連続音声中の母音による男声女声判別

Male or Female Detection Using Vowel Feature Parameters in Continuous Speech

古市　千枝子　大貫　雅之

桐蔭横浜大学工学部電子情報工学科

(2007年3月1日　受理)

1　まえがき

人の話し声には、主に、何を話しているかという言語情報と誰の声であるかという話者情報が含まれている。音声情報処理の分野では前者の言語情報を抽出することを音声認識、後者の話者情報を抽出することを話者認識と呼ぶ。音声認識や話者認識の方法としては、音声の音響的・音韻的特徴が効率よくしかもかなり正確に表現された音響モデルを基にして、未知音声に対するゆう度を最大化する方法が一般的に用いられている。

ところで最近、個人の情報を守り、被害を防止する技術の要請が高まり、個人の身体的特徴によって本人を確認する方法がいくつか開発されつつある。暗証番号やパスワードなどに比べ、原理的に極めて「なりすまし」がし難いため本人認証の方式として関心が集まっている[1]。

本稿では人の声による本人確認システムの前処理として用いるために、未知音声を先ず男声と女声に大分類する手法を提案し、その有効性を検討することを目的とする。提案法によって得られた結果は本人確認システムだけでなく、音声認識システムの前処理として用いることによって、音響モデルの高精度化にも適用可能である。

話者の特徴は一般に子音より母音の方に多く含まれていることが知られている。そこで提案法では、すでに開発した音素セグメンテーションシステムによって得られる連続音声中の母音部から自動的に抽出した特徴パラメータを用いる。従って発声内容に依存しないより汎用性の高いテキスト独立型の男声女声判別システムの構築が可能である。さらに本提案法では話者認識系と同じ音響分析法を用いているので男声女声判別のためのパラメータをあらたに抽出する必要がなく話者認識システムの前処理として適している。

2　男声女声判別システム

図1に男声女声判別システムの構成を示す。以下では各部の詳細について述べる。

2.1　音素セグメンテーション

音素セグメンテーションのための特徴パラメータとして、不偏メルケプストラム[2]と零交差数から抽出した3種の静的なパラメータ(有声音検出パラメータ、零次メルケプストラム係数、零交差数)と3種の動的なパラメータ（零次メルケプストラム時間変化パラメータ、零交差数時間変化パラメータ、スペクト

Chieko Furuichi and Masayuki Oonuki ; Dept. of Electronics and Information Engineering, Faculty of Engineering, Toin University of Yokohama, 1614, Kurogane-cho Aoba-ku, Yokohama, 225-8502

ル包絡時間変化パラメータ）を用いる[3][4][5]。

先ず音声信号を有声音検出パラメータによって有声部と非有声部に分け、それぞれの区間において最適な特徴パラメータとアルゴリズムを用いて音素境界を推定する。さらに検出したセグメントに対して、有声部のセグメントには、母音（W）、有声子音（C）、いずれにも決められない有声音（V）の3種類、非有声部のセグメントには、摩擦性（F）、非摩擦性（U）、無音（S）の3種類の音韻性を表す記号でラベリングを行う。

本システムでは母音（W）セグメントから母音の特徴パラメータを抽出する。

2.2　母音の特徴パラメータの抽出

話者の識別に有効な情報は、母音では一般的に音響的特性が安定しているスペクトルの定常部分にある。連続音声中の母音音素は単独発声された単母音より、前後の音素との調音結合の影響をうけてスペクトルが変動する。そのため母音（W）セグメントのスペクトル包絡の定常部はスペクトル包絡の時間変化量 d_i を利用して検出する。この d_i はメルケプストラムの低次の項を用いて

$$d_i=\left(\sum_{m=1}^{2}(d_i^m)^2\right)^{1/2}$$

$$d_i^m=K_4\sum_{n=-4}^{4}w_n n g_{i+n}^m,\quad K_4=\left(\sum_{n=-4}^{4}w_n n^2\right)^{-1} \qquad (1)$$

で与えられる。ここで g_i^m は i 番目の分析フレームの m 次の不偏メルケプストラム係数[2]を表す。w_n は偶関数の窓関数で対象とする範囲の速さの時間変化を抽出して、不要な細かな時間変化を減衰させる働きをもつ[5]。ここでは長さ9のブラックマン窓により約3Hzから20Hzの周波数に相当するスペクトル包絡の時間変化を検出している。

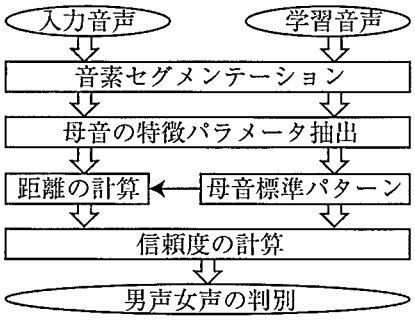


図1　男声女声判別システムの流れ図

Wセグメントにおいて d_i の値が極小となるフレームを中心として、1分析フレームおきにとった前後5フレーム分の1次から12次までの不偏メルケプストラム係数[2]を母音の特徴パラメータとする。

