社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2004－ＳＬＰ－51 （６）
2004／5／28

雑音に頑健な話者照合のための基本周波数情報の利用

浅見 太一† 岩野 公司† 古井 貞熙†

† 東京工業大学大学院 情報理工学研究科 計算工学専攻
〒 152-8552 東京都目黒区大岡山 2-12-1
E-mail: †{taichi,iwano,furui}@furui.cs.titech.ac.jp

あらまし 本稿では，雑音環境下での話者照合性能を向上させるために，韻律情報を利用する手法を提案する．韻律特徴量として $\log F_0$ と $\Delta \log F_0$ を利用し，ケプストラムなどの音韻特徴量と結合して用いる．F_0 は，時間-ケプストラム平面に対してハフ変換を適用することによって雑音に頑健に抽出される．音韻と韻律を融合したモデルは，音節を単位としたマルチストリーム HMM によって構築する．提案手法の有効性を確認するため，様々な SNR 条件下で日本語 4 桁連続数字発声による話者照合実験を行った．実験の結果，提案手法によって全ての SNR 条件で等誤り率の削減が確認された．SNR が 10dB の時に最も大きく性能が改善し，韻律情報を利用しない場合に比べ相対的に 39.9%等誤り率が削減された．

キーワード 話者照合，耐雑音，基本周波数，ハフ変換，マルチストリーム HMM

Use of F_0 information for noise-robust speaker verification

Taichi ASAMI†, Koji IWANO†, and Sadaoki FURUI†

† Department of Computer Science, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
E-mail: †{taichi,iwano,furui}@furui.cs.titech.ac.jp

Abstract This paper proposes a noise-robust speaker verification method using prosodic information. This method uses $\log F_0$ and $\Delta \log F_0$ as prosodic features. They are combined with segmental features such as cepstral parameters. F_0 is extracted by a noise-robust method using the Hough transform which is applied to time-cepstrum images. The segmental and prosodic features are combined and modeled by multi-stream HMMs. Speaker verification experiments were conducted using four-connected-digit utterances of Japanese, contaminated by white noise with various SNRs. Experimental results show that equal error rates were reduced in all SNR conditions. The best reduction was observed at 10dB SNR condition; the error rate was reduced by 39.9% from the baseline method using only segmental features.

Key words speaker verification, noise robustness, fundamental frequency, Hough transform, multi-stream HMM

1. はじめに

近年，高精度な話者照合システムの需要が高まっており，特に話者照合の雑音に対しての頑健性を向上させることが重要な課題となっている．

高性能な話者認識システムの実現のために，基本周波数 (F_0) 情報をスペクトルやケプストラムといった音韻的な特徴量と組み合わせて用いる様々な手法がこれまで提案されてきた [1-9]．F_0 情報は音韻情報よりもチャネル歪みや加算性の雑音に対して頑健であるとされ [2]，話者認識の雑音に対する頑健性を向上させるのに役立つと考えられる．文献 [2] では，電話音声における受話器の種類の変動による性能劣化に対処するために F_0 情報を利用した頑健な話者認識手法を提案している．文献 [4] では，ベクトル量子化に基づく話者識別手法において，F_0 情報を利用することによって加算性雑音に対する頑健性が向上することが報告されている．しかし，話者照合の耐雑音性向上のために F_0 情報を用いた研究はこれまでほとんど報告されていない．

そこで本研究では，F_0 情報を利用した雑音に頑健な話者照合手法を提案する．我々の先行研究 [10] では，韻律情報を用いた雑音に頑健な音声認識手法について報告している．この手法では，画像処理技術の一つであるハフ変換 [11] を用いて雑音に対して頑健に F_0 を抽出し，得られた F_0 情報とケプストラムをマルチストリーム HMM によって融合している．様々な雑音環境下での音声認識実験において，ハフ変換を用いた F_0 抽出

