作成し，Hidden Markov Model(以下HMM)法により話者認識実験を行った．結果，周波数帯域が広がるにつれて認識率は向上するが，周波数帯域が2倍になっても認識率にはあまり影響与えていないことがわかった．また，1000Hz以下の第1，2フォルマント周波数を含む周波数帯域が話者認識率に影響を与えており，音声が1000Hz以上の周波数帯域を多く含んでいても認識率にはあまり影響しないことがわかった．

1．はじめに

人の最も身近な情報伝達の手段の１つである音声には，意味内容を示す音韻性情報，話者を示す個人性情報，話者の感情を示す情緒性情報など様々な情報が含まれている．その中でも音韻性情報を用いた認識が音声認識，個人性情報を用いた認識が話者認識であり，音声認識を用いたシステムは既に実用化されている．しかし，話者認識は未だ研究段階であり，話者認識システムはほとんど実用化されていない．これは，話者認識がセキュリティなどのような分野で利用されることを考えると，かなり高い認識率を必要とするためであると考えられる．しかし，音声には揺らぎが存在し，異なる話者のみでなく，同一話者が同じ言葉を話してもその発声時期等により異なることがあり，このような揺らぎが話者認識を更に難しくしている．一方，音声認識は，多少誤認識があっても言い直しやキーボードによる入力の修正が可能であるという違いがあり，この点が既にシステムが実用化されている理由であると考えられる．

さて，従来の話者認識の研究では音声の低域部分のみでなく，高域部分(この場合の“高域”とは電話で使用されている周波数帯域以上の周波数帯域を指す)にも個人性情報が含まれおり，音声の高域に存在する個人性情報の話者認識への貢献度についての研究が報告されている[1]．しかし，この研究は単語音声を話者認識に用いており，母音のどの周波数帯域に個人性情報が多く含まれているか明らかにしていない．また，認識法として，マルチテンプレート法を使用している．そのため，近年音声認識や話者認識の分野で使用されているHMM法[2], [3]による話者認識ではない．また，別の研究では，HMMの１つである連結音韻モデルを使用した話者認識[4], [5]が行われているが，音声の周波数帯域による認識率の違いは述べられていない．話者認識の認識率向上のためには，どの周波数帯域に個人性情報が多く含まれているかを知る必要がある．言い換えると話者認識率と周波数帯域の関係がわかれば，認識率の向上につながるといえる．そこで，本研究では，LPF，HPF，BPFを用い周波数帯域を制限した孤立発声母音の話者認識率の違いをHMM法を用い調べることとした．周波数帯域を制限した音声の話者認識率を調べることで，定量的に抽出することが困難な個人性情報がどの周波数帯域に多く含まれるかを明らかにすることが出来る．

以下，２章では，HMM法を中心に話者認識法について述べ，３章では話者認識実験手順をはじめとする話者認識実験について述べ，４章では話者認識実験結果，５章では考察，６章ではまとめを述べる．

2. 話者認識法

音声中に含まれる個人性情報を抽出し，話者が誰であるかを認識することを話者認識という．しかし，音声中には個人性情報のみでなく音韻性情報や情緒性情報などが複雑に絡み合った形で存在するため個人性情報のみを抽出することは困難である．

更に，話者認識を難しくしている要因のひとつに音声の揺らぎがある．音声の揺らぎは，異なる話者が同じ言葉を発した際に話し方の違いや発声器官の違いのため，各々の音声に音響的性質の差が生じる“個人差”と呼ばれるもののほかに，同一話者が同じ言葉を発声した場合にも生じる．特に話者認識が音声認識ほど一般的になっていない理由としては，話者認識がセキュリティなどの分野に利用されることを考えると，音声認識の認識率に比べ，話者認識では，かなり高い認識率を必要とするためではないかと考えられる．そこで，話者認識率向上のためには，どの周波数帯域に個人性情報が多く含まれているかを知る必要があると考えた．言い換えれば，周波数帯域と話者認識率の関係がわかれば，認識率のよい周波数帯域のみを話者認識に用いることで認識率の向上につながると考えられる．