2.3　スペクトル距離の計算

入力音声から母音の特徴パラメータを抽出するまでの処理は、学習過程で標準パターンを作成する操作と全く同じ方法で行われる。

入力パターンTと標準パターンRの特徴パラメータは2.2で述べたようにそれぞれ5個のフレームのパラメータセットで表される。1フレーム分の特徴パラメータは1次から12次までの不偏メルケプストラム係数で構成され、それぞれ t_i^m, r_i^m $(m=1, 2,\cdots 12)$ で表す。ここで i はフレーム番号、m はメルケプストラムの次数である。入力パターンTと標準パターンRの i 番目のフレーム間のスペクトル距離 q_n はメルケプストラムのユークリッド距離で定義し、次式で計算する[6]。

$$q_n=\left(\sum_{m=1}^{12}(t_i^m-r_i^m)^2\right)^{1/2}*4.343\quad [dB] \qquad (2)$$

上式では q_n の単位がdBで表されるように定数を乗じている。

入力パターンTと標準パターンRのスペクトル距離Dは次式で計算する。

$$D=\frac{q_{11}+2q_{22}+2q_{33}+2q_{44}+q_{55}}{8} \qquad (3)$$

パターン間の距離計算は特徴パラメータの時系列の区間両端の切断の影響を小さくするために、中央部のフレームに重みをかけてパターンの距離を求めている。

2.4　男声女声判別のアルゴリズム

入力パターンと全ての標準パターンとのスペクトル距離を2.3の定義に従って計算し、距離の近い順に並べて D_1, D_2,…,D_N で表す。

ここで、D_iは入力パターンとの距離がi番目に近い距離の値である。Nは入力パターンとの距離が一番近いD_1から1 dB以内にある標準パターンの個数を表す。

入力パターンと標準パターンのスペクトル距離に対するスコアを次式で定義する。

$$S_i=\frac{1}{(D_i-D_1+1)},\quad (i=1,2,\cdots,N) \qquad (4)$$

式（4）から距離が最も近いD_iに対するスコアは1となり、距離が大きくなるに従ってD_iに対するスコアS_iの値は小さくなる。

次に、同一話者に対するスコアをS_j（$j\in\{1,\ 2,\ \cdots,\ N\}$）としたとき、その話者の信頼度を次式で定義する。

$$P_k=\left(\sum_j S_j\right)\Big/\left(\sum_{i=1}^{N} S_i\right),\ (k=1,\ 2,\ \cdots,\ K) \qquad (5)$$

$$\sum_{k=1}^{K} P_k=1 \qquad (6)$$

式（5）の分子は同一話者のスコアを加算した値となっている。KはN個の標準パターンに含まれる話者の異なりの個数を表す。N個のスコアの話者がすべて同じ話者の場合は$K=1$となり、式（5）の分子と分母は等しいので$P_k=1$（$k=1$）となる。

式（5）、（6）から信頼度P_kの最大値は1で$P_k(1\leq k\leq K)$の全加算値は1であるから、P_kの値が1に近いほどその話者は入力パターンの有力候補となる。従って、P_kの値は候補話者の信頼性を表す尺度と考えることができる。

候補話者を男声と女声に分け、それぞれの信頼度P_kを合計した値を男声女声の判別の信頼性を表す尺度として用いる。

表1　音声資料

話者数	男声　20名（m01～m20） 女声　13名（f01～f13）
テキスト	4桁数字からなる10単語
学習話者数	男声　5名、女声　5名
学習音声	10単語
入力音声	1単語または3単語

3　評価実験

3.1　音声資料と分析条件

評価実験では、表1で示した話者認識用データベースを用いた。テキストとして4桁のランダムな数字を連続発声したものを1単語とした10種類の単語音声を使用した。実験に用いた話者数は男声20名（m01～m20）、女声13名（f01～f13）である。学習話者としては入力話者とは異なる男女各5名を選び、10個の単語音声から抽出した母音の特徴パラメータを標準パターンとした。

表2に音声の分析条件を示す。

3.2　実験

図2は入力話者m01が単語「2934」(/nikyusaNyoN/)を発声したときの学習話者男女各5名（m11～m15、f06～f10）の信頼度を示したものである。横軸は学習話者番号、縦軸は各学習話者の信頼度を表す。この場合、男声信頼度の合計は0.591、女声信頼度の合計は0.409となり入力話者が男声である信頼性は0.591である。

表2　音声の分析条件

サンプリング周波数	10kHz
量子化数	12bit
分析フレーム長	25.6 msec ブラックマン窓
フレーム周期	10.0 msec
特徴パラメータ	12次不偏メルケプストラム
フレーム数	5フレーム
ケプストラムの抽出	不偏推定法（加速定数1.0、繰り返し3回、次数20次
メルケプストラムの抽出	不偏推定法（加速定数1.0、繰り返し2回、次数12次）