法とモデルの融合法の有効性が確認されていることから，本研究においても，これらの手法を話者照合に適用した．

本稿では，音韻・韻律情報を融合したモデルを用いた話者照合システムを構築し，その雑音環境下での性能を実験によって評価する．

以下，ハフ変換を用いた F_0 の抽出法，音韻情報と韻律情報を融合した話者照合手法，雑音環境下での評価実験について述べる．

2. ハフ変換を用いた雑音に頑健な F_0 抽出法

ハフ変換は画像処理の分野で利用される手法で，雑音を含む画像から直線，円，楕円といったパラメトリックな図形を抽出するのに有効な手法である [11]．

音声の F_0 パターンはある程度の時間連続性を有している．時間-ケプストラム平面に現れる F_0 に相当するピーク値の軌跡は，背景雑音によって，ばらついたり，はっきりと現れなくなったりするため，適当な窓幅で時間-ケプストラム平面を切り出し，その中の最も優位な直線成分を取り出すことで，時間連続性が考慮された雑音に頑健な F_0 値を抽出することができる．従来，フレームごとに F_0 を抽出，確定した後に，その値に対して各種の平滑化を行う手法が考えられてきたが，これらの手法は，F_0 抽出誤りに敏感である．ハフ変換を用いる抽出法は，F_0 抽出前の時間-ケプストラム平面において時間連続性を考慮することから，頑健性が向上する [10]．

2.1 ハフ変換

変換対象画像 (x-y 平面) に n 個の画素 (x_i, y_i) $(i = 1, \cdots, n)$ が存在したとする．この時，各点を次式を用いて m-c 平面上の直線に変換する．

$$c = -x_i m + y_i \quad (i = 1, \cdots, n) \tag{1}$$

この時，m-c 平面の直線上の点に，点 (x_i, y_i) の輝度を累積する．この操作を m-c 平面への投票と呼ぶ．x-y 平面上の全ての点を m-c 平面に投票した後で，m-c 平面上で投票値の累積が最大となる点 (m, c) を選び，以下の式で逆変換することで，最も優位な x-y 平面での直線成分を抽出することができる．

$$y = mx + c \tag{2}$$

図 1 にハフ変換による直線成分抽出の様子を示す．

2.2 ハフ変換を利用した基本周波数抽出法

サンプリング周波数 16kHz の音声データを分析窓長 32ms，フレーム周期 10ms で 256 次元のケプストラムに変換する．今回の実験では，男性話者の発声のみを使用するため，ピークの探索範囲をケプストラムの 60 次元以上 (F_0 で 270Hz 以下) に限定する．さらに，雑音の重畳した音声のケプストラムは，低次部分ほどピーク値が大きくなる傾向があるため，探索領域の低次部 (60 ～ 140 次元) の d 次のケプストラムに次式で示す値を乗算する [10]．

$$0.6 + 0.4 \sin\left(\frac{d - 60}{140 - 60} \times \frac{\pi}{2}\right) \tag{3}$$

次に，F_0 を求めたいフレームを中心に，前後 4 フレーム，計

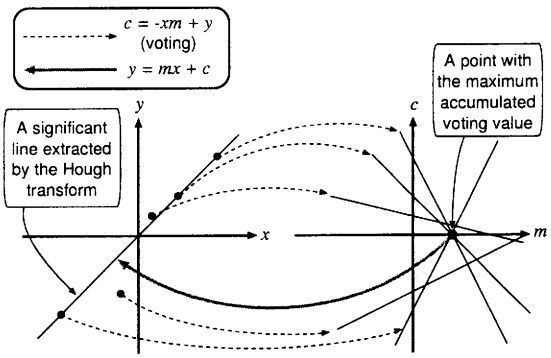

図 1 ハフ変換による直線成分の抽出．

9 フレームの時間-ケプストラム画像を切り出し，ハフ変換を行う．この時，各画素の輝度値はケプストラムの値であり，この値が投票値となる．ただし，全ての画素について投票を行うことは効率的ではないため，一定の閾値以上の値を有する点のみを投票に用いる．本実験では閾値は 0.05 とした．ハフ変換によって得られた直線の中点のケプストラム次数から F_0 の値を計算する．この操作を全てのフレームについて行うことで，9 フレーム分の連続性が考慮された F_0 値が抽出される．