話者認識法は現在種々のものが提案されているが，本研究では，近年音声認識や話者認識の分野で研究が進み実用化が進んでいるHMM法を用いた． HMM法の詳細については，2.1節で詳細を述べる．

2.1　HMM法

音声生成の過程をマルコフ過程としてモデル化した手法がMarkov Model法である．Markov Modelは，N個の状態S_1，S_2，…，S_i，…，S_Nを持ち一定周期毎に状態を遷移し，その遷移の際に1つずつ音声の特徴パラメータ系列を出力する．それぞれのHMMは，遷移確率と出力確率によって確率的に決められている．この出力は観測できるが，状態S_iの遷移は観測できないため，“Hidden”と呼ばれている．また，出力が有限個の記号(ラベル)の場合を“離散分布HMM”といい，出力が多次元正規分布などの連続的な確率分布に従う時“連続分布HMM”という．本研究で使用したHMMはこの連続分布HMMを使用した．

HMM法は，

・実際の音声の揺らぎを正確に反映しやすい．
・統計理論や情報理論による理論的発展がしやすい．
・確率の概念を用いて言語処理等を統合しやすい．
・DPマッチングを特殊な場合として含み拡張されている．

などの特徴から近年着目され，音声認識や話者認識に応用されている認識方法である[3]．

本研究では，以上のような音声の揺らぎを反映しやすいという観点から，話者認識法としてHMM法を用いた．

次に使用したHMMのモデルについて説明する． 本研究では，学習時に各話者の音韻クラスHMMを作成する．作成した各話者の音韻モデルを以下では，“話者音韻HMM”と呼ぶこととする．話者音韻HMMは，各話者別に作成し，1つのHMMは3状態($N=3$)で1ガウス分布の連続分布HMMとする．また，モデルは，音声の不可逆性を考慮しleft to right 型のBakisモデルとした．

HMM法を用いた話者認識法は，3.1節の話者認識実験の手順でも詳細を述べるが，学習と認識の部分に分けられる．HMMの学習時に各単語を数回発声し，その発声をモデル化したHMMを作成する．認識時には入力系列を出力する確率(尤度)が最大になるHMMを探し，認識結果とする方法である．

3．話者認識実験

3.1　話者認識実験手順

実験手順を図1に示す．

第1に実験に使用する音声資料を収集した．音声資料については3.2節で更に詳細を述べる．

第2に音声分析を行った．音声は，帯域制限をしていない原音声をLPF，HPF，BPFのそれぞれのフィルタに通し周波数帯域を制限した音声を作成した．また，比較のため，原音声も話者認識実験に使用した．次にLPCケプストラム分析を行い，特徴パラメータを求めた．音声分析の仕様については3.3節で述べる．

第3にHMMの学習を行い，話者音韻HMMを作成した．

第4に作成した話者音韻HMMと特徴パラメータを比較し，モデルの評価，尤度計算を行った．但し，学習には10回収集した音声のうち3回分を使用し，話者認識実験には，学習に使用した3回分を含む10回分全てを認識用として用いた．

フィルタの遮断周波数は1000Hzから5000Hzまで1000Hz刻みで変化させていった．また，BPFにおいても同様に遮断周波数を1000Hz刻みとし，通過域の幅が1000Hzとなるように設定した．

3.2　音声資料

成人男性10名が発声した孤立発声5母音（/ａ/，/ｉ/，/ｕ/，/ｅ/，/ｏ/）それぞれ10回をSONY製マイクロフォン（ECM-330）を用い収集した．収集した音声は，11025Hzでサンプリングし，16bit/sampleで量子化し