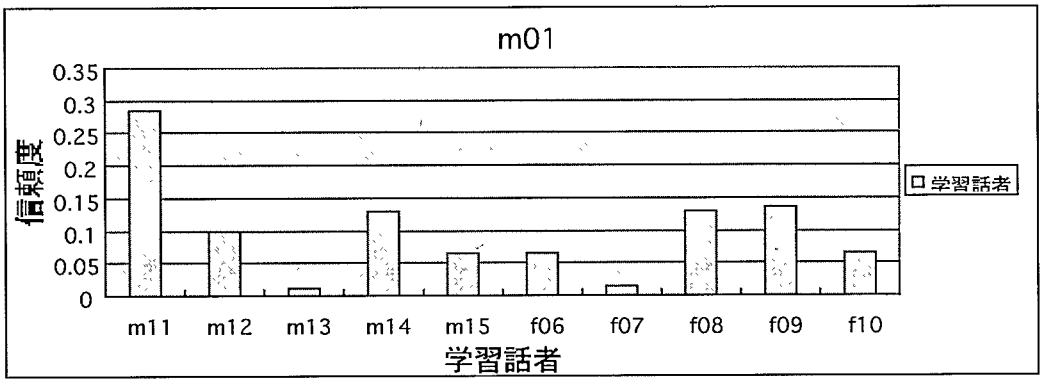


図2　入力話者 m01 に対する学習話者の信頼度：発声単語「2934」(/nikyusaNyoN/)

3.2.1　入力単語の影響

先ず男性話者 10 名の入力音声について実験を行った。図 3 (a)、(b)、(c) は入力音声の単語をそれぞれ「0712」、「1823」、「2934」と替えた場合の男声女声信頼度の変化を調べたものである。横軸は入力話者番号、縦軸は各入力話者に対する男声と女声の信頼度の合計を表す。学習話者として男声 m11 ～ m15、女声 f06 ～ f10 を用いた。入力単語を替えることによって男声女声信頼度が変化するが、いずれの入力単語でも入力話者 m02、m03、m05、m06、m07、m09、m10 は男声信頼度の方が高い値を示している。一方、入力話者 m01、mo4、m08 は入力単語によっては女声信頼度の方が高い値を示す。

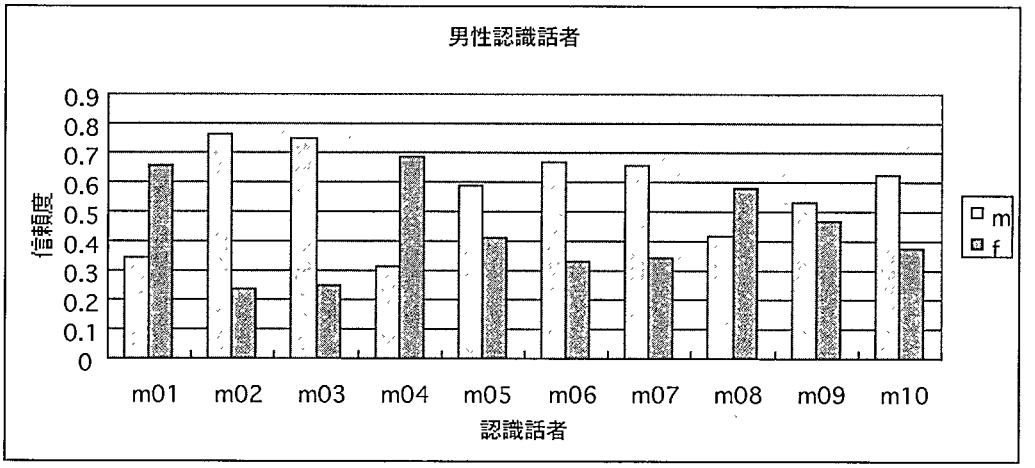
単語ごとの全話者の男声信頼度の平均値は、入力単語「0712」が 0.564、「1823」が 0.612、「2934」が 0.726 である。このことから、3 単語の中では「2934」が入力単語として最も適しているといえる。図 3 (c) では男声女声判別の難しい話者 m01、m04、m08 も男声信頼度の方が高い値を示している。

次に、女性話者 8 名の入力音声に対して実験を行った。男声に対する実験と同じ入力単語を用いた場合の結果を図 4 (a)、(b)、(c) に示す。学習話者は図 3 の実験と同一にした。単語ごとの全話者の女声信頼度の平均値は、入力単語「0712」が 1.000、「1823」が 0.964、「2934」が 0.975 である。女声の場合、入力単語に依存せず、いずれの話者も常に高い信頼度が得られることが分かった。