3. F_0 情報を用いた話者照合

3.1 音韻・韻律特徴量の融合

音韻特徴量は，MFCC12 次元・Δ MFCC12 次元・Δパワーの計 25 次元を用いる．特徴量はフレーム長 25ms，フレーム周期 10ms で抽出し，入力音声ごとに CMS を行っている．

韻律特徴量も音韻特徴量と同じフレーム周期で抽出される．特徴量としては，$\log F_0$ と $\Delta \log F_0$ の 2 つの値を用いる．$\Delta \log F_0$ は，

$$\begin{aligned} \Delta \log F_0 &= \frac{d \log F_0}{dt} \\ &= \frac{d \log F_0}{dF_0} \cdot \frac{dF_0}{dt} \\ &= \frac{1}{F_0} \cdot \Delta F_0 \end{aligned} \tag{4}$$

のように展開される．ΔF_0 はハフ変換によって得られた直線の傾きから直接求めることができる．また，本研究では，比較として，ハフ変換を用いず，ケプストラム法によって F_0 を抽出した場合においても実験を行っている．その際には，$\log F_0$ の値を 9 フレーム幅で最小 2 乗近似することによって得られる直線の傾きを $\Delta \log F_0$ として用いる．最終的には，音韻特徴量と韻律特徴量を各フレームごとに結合することで，融合特徴量を作成する．

本研究では，2 つの韻律特徴量 ($\log F_0$ と $\Delta \log F_0$) それぞれの効果を確認するために，韻律特徴量として，$\log F_0$ のみを用いた場合，$\Delta \log F_0$ のみを用いた場合，$\log F_0$ と $\Delta \log F_0$ の両方を用いた場合の 3 通りについて検討する．そのため，融合特徴量は合計 26 または 27 次元となる．

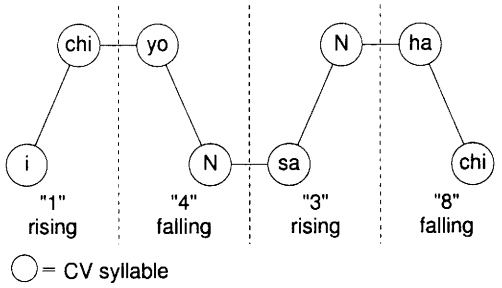

図 2　4 桁連続数字発話の F_0 パターン.

3.2 音韻・韻律モデルの融合

話者照合実験には日本語 4 桁連続数字音声を利用する．日本語連続数字発声では，CV 音節を単位として韻律 (F_0) のパターンを表現するのが容易である．4 桁連続数字音声の F_0 遷移パターンを見ると，図 2 のように，上昇部分と下降部分に分けることができる．そこで，F_0 遷移パターンを CV 音節を単位として表現し，各音節に「上昇」あるいは「下降」という韻律ラベルを付加することによってモデリングする．なお，図中の点線は数字境界を示している．音韻モデルも音節単位で作成し，マルチストリーム HMM によって，F_0 遷移パターンを考慮した音韻・韻律融合モデル (SP-HMM: Segmental-Prosodic HMM) を構築する．

3.2.1 音節単位でのモデリング

全ての数字は 2 つの CV 音節 (2 モーラ) で構成される (「2」は /ni:/ ,「5」は /go:/ と最終母音が長音化した形で扱う)．ここでは数字内部の音韻環境のみを考慮する．したがって，融合モデルは左右どちらかのコンテキストにのみ依存する音節モデルとなる．そこで，融合モデルを，左コンテキスト (LC) 依存の音節 (SYL)「LC-SYL, PM」と，右コンテキスト (RC) 依存の音節 (SYL)「SYL+RC, PM」と表現する．ここで「PM」は F_0 パタンの遷移を示し，上昇 (U)・下降 (D) のいずれかとなる．例えば,「上昇型数字 1 (/ichi/) の第一音節 /i/」は「i+chi, U」と表記される．表 1 に融合モデルの一覧を示す．sil は数字間の長い無音区間および連続数字の最初と最後に入る無音区間を表現し，sp は数字間に入る短い無音区間を吸収するモデルである．