た．収録した音声は，目視により音声区間の切り出しを行った．音声は時期による変動(経時変化)を伴うが，今回は同一時期に収集した音声を用いた．

3.3　音声分析

フィルタを通していない原音声とLPF，HPF，BPFを通し帯域制限した音声に対し音声分析を行った．LPF，HPF，BPFは，フーリエ級数法によるディジタルフィルタ(FIR型)で，タップ数11025とし作成したプログラムを使用した．音声分析には，LPCケプストラム分析[6]を使用した．音声分析仕様を表1に示す．音声は非定常信号であるが，短い区間であれば定常信号とみなすことが出来る．そのためフレーム長は経験的に15msから30ms，フレーム更新周期は5msないし10msにすることが多い．これを踏まえてフレーム長256点(約23ms)，フレーム更新周期88点（約8ms）に設定した．LPCケプストラム分析により得られた16次の係数を特徴パラメータとし，この特徴パラメータをHMM法の学習に用いた．

4．話者認識実験結果

図2にLPFの遮断周波数と話者認識率の関係を示す．以下の図の“原音声”は，フィルタを通していない5母音全ての話者認識率を平均したものである．

図2よりLPFでは遮断周波数が上がるにつれて認識率が向上している．これは，遮断周波数が上がるにつれて，音声の周波数帯域が広がったため含まれる個人性情報の量が増え，話者認識率が向上したと考えられる．

しかし，1000Hzと2000Hz，2000Hzと4000Hzを比較すると周波数帯域は2倍に広がっているが，認識率はそれほど急激に向上していない．また，遮断周波数1000Hz，2000Hz，3000Hz，4000Hz，5000Hzとそれぞれの認識率の差を比較すると，1000Hzと2000Hzの間で認識率の差が4.8%と他の周波数間よりも大きいことがわかる．このことから，1000Hzと2000Hzの間で最もよく認識率が向上したといえる．以上の結果から，

・周波数帯域が広がれば認識率は向上するが，単純に周波数帯域が2倍に広がっても認識率は急激には向上はしない．

・1000Hzと2000Hzの間で認識率が最もよく向上した．

ことがわかった．

但し，4000Hz付近で認識率が低下していることに関しては本節後半で詳細を述べる．

図3にHPFの遮断周波数と話者認識率の関係を示す．図3より1000Hzから4000Hzと遮断周波数を上げるにつれてHPFでは認識率は徐々に低下していくことがわかる．これは，周波数帯域が狭まるにつれて，含まれる個人性情報の量が減少し，話者認識率が低下していくことを示しており，LPFの結果と一致する．

但し，図3からもわかるように5000Hzで認識率が向上している．この点に関してはLPFと同様に本節後半で詳細を述べる．

図4にBPFの遮断周波数と話者認識率の関係を示す．尚，図4において，例えば“1000－2000”とは，“1000Hzから2000Hzまでの周波数通すBPFを使用した”ことを示す．

図4のBPFの結果から1000Hz以上では遮断周波数の違いによる話者認識率はほとんど差がないことがわかる．

LPF，HPF，BPFの結果をまとめると，周波数帯域が広がるにつれて個人情報を多く含むようになるため認識率は向上するが，周波数帯域が2倍になってもあまり認識率は向上しないといえる．

図5に遮断周波数と母音別話者認識率の関係（LPF）を示す．周波数を上げるにつれて認識率が向上するのはかわらないが，話者認識率には母音毎に違いがあることがわかった．特に，/a/において4000Hzでの認識率の低下が大きかったため，全体の認識率(図2)に影響を及ぼしたものと考えられる．しかし，この4000Hzで認識率が急に低下した原因についてはっきりと特定できなかった．

次に，図6に遮断周波数と母音別話者認識率の関係（HPF）を示す．LPFの時と同様に話者認識率には母音毎に違いがあることがわかった．特に/a/,/o/を除き，遮断周波数が5000Hzで一様に認識率が向上していることがわかった．これが全体の認識率（図3)に影響したものと考える．しかし，5000Hzで認識率が向上した原因については，はっきりと特定できなかった．