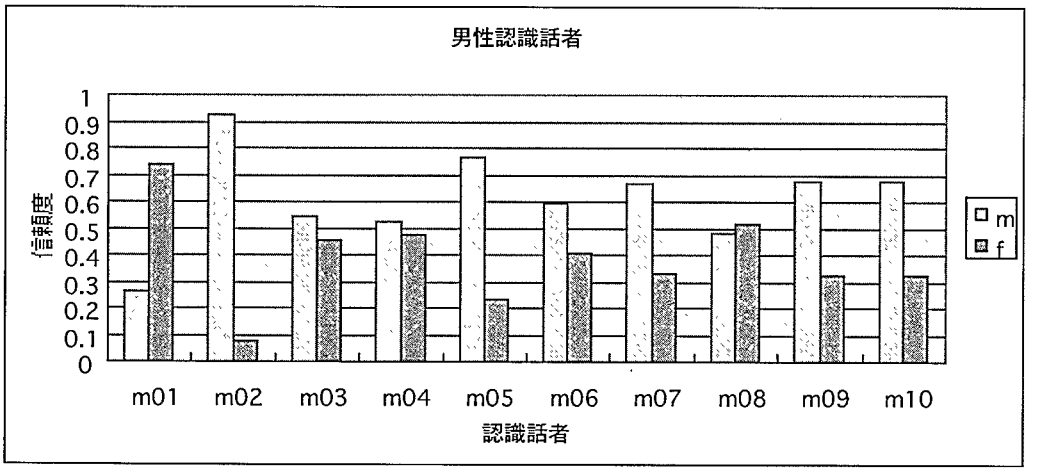
3.2.2　学習話者の影響

3.2.1 では、女声話者に対しては入力単語に依存せずに高い女声信頼度が得られることが分かったので、以降は男性話者に対する信頼度の影響を調べた。まず学習話者の選定が男声女声信頼度に及ぼす影響を調べる実験を行った。図 5 は入力単語を「2934」に固定して、学習話者を (a) では男声 m11 ～ m15、女声 f06 ～ f10、(b) では男声 m16 ～ m20、女声 f01 ～ f05、(c) では男声 m11、m12、m16、m17、m20、女声 f02、f05、f06、f07、f10 とした場合の結果である。学習話者の違いによる全話者の男声信頼度の平均値は、(a)0.726、(b)0.740、(c)0.821 である。(c) の学習話者の組合せが男声信頼度の平均値が最も高く、判別の難しい話者 m01、m04、m08 も男声信頼度の方が高い値を示している。

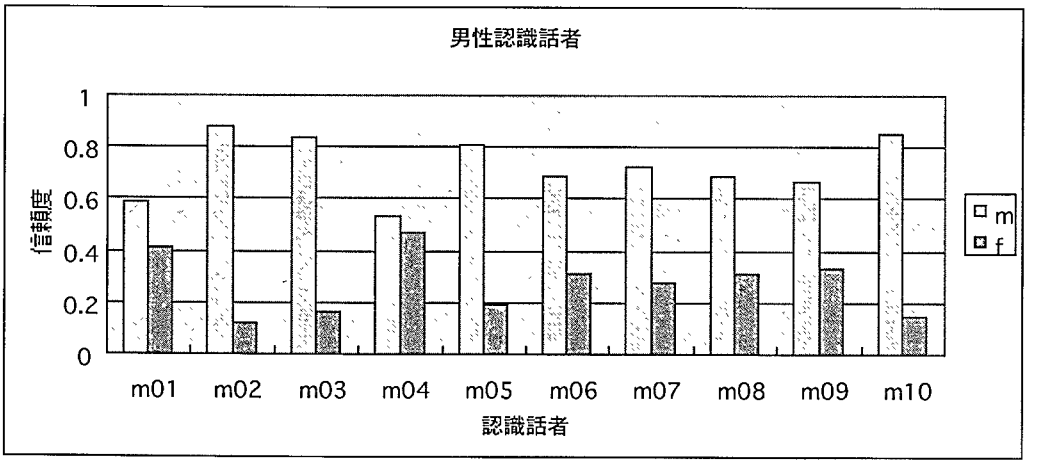
次に女声の学習話者の影響を調べた。図 6 は入力単語を「2934」として、入力話者数を 20 名にした場合の実験結果である。(a)、(b) いずれも入力話者 m01 ～ m10 と m11 ～ m20 に対する男声学習話者はそれぞれ m11 ～ m15 と m01 ～ m05 とし、女声学習話者として (a) では f06 ～ f10、(b) では f01 ～ f05 を使用した。全話者の男声信頼度の平均値は、(a)0.815、(b)0.787 となり女声の学習話者の選