3.2.2 マルチストリーム HMM

融合モデルはマルチストリーム HMM によってモデル化される．音韻と韻律特徴量を 2 つのストリームに分け，それぞれから得られる出力確率を重み付けし，合わせることで，融合特徴量の出力確率を得る．融合特徴量ベクトル $\boldsymbol{O_{sp}}$ が与えられたときの状態 j における出力確率 $b_j(\boldsymbol{O_{sp}})$ は以下の式で与えられる．

$$b_j(\boldsymbol{O_{sp}}) = b_j(\boldsymbol{O_s})^{\lambda_s} \cdot b_j(\boldsymbol{O_p})^{\lambda_p} \tag{5}$$

ここで $b_j(\boldsymbol{O_s})$, $b_j(\boldsymbol{O_p})$ はそれぞれ状態 j で音韻特徴量 $\boldsymbol{O_s}$, 韻律特徴量 $\boldsymbol{O_p}$ の出力確率である．λ_s, λ_p はそれぞれ音韻・韻律ストリーム重みであり，$\lambda_s + \lambda_p = 1$ $(0 \leq \lambda_s, \lambda_p \leq 1)$ とする．

3.2.3 融合モデルの構築

まず，音韻特徴量を用いて音韻モデル (S-HMM: Segmental HMM) を，韻律特徴量を用いて韻律モデル (P-HMM: Prosodic HMM) をそれぞれ構築し，混合ガウス分布を共有化することで融合モデルを作成する．具体的には以下のような手順で構築する．

(1)　まず，音韻特徴量のみを用いて音節単位の音韻モデル (S-HMM) を学習する．各音節モデルは韻律情報を考慮しないため,「i+chi, *」「i-chi, *」のようにワイルド・カード記号「*」を用いて表される．sil モデルと sp モデルを合わせて合計 22 のモデルを作成する．状態数は，音素数× 3 とし，sil モデルは 3 状態，sp モデルは 1 状態とした．

(2)　作成した音節モデルを用いて，学習データの強制切り出しを行い，時間ラベルを作成する．

(3)　得られた時間ラベルの各数字に，人手によって上昇・下降の韻律ラベルを付与し，このラベル情報と韻律特徴量を用いて，韻律モデル (P-HMM) を学習する．韻律モデルは音韻情報を考慮しないため,「上昇型数字の第一音節」は「*+*, U」,「上昇型数字の第二音節」は「*-*, U」と表記される．sil, sp を含め合計 6 モデルを作成し，状態数は全てのモデルで 1 とする．

(4)　融合モデル (SP-HMM) は，各状態の音韻・韻律ストリームの混合ガウス分布を，音韻・韻律モデルそれぞれの混合分布と共有することで構築される．例えば，融合モデル「i+chi, U」の音韻ストリームの混合分布は音韻モデル「i+chi, *」の混合分布と共有し，韻律ストリームの混合分布は韻律モデル「*+*, U」と共有する．なお，混合モデルの状態数は音韻モデルと同じ (音素数× 3) とする．韻律モデルの状態数は 1 であるので，融合モデルの全状態は，この 1 状態のみと混合分布の共有を行う．例として,「1」のモデルを融合する様子を図 3 に示す．