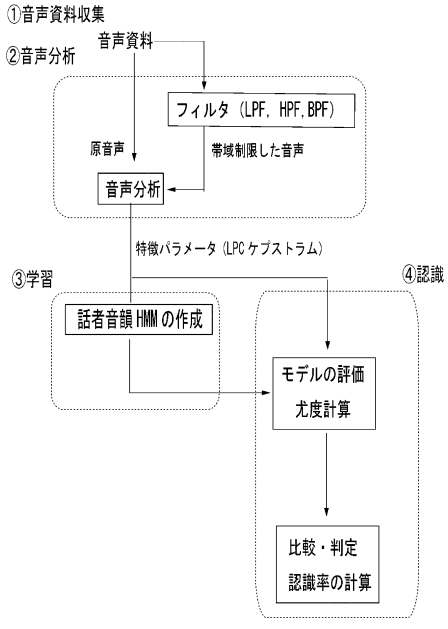

図1　話者認識実験手順

図７に遮断周波数と母音別話者認識率の関係（BPF）を示す．やはり，BPFにおいても母音によりかなり話者認識率に違いがあることがわかった．

図８に母音(原音声)と話者認識率の関係を示す．図８より話者認識率は，母音によって異なり，/i/と/e/は認識率がよく，逆に/o/では認識率が他の母音に比べて低いことがわかる．これは，LPF，HPF，BPFにおいても同様で，/i/と/e/の認識率は比較的高く，/o/は比較的認識率が低い傾向にあった．

5. 考　察

LPFにおいて遮断周波数が1000Hz以上では認識率は大きく向上しない原因について考察した．予備実験からフォルマント周波数と呼ばれる音韻を特徴づける周波数は，個人性情報も関係が有ると考えていた．フォルマント周波数は，母音により異なるが，一般には第１フォルマント周波数で1.4kHz以下，第２フォルマントで4.0kHz以下の値とされる．但し，このフォルマント周波数は，同じ音韻でも発声者により大幅に変化するとされている．また，連続発声と孤立発声によっても多少異なると考えられる．そこで，追加実験として第1，2フォルマント周波数を求めた結果，本実験で使用した母音の第1，2フォルマント周波数には1000Hz以下が多く，そのため1000Hz以上の周波数帯域では認識率にほとんど影響が出なかったと考えられる．言い換えると，母音の第1，2フォルマント周波数を含む周波数帯域が個人性情報を多く含み，話者認識率にも影響を与えている．そのため，話者認識には第1，2フォルマント周波数帯域を含むような音声を用いることが有効であり，それ以上の周波数帯域を含んでいても認識率はあまり変化しないと考えられた．今後は，遮断周波数を1000Hz以下の第1，2フォルマント周波数に設定し実験を行い，検証することが必要である．

表1　音声分析（LPCケプストラム分析）仕様

項目	仕様
フレーム長	256点
フレーム更新周期（インターバル）	88点
窓関数	ハミング窓
分析次数	16次

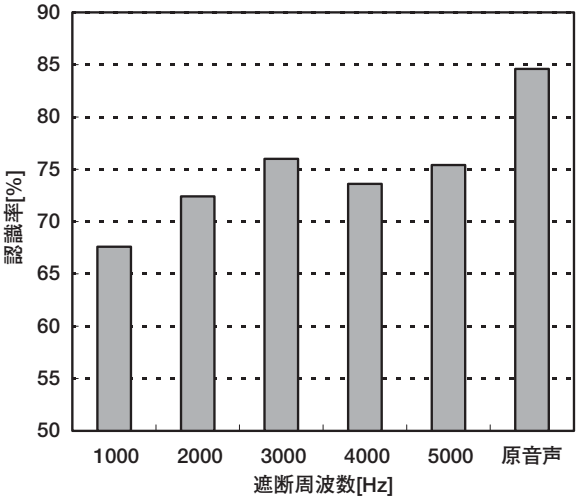

図2　遮断周波数と話者認識率の関係(LPF)