（a） 入力単語「0712」（/zeronanaicini/）

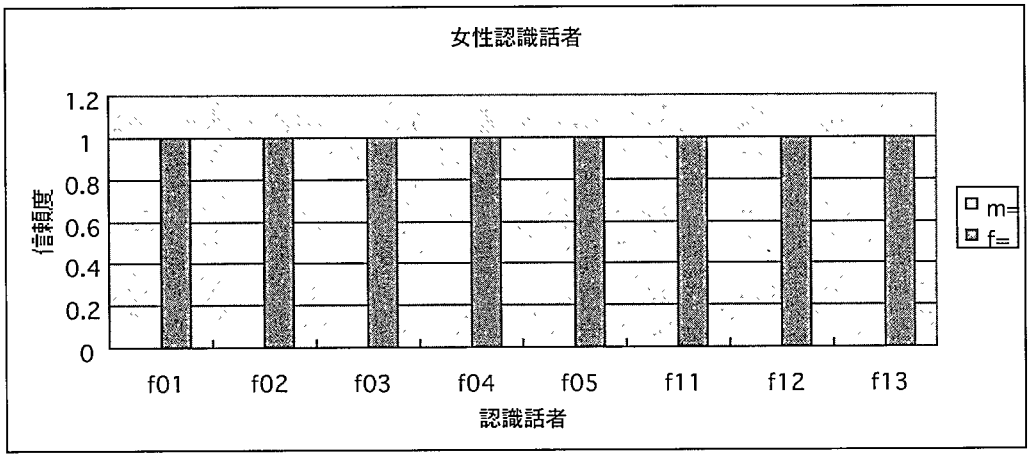


（b） 入力単語「1823」（/icihacinisaN/）

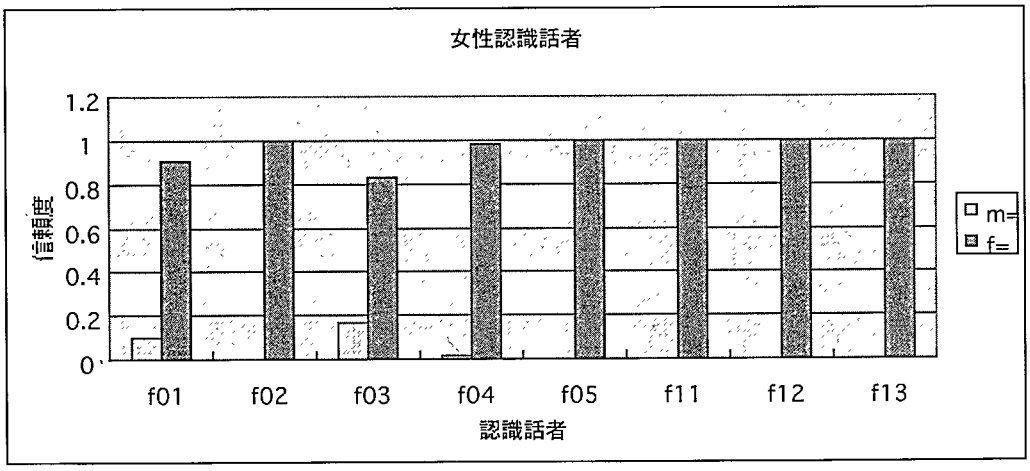


（c） 入力単語「2934」（/nikyusaNyoN/）

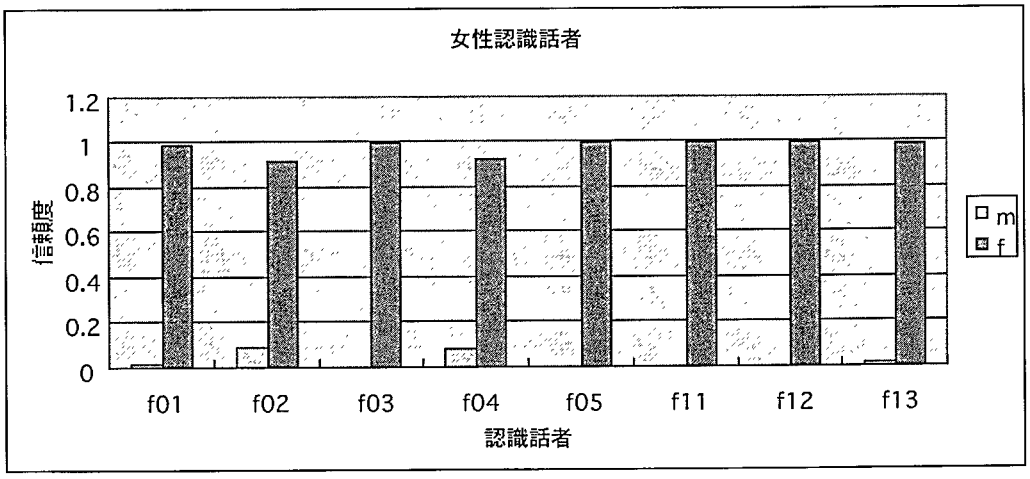
図3 入力単語による男声女声信頼度の比較（男性話者）



（a）　入力単語「0712」（/zeronanaicini/）

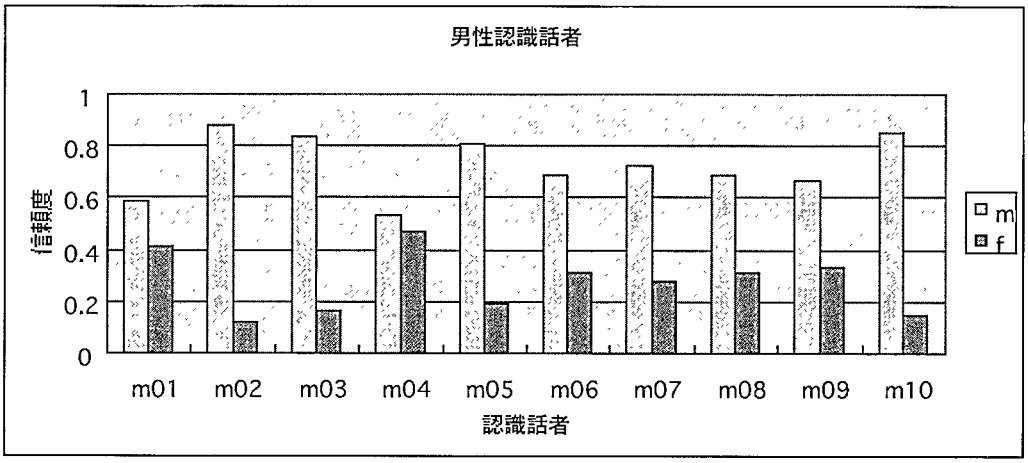


（b）　入力単語「1823」（/icihacinisaN/）

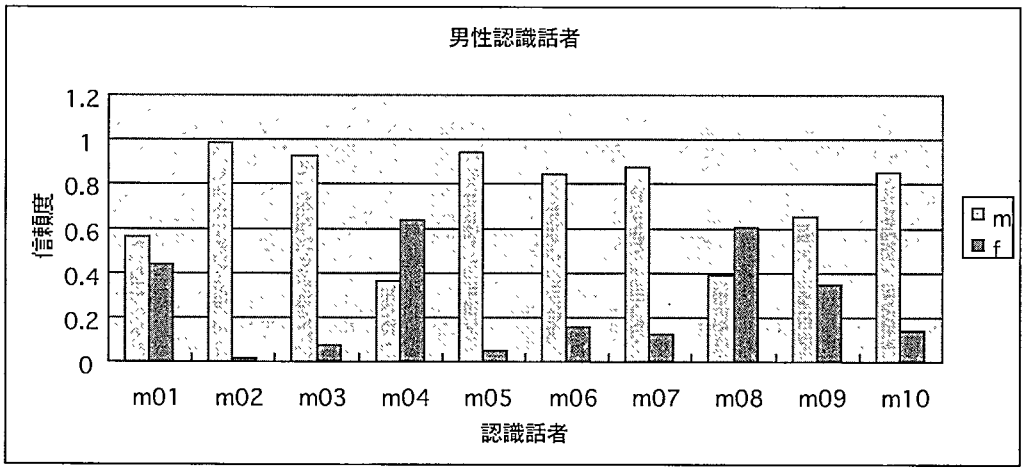


（c）　入力単語「2934」（/nikyusaNyoN/）

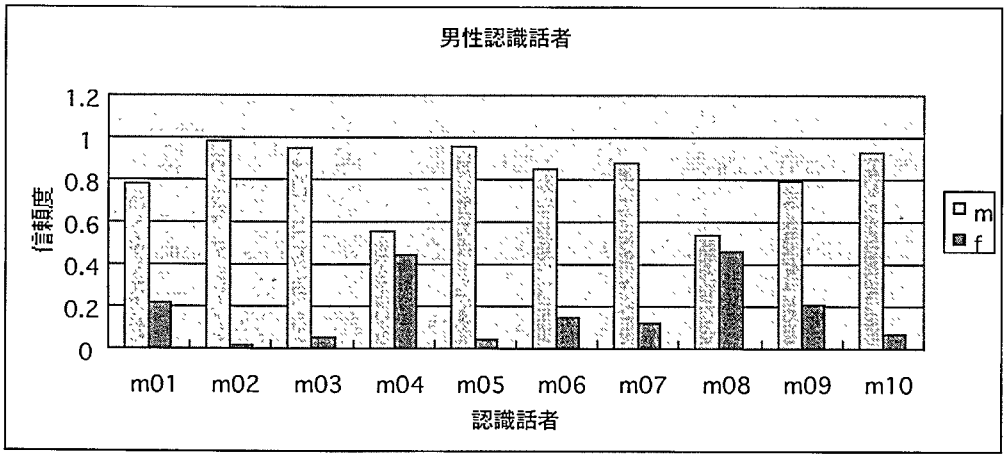
図4　入力単語による男声女声信頼度の比較（女性話者）