3.3 話者照合スコア

特徴量 x が入力されたとき，申告話者 S^c である確率 $p(S^c|x)$ は以下のように定義される．

$$p(S^c|x) = \frac{p(x|S^c)p(S^c)}{p(x)} \tag{6}$$

ここで，音声特徴量の生起確率 $p(x)$ を，不特定話者モデルからの特徴量の出現確率 $p(x|S^g)$ を用いて表すと，

$$p(S^c|x) = \frac{p(x|S^c)p(S^c)}{p(x|S^g)p(S^g)} \tag{7}$$

となる．各話者について，申告話者の出現確率 $p(S^c)$ は共通であると仮定し，さらに不特定話者モデルの生起確率は定数となるため，

$$p(S^c|x) \propto \frac{p(x|S^c)}{p(x|S^g)} \tag{8}$$

となる．これは，特定話者モデルから得られた尤度を不特定話者モデルから得られた尤度で正規化することを意味している．式 (8) の右辺は，

表 1 融合モデル (SP-HMM) の一覧．融合モデルは「LC-SYL,PM」「SYL+RC,PM」と表記され，「LC-SYL,PM」は左コンテキスト依存の音節モデル，「SYL+RC,PM」は右コンテキスト依存の音節モデルとなる．「PM」は F_0 パターンの遷移を示し，上昇（「U」）・下降（「D」）で表現される．

digit	model	
0	ze+ro,U	ze+ro,D
/zero/	ze-ro,U	ze-ro,D
1	i+chi,U	i+chi,D
/ichi/	i-chi,U	i-chi,D
2	ni+i,U	ni+i,D
/ni:/	ni-i,U	ni-i,D
3	sa+N,U	sa+N,D
/saN/	sa-N,U	sa-N,D

digit	model	
4	yo+N,U	yo+N,D
/yoN/	yo-N,U	yo-N,D
5	go+o,U	go+o,D
/go:/	go-o,U	go-o,D
6	ro+ku,U	ro+ku,D
/roku/	ro-ku,U	ro-ku,D
7	na+na,U	na+na,D
/nana/	na-na,U	na-na,D

digit	model	
8	ha+chi,U	ha+chi,D
/hachi/	ha-chi,U	ha-chi,D
9	kyu+u,U	kyu+u,D
/kyu:/	kyu-u,U	kyu-u,D
	sil	sp

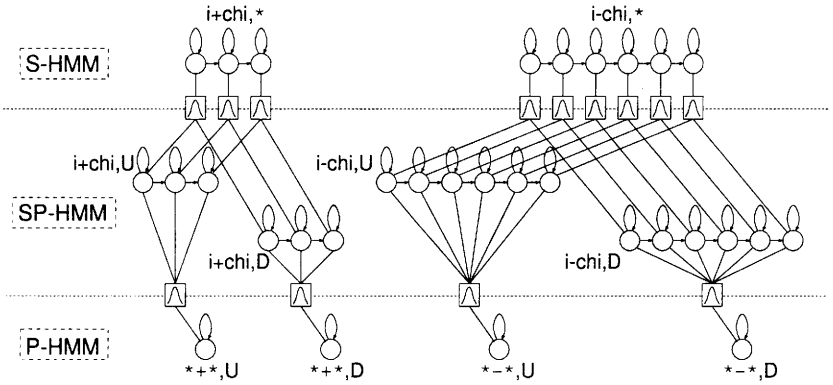

図 3 mixture 共有による融合モデル (SP-HMM) の構築．音韻モデル (S-HMM) は音韻特徴量のみから，韻律モデル (P-HMM) は韻律特徴量のみから学習される．

$$\frac{p(x|S^c)}{p(x|S^g)} = \frac{\Sigma_w p(x|S^c,w)p(w)}{\Sigma_w p(x|S^g,w)p(w)} \approx \frac{\max_w p(x|S^c,w)}{\max_w p(x|S^g,w)} \quad (9)$$

のように計算される．w は 4 桁連続数字列である．これは，通常の音声認識に用いられるのと同様の尤度計算を特定話者モデルと不特定話者モデルに対して行い，得られた 2 つの尤度を照合に用いることを意味している．

