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音声認識
Title(English)	
著者(和文)	岩野 公司, 関 高浩, 古井 貞熙
Authors(English)	Koji Iwano, SADAOKI FURUI
出典(和文)	日本音響学会 2002年秋季講演論文集, Vol. , No. 1-9-12, pp. 23-24
Citation(English)	, Vol. , No. 1-9-12, pp. 23-24
発行日 / Pub. date	2002, 9

ハフ変換による基本周波数情報を用いた雑音に頑健な音声認識 *

◎岩野 公司　関 高浩　古井 貞熙 (東工大)

1　はじめに

本稿では，雑音環境下での音声認識性能を，韻律情報を利用して向上させる手法を提案する．

これまで，韻律情報を利用した連続音声認識の研究としては，話し言葉音声認識性能の向上に利用したもの [1] などがあるが，耐雑音性の向上に利用した研究例はほとんど報告されていない．しかし，韻律情報の一つである基本周波数（F_0）情報は，句や単語境界の推定に役立つとされ，雑音環境下で頑健に抽出することが出来れば，雑音重畳音声の認識性能向上に有効であると考えられる．

そこで，韻律特徴量として，時間–ケプストラム平面をハフ変換 [2] することで得られる F_0 情報を利用する．ハフ変換は雑音を含む画像から頑健に直線成分を抽出することが可能な手法であり，時間–ケプストラム平面に適用することによって雑音に頑健な F_0 情報を得ることができる．

以下では，ハフ変換を用いた韻律特徴量の抽出法，韻律情報と音韻情報の融合，さらにそのモデルを利用した連続数字音声認識における耐雑音性の評価について報告する．

2　ハフ変換による F_0 情報の抽出

16kHz サンプリング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