（a） 学習話者：男声（m11～m15）、女声（f06～f10）

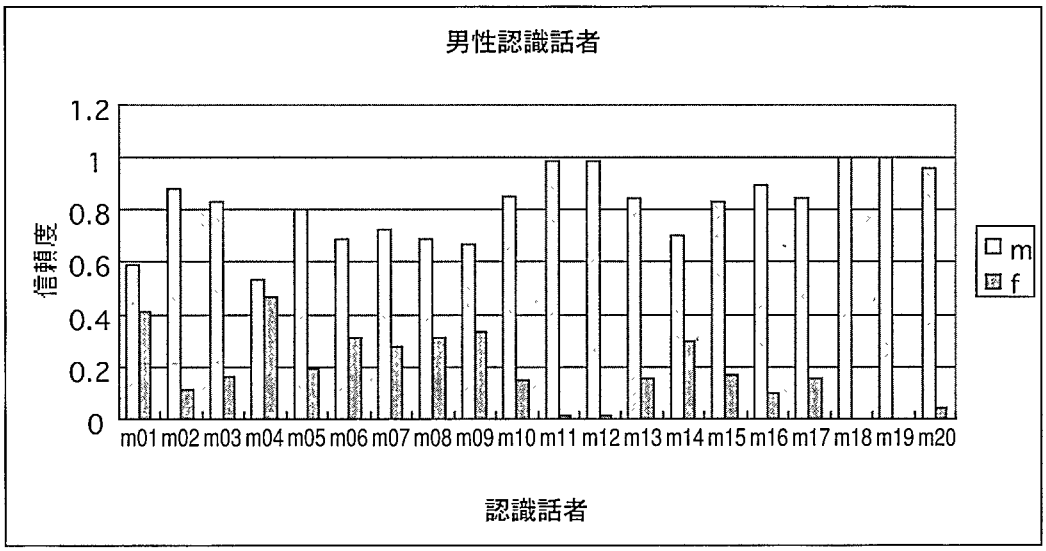


（b） 学習話者：男声（m16～m20）、女声（f01～f05）

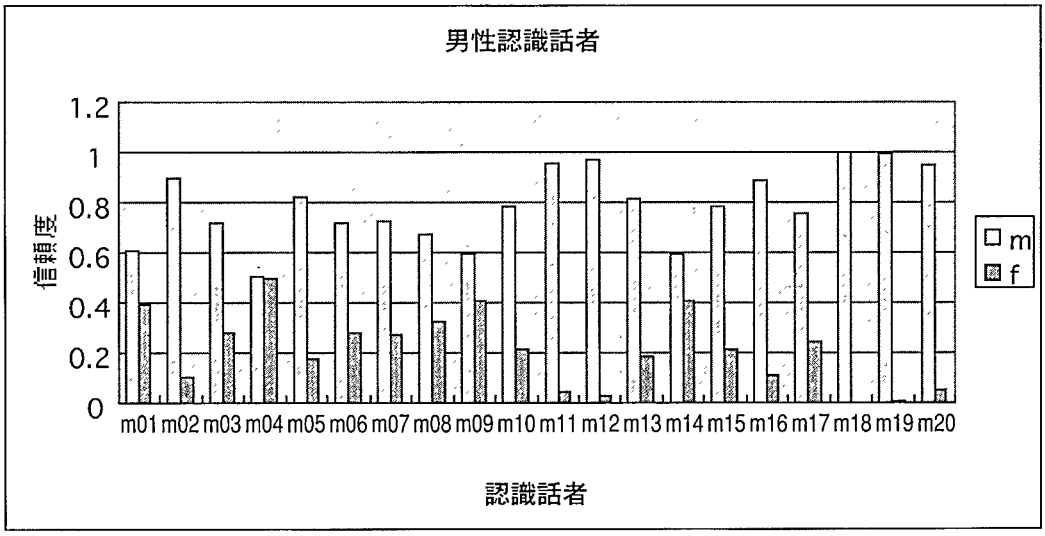


（c） 学習話者：男声（m11,m12,m16,m17,m20）、女声（f02,f05,f06,f07,f10）

図5 学習話者の違いによる男声女声信頼度の比較（男性入力話者、入力単語「2934」）



（a）学習話者　女声（f06～f10）



（b）学習話者　女声（f01～f05）

図6　女声学習話者の違いの影響（男性入力話者、入力単語「2934」）

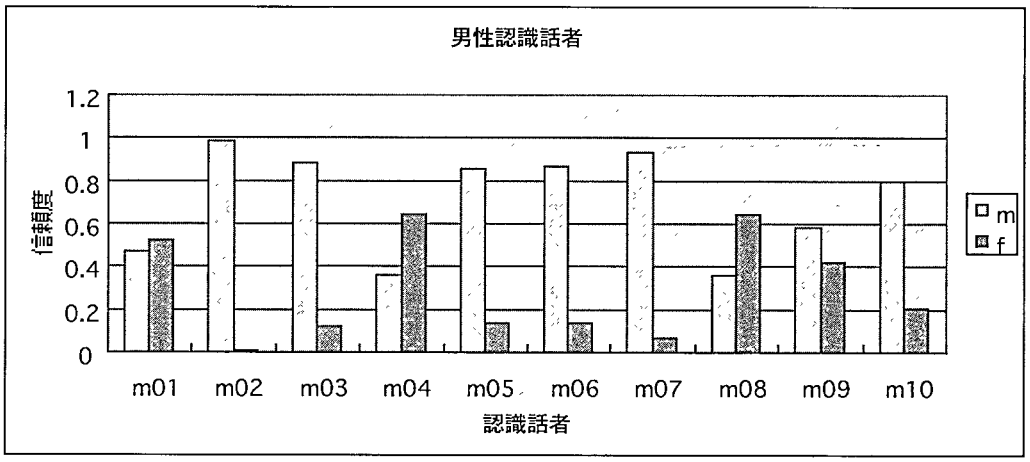
択の影響は少ないことが分かった。

3.2.3　入力単語数の影響

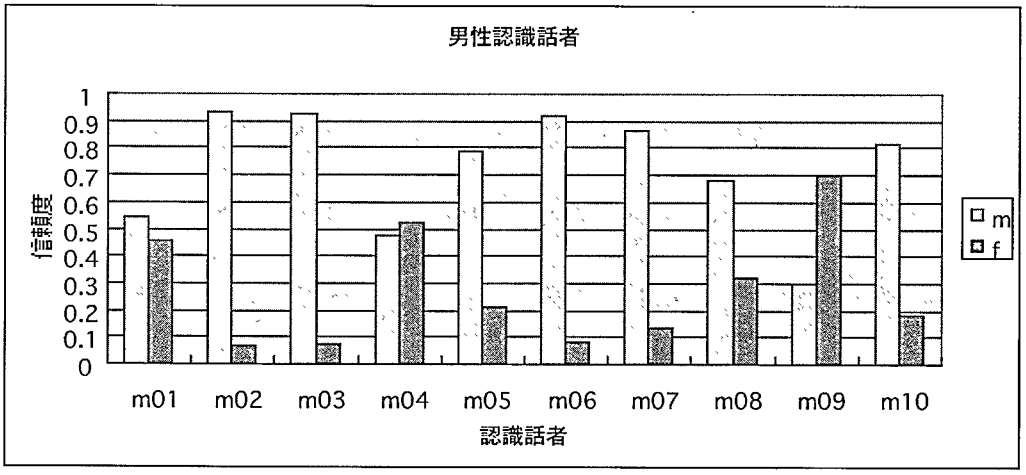
