論文 / 著書情報
Article / Book Information

論題(和文)	ハフ変換による基本周波数情報を用いた雑音に頑健な話者照合
Title(English)	
著者(和文)	浅見 太一, 岩野 公司, 古井 貞熙
Authors(English)	Koji Iwano, SADAOKI FURUI
出典(和文)	日本音響学会 2004年春季講演論文集, Vol. , No. 3-Q-17, pp. 177-178
Citation(English)	, Vol. , No. 3-Q-17, pp. 177-178
発行日 / Pub. date	2004, 3

ハフ変換による基本周波数情報を用いた雑音に頑健な話者照合 *

◎浅見 太一　岩野 公司　古井 貞熙 (東工大)

1　はじめに

韻律情報の一つである基本周波数 (F_0) 情報は個人性を含むとされ，これまでに韻律情報を利用した話者認識の研究が幾つかなされている [1-3]．一方，音声認識に F_0 情報を用いることによって認識性能の耐雑音性が向上することが報告されている [4]．それに対して，これまで話者照合性能の耐雑音性を向上させるために F_0 情報を利用した研究例は少ない．

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について論ずる．話者照合に用いる韻律特徴量として，時間-ケプストラム平面をハフ変換 [5] することで得られる雑音に頑健な F_0 情報を利用する．音韻情報と韻律情報をマルチストリーム HMM によって融合し，申告話者・不特定話者モデルとして利用する．

以下，ハフ変換を用いた韻律特徴量の抽出法，音韻情報と韻律情報を融合した話者照合手法，提案手法の耐雑音性を確認する評価実験について述べる．

2　ハフ変換による F_0 情報の抽出

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し，雑音の影響を低減するため低次のケプストラムを小さく見積もるようにリフタリング処理を行う [4]．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 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．

ハフ変換は以下のように行われる．まず，対象画像 (x-y 平面) に n 個の画素 $(x_i, y_i)(i = 1, ..., n)$ が存在するとき，各点を次式を用いて m-c 平面上の直線に変換する．

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

このとき，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．次に，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 平面上の直線を抽出する．

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

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．

韻律特徴量は，$\log F_0$ 1 次元を用いる．抽出方法としてハフ変換を利用した場合 (P_h) と，従来のケプストラム法を使った場合 (P_{nh}) の 2 通りの韻律特徴量について検討する．

音韻特徴量と韻律特徴量は同じフレーム周期であり，両者を各フレーム毎に結合することで，合計 26 次元の融合特徴量を作成する．

3.2　音韻・韻律モデルの融合

本研究では，4 桁連続数字音声に対する話者照合をタスクとしている．連続数字発声では CV 音節を単位として韻律 (F_0) のパタンが表現される [4]．そこで，音韻・韻律の融合モデル (SP-HMM: Segmental-Prosodic HMM) を音節単位で構築する．

融合モデルは，数字内部の音韻環境のみを考慮し，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)「RC+SYL, PM」と表現される．ここで「PM」は F_0 パターンの遷移を示し，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．

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．

$$b_j(\boldsymbol{O_{sp}}) = b_j(\boldsymbol{O_s})^{\lambda_s} \cdot b_j(\boldsymbol{O_p})^{\lambda_p} \tag{3}$$

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．

融合モデルは，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: Segmental 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．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する sil，数字間に入る短い無音区間を吸収するための 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: Prosodic 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

∗ Noise robust speaker verification using F_0 information extracted by Hough transformation

By Taichi Asami, Koji Iwano, and Sadaoki Furui (Tokyo Institute of Technology)

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．

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．

$$p(S^c|x) = \frac{p(x|S^c)p(S^c)}{p(x)} \tag{4}$$

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，

$$p(S^c|x) = \frac{p(x|S^c)p(S^c)}{p(x|S^g)p(S^g)} \tag{5}$$

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，

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

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．そこで，話者照合スコアを，

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

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．

4 話者照合実験

4.1 音声データ

音声データは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．男性話者 37 名が 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．

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．

4.2 実験結果

融合モデル (SP-HMM) と F_0 情報を用いていない音韻モデル (S-HMM) の各 SN 比における話者照合の等誤り率 (Equal error rate) を図 1 に示す．図中の SP-HMM-nh は韻律特徴量として P_{nh} を利用した場合，SP-HMM-h は P_h を利用した場合を示している．各 HMM の混合数は，30dB の雑音環境で行った予備的な実験から，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．音韻・韻律ストリーム重みは各 SN 比での実験ごとに事後的に最適値を設定した．

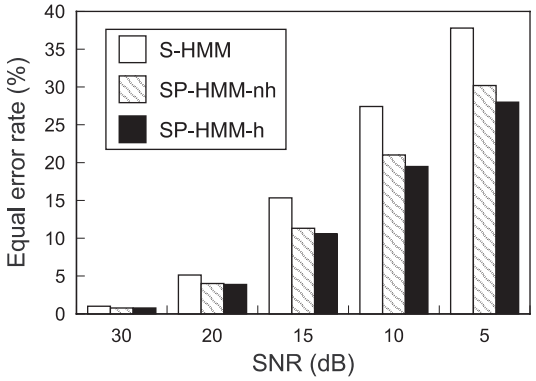

図 1. 各 SN 比における融合モデル (SP-HMM) と音韻モデル (S-HMM) の等誤り率の比較

全ての SN 比において，音韻情報のみのモデルで照合を行うよりも，音韻・韻律の融合モデルを用いた方が照合性能が高くなっていることがわかる．また，韻律特徴量として P_{nh} よりも P_h を用いた方が誤り率が小さくなる傾向がある．これは，ハフ変換による F_0 抽出方法がケプストラム法による抽出に比べて雑音に対して頑健であることを示している．

照合性能が最も大きく改善したのは，実環境に近い 15dB の雑音条件の時であり，31.1% の誤り率削減が確認された．

5 まとめ

ハフ変換によって抽出した F_0 情報を用いた話者照合手法を提案し，4 桁連続数字音声を用いた実験において本手法の有効性を示した．提案手法を用いることにより，音韻情報のみによる照合と比べて，全ての SN 比において等誤り率が減少した．

今後の課題としては，$\log F_0$ 以外に利用可能な韻律特徴量の検討，最適ストリーム重みの自動設定手法の導入，融合モデルのトポロジーの検討などが挙げられる．

参考文献

[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] 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).

[3] F. Weber, L. Manganaro, B. Peskin, and E. Shriberg, “Using prosodic and lexical information for speaker identification,” Proc. ICASSP, vol.1, pp.141-144 (2002-5).

[4] 岩野公司，関 高浩，古井貞熙，“雑音に頑健な音声認識のための韻律情報の利用，” 情処研報, vol.2003, no.58, pp.55-60 (2003-5).

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