話者照合スコアは，対数を用いて，

$$p = \log p(x|S^c) - \log p(x|S^g) \quad (10)$$

と定義し，このスコアが閾値を越えた時に，申告者本人であると判断する．申告話者・不特定話者モデルにマルチストリーム HMM による融合モデルを用いる．

本実験で用いた話者照合システムの処理の流れを図 4 に示す．申告話者・不特定話者モデルに，マルチストリーム HMM による融合モデルを用いる．

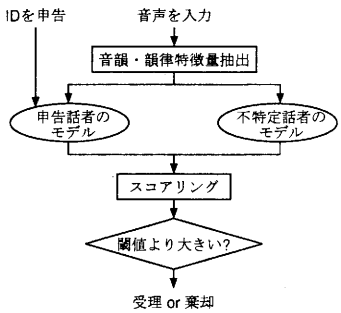

図 4 音声入力から照合までの流れ．

4. 話者照合実験

4.1 実験条件

4.1.1 実験方法

実験に用いた音声データは，男性話者 37 名から，時期差による変化を考慮し，1 ヵ月毎に 5 時期に渡って収録を行ったものである．各話者が 1 時期に 50 個の 4 桁連続数字を発声しており，音声は 16kHz，16bit で標本化・量子化した．

1 ～ 3 時期目のデータを学習データ，4，5 時期目のデータを評価データとする．不特定話者モデルの学習データに含まれている詐称者と含まれていない詐称者を用意するため，学習データは 18 名と 19 名の 2 グループに分ける．例えば，第 1 グループに属する話者を申告話者として照合実験を行う場合は，第 2 グループに属する全ての話者のデータで学習した不特定話者モデルを利用して尤度の正規化を行う．こうすることで，話者ごとの評価データは，「本人のデータ (1 名分)」「不特定話

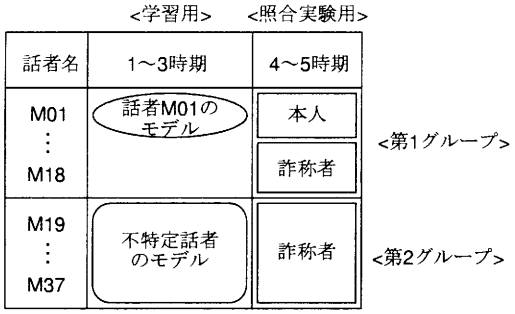

図 5　申告話者が M01 の時の学習データと評価データ.

者モデルの学習に含まれている詐称者のデータ (19 名分)」「不特定話者モデルの学習に含まれていない詐称者のデータ (17 名分)」となる．話者 M01 に対して照合実験を行う際のデータの用い方を図 5 に示す．

学習データには SN 比 30dB の白色雑音を付加させ，評価データには SN 比 5，10，15，20，30dB の白色雑音を付加させたものを用いる．

4.1.2　辞書・文法

本実験における話者照合は，テキスト独立型である．単語辞書には，各数字が 2 種類定義されている．例えば，数字 1 は「i+chi,U i-chi,U sp」「i+chi,D i-chi,D sp」の 2 種類の読みが定義されている．韻律の変化は図 2 のように数字境界で生じていることから，数字内で変化することはないとしている．

文法は，任意の 4 桁連続数字とし，数字-数字での韻律パターンの変化は任意とした．

4.2　実験結果

4.2.1　韻律特徴量による等誤り率の比較

韻律特徴量である $\log F_0$，$\Delta \log F_0$ それぞれの効果を確認するために，各 SNR において話者照合性能の比較を行った．音韻情報のみのモデル (S-HMM) の性能は本研究でのベースラインとなる．融合モデル (SP-HMM) では，**H-L**，**H-D**，**H-LD** の 3 種類の韻律特徴量について実験を行った．各韻律特徴量の構成は表 2 に示すとおりである．音韻・韻律ストリーム重みは SNR ごとに事後的に最適値に設定した．HMM の混合数は，S-HMM の特定話者・不特定話者モデル，P-HMM の特定話者・不特定話者モデルともに 4 とした．これは SNR が 30dB のときに最も高い性能が得られた混合数である．

結果を図 6 に示す．全ての SNR において，どの韻律特徴量を用いた場合でも，音韻情報のみのモデルによる照合よりも性能が向上していることが分かる．また，**H-D** よりも **H-L** を使った時の方が照合性能が高いことから，$\log F_0$ の方が $\Delta \log F_0$ よりも話者性を有していることが分かる．さらに **H-LD** を用いた時の性能が最も高くなっていることから，2 つの韻律特徴量は相補的に働いているといえる．**H-LD** を利用したとき照合性能が最も大きく改善したのは SNR が 10dB の時であり，相対的に 39.9% の等誤り率の削減が確認された．後の実験では，**H-LD** を韻律特徴量として用いた．

表 2　ハフ変換によって抽出した 3 種類の韻律特徴量.