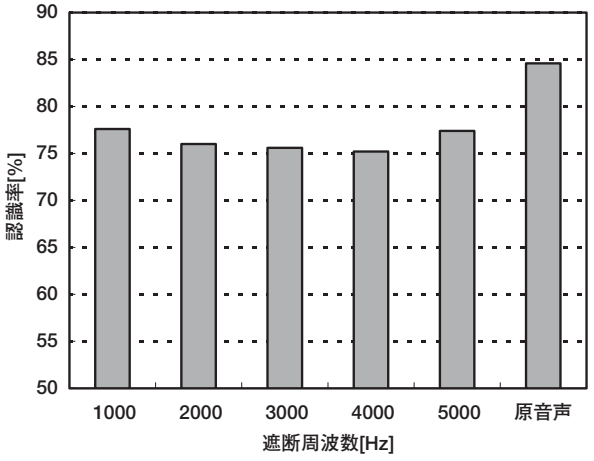

図3　遮断周波数と話者認識率の関係(HPF)

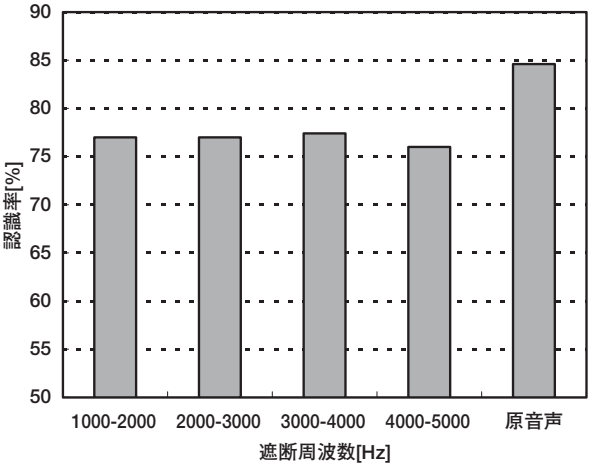

図4　遮断周波数と話者認識率の関係(BPF)

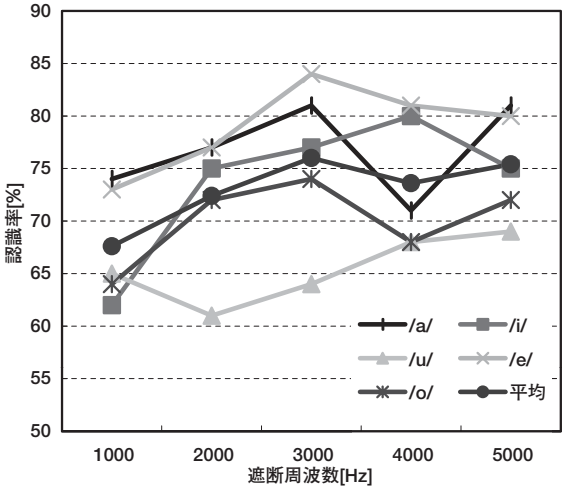

図5　遮断周波数と母音別話者認識率の関係(LPF)

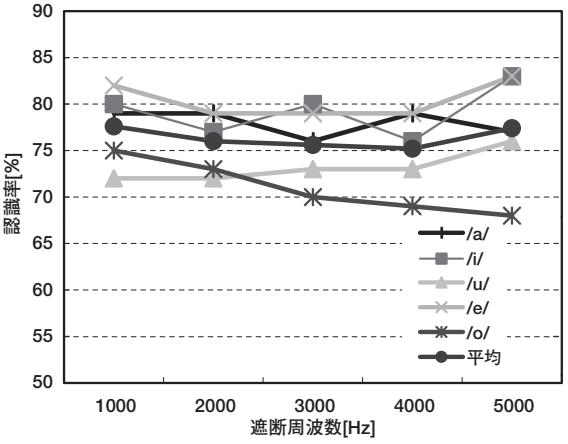

図6　遮断周波数と母音別話者認識率の関係(HPF)