図7は入力単語を3単語とした場合の結果である。(a)は「0712」、「1823」、「2937」、(b)は「3045」、「4156」、「5267」を入力単語とし、学習話者は男声 m16 ～ m20、女声 f01 ～ f05とした。この場合の全話者の男声信頼度の平均値は、(a)0.710、(b)0.724となり入力単語を1単語とした時とほとんど変化がない。

4　むすび

連続音声中の母音の特徴パラメータを用いて男声女声の判別システムを提案し評価実験を行った。女性話者では入力音声に用いる単



（a）入力単語「0712」「1823」「2937」



（b）入力単語「3045」「4156」「5267」

図7　入力単語数の影響：学習話者　男声（m16～m20）、女声（f01～f05）

語の種類や学習話者の選び方に関わりなく、ほぼ100%女声と判定できた。男性話者では種々の条件に関わりなく比較的安定して男声と判定できる話者とそうでない話者が存在する。後者の話者は20名中4名おり、これらの話者に対して男声信頼度を上げる条件として入力単語の種類や学習話者の選定が関わっていることが分かった。今後は母音の種類による男声女声判別の精度を調べることによって入力音声に適した単語を選定し、男性入力話者に対して男声信頼度を上げるような学習話者を選択してより安定した男声女声判定システムの構築を目指す。

参考文献

[1] 松井知子，黒岩眞吾：“音声による個人認証技術の現状と展望　－今、なすべきことは何か！－”，電子情報通信学会誌，vol.87, 4, pp.314-321, (2004)
[2] 今井聖，古市千枝子：“対数スペクトルの不偏推定法”，電子情報通信学会論文誌，vol.J70-A，3，pp.471-480，(1987)
[3] 今井聖，古市千枝子：“連続音声の音素的単位へのセグメンテーション”，電子情報通信

学会論文誌，vol.J72-D-II，1, pp.11-21，(1989)
[4] 古市千枝子，今井聖："多様な音韻環境における音素的単位のセグメンテーション"，電子情報通信学会論文誌，vol.J72-D-II，8, pp.1221-1227，(1989)
[5] 古市千枝子，相澤桂，井上和彦，今井聖："音声認識におけるルールベース法による話者独立音素セグメンテーション"，日本音響学会誌55巻10号，pp.707-716，(1999)
[6] 古市千枝子，今井聖："特定話者任意語い連続音声の音素認識"，電子情報通信学会論文誌，vol.J70-A，3，pp.471-480，(1987)