韻律特徴量	特徴量の構成要素 (次元数)
H-L	$\log F_0$ (1)
H-D	$\Delta \log F_0$ (1)
H-LD	$\log F_0$, $\Delta \log F_0$ (2)

4.2.2　ハフ変換の雑音に対する効果の検証

ハフ変換を利用したことによる効果を確認するために，F_0 の抽出にハフ変換を用いた場合と，従来のケプストラム法を用いた場合の各 SNR における照合性能の比較を行った．韻律特徴量 **NH-LD** はハフ変換を用いずに抽出した $\log F_0$ と $\Delta \log F_0$ からなる．

韻律特徴量 **H-LD** と **NH-LD** による照合実験結果を図 7 に示す．全ての SNR において，**H-LD** を用いた方が **NH-LD** を用いるよりも等誤り率が低くなっている．さらに，より多くの雑音が重畳するほど，ハフ変換を用いることによる照合性能の向上率が大きくなっており，ハフ変換を用いた F_0 抽出が話者照合の耐雑音性を向上させるのに有効であることが分かる．

4.2.3　ストリーム重みによる等誤り率の比較

各 SNR において，韻律ストリーム重み (λ_p) を変化させた時の等誤り率の推移を図 8 に示す．全ての SNR において，$0.0 < \lambda_p < 0.9$ という広い範囲において，ベースライン ($\lambda_p = 0$) からの性能改善が見られる．このことから，提案手法はストリーム重みのずれに対して頑健であるといえる．

また，表 3 は，各 SNR において，最も高い性能を得られた韻律ストリーム重みを示したものである．この結果から，雑音がより多く重畳するほど，韻律情報が照合に役立っていることが分かる．

5.　ま　と　め

本稿では，雑音環境下での 4 桁連続数字音声による話者照合において，基本周波数情報の利用を検討した．時間-ケプストラム平面のハフ変換によって抽出された雑音に頑健な F_0 情報を有効に利用するために，音節単位のマルチストリーム HMM を用いて音韻・韻律の融合モデルを構築し，そのモデルによる雑音環境下での話者照合性能を評価実験によって確認した．実験の結果，融合モデルを用いることによって，音韻情報のみのモデルに比べて，全ての SNR 条件において等誤り率が減少することを確認した．また，韻律特徴量として，$\log F_0$ と $\Delta \log F_0$ の両方を用いた場合に最も性能が高くなることから，それぞれの特徴量が相補的に照合性能の改善に貢献していることが確認された．さらに，ハフ変換による F_0 抽出法が話者照合の頑健性向上に役立つこと，提案する話者照合法がストリーム重みのずれに対して頑健であることを確認した．

今後の課題としては，1) F_0 情報以外の韻律特徴量の検討，2) 融合 HMM のトポロジーの改善，3) 有声・無声情報の効果的な利用法の検討，4) ストリーム重みの自動最適化手法の導入，5) 白色雑音以外の雑音における本手法の効果の確認などが挙げられる．

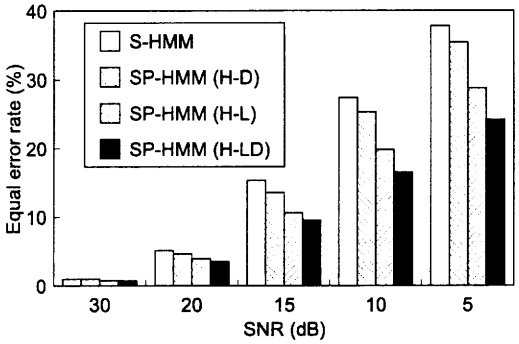

図 6 各 SNR における融合モデル (SP-HMM) と音韻モデル (S-HMM) の等誤り率の比較.