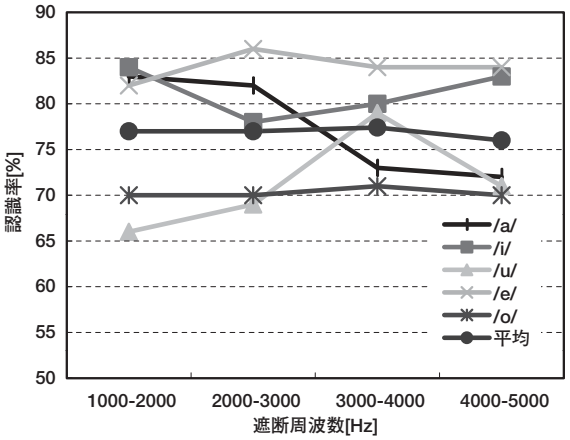

図7　遮断周波数と母音別話者認識率の関係(BPF)

6. おわりに

本研究では，音声の周波数帯域と話者認識率の関係をHMM法を用いた話者認識実験により明らかにした．結果，周波数帯域が広がるにつれて認識率は向上するが，遮断周波数が1000Hz以上では認識率はあまり向上しないことがわかった．また，音声の第1，2フォルマント周波数を含む帯域を用いることが話者認識の認識率向上には有効であることが示された．

今後の課題として次のような音声資料による実験が必要である．

・連続発声母音及び子音を含めた連続発声音声での実験

・女性話者の音声による実験

更に今回の音声資料収集は，多くの計算機が置かれた，学生が多く出入りする実験室で行われた．そのため，雑音環境下での音声収集にも問題があったと考えられる．認識率の低下には雑音の影響も大きいといわれ，特に高域はパワーが少ないため雑音による影響を受けやすいと考えられる．今回のLPFの遮断周波数4000Hzで認識率が低下した原因とHPFの遮断周波数5000Hzで認識率が向上した原因が雑音によるものかどうかはわからないが，今後は音声資料収集の場所と方法についても検討していきたい．また，フォルマント周波数の詳細な抽出を検討に入れ，追加実験が必要である．

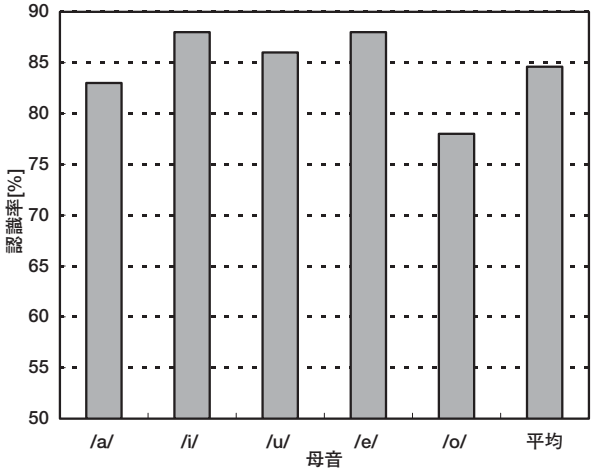

図8　母音(原音声)と話者認識率の関係

謝　辞

本研究にあたり，音声資料収集，実験でご協力いただいた平成16年度卒業生橋口幸治君，平成17年度卒業生川﨑望さんに深く感謝いたします．

また，音声資料収集にご協力いただいた奈良工業高等専門学校学生の皆様に深く感謝いたします．

参考文献

［1］早川昭二，板倉文忠，“音声の高域に含まれる個人性情報を用いた話者認識”，日本音響学会誌，51，11，pp．861-868，(1995)

［2］中川聖一，“確率モデルによる音声認識”，電子情報通信学会，pp．29－89，(1988)

［3］大河内正明，“Hidden Markov Modelに基づいた音声認識”，日本音響学会誌，42，12，pp．936-941，(1986)

［4］松井知子，古井貞熙，“連結音韻HMMによる話者認識”，電子情報通信学会秋季大会講論集，SA-6-2，(1992)

［5］松井知子，古井貞熙，“連結音韻モデルによる話者認識”，信学技報，SP92-128，pp．41-47，(1993)

［6］中川聖一，坂本光範，“音声認識・話者認識のためのFFTケプストラムおよびLPCケプストラムの検討”，電子情報通信学会論文誌A，Vol.J66-A，No. 12，pp．1199-1206，(1988)