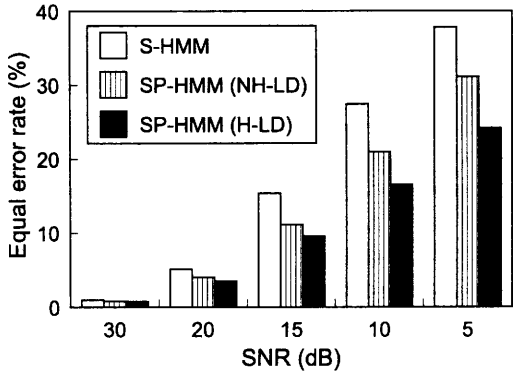

図 7 基本周波数抽出にハフ変換を用いた場合 (**H-LD**) とケプストラム法を用いた場合 (**NH-LD**) の等誤り率の比較.

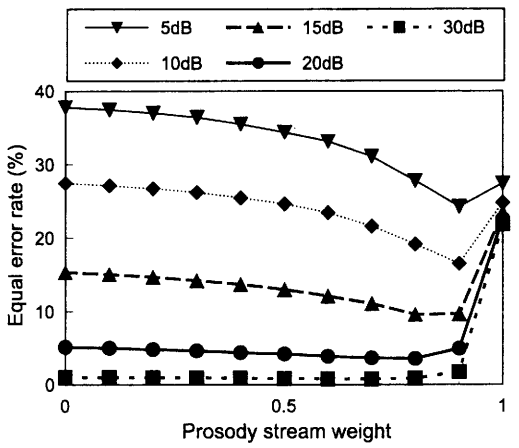

図 8 各 SNR において韻律ストリーム重み (λ_p) を変化させたときの等誤り率の推移.

表 3 各 SNR における最適ストリーム重み.

SNR	Optimal prosody weight
30dB	0.7
20dB	0.8
15dB	0.8
10dB	0.9
5dB	0.9

文 献

[1] 松井知子, 古井貞熙, "音源・声道特徴を用いたテキスト独立形話者認識," 信学論, vol.J75-A, no.4, pp.703-709 (1992-4).

[2] M.J. Carey, E.S. Parris, H. Lloyd-Thomas, and S. Bennett, "Robust prosodic features for speaker identification," *Proc. ICSLP96*, vol.3, pp.1800-1803, Philadelphia, Pennsylvania (1996-10).

[3] M.K. Sönmez, L. Heck, M. Weintraub, and E. Shriberg, "A lognormal tied mixture model of pitch for prosody-based speaker recognition," *Proc. Eurospeech97*, vol.3, pp.1391-1394, Rhodes (1997-9).

[4] Y.-J. Kyung and H.-S. Lee, "Text independent speaker recognition using micro-prosody," *Proc. ICSLP98*, vol.1, pp.157-160, Sydney (1998-12).

[5] Y. Cheng and H.-C. Leung, "Speaker verification using fundamental frequency," *Proc. ICSLP98*, vol.1, pp.161-164, Sydney (1998-12).

[6] K.P. Markov and S. Nakagawa, "Text-independent speaker recognition using multiple information sources," *Proc. ICSLP98*, vol.1, pp.173-176, Sydney (1998-12).

[7] 服部陽介, 徳田恵一, 益子貴史, 小林隆夫, 北村 正, "多空間ガウス混合モデルを用いた話者認識," 音講論, vol.1, pp.99-100 (2000-3).

[8] F. Weber, L. Manganaro, B. Peskin, and E. Shriberg, "Using prosodic and lexical information for speaker identification," *Proc. ICASSP2002*, vol.1, pp.141-144, Orlando, Florida (2002-5).

[9] D. Reynolds, et al., "The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition," *Proc. ICASSP2003*, vol.4, pp.784-787, Hong Kong (2003-4).

[10] 岩野公司, 関 高浩, 古井貞熙, "雑音に頑健な基本周波数抽出法とその音声認識への適用," 信学技報, vol.102, no.35, pp.37-42 (2002-4).

[11] P.V.C. Hough, "Method and means for recognizing complex patterns," U.S. Patent #3069654 (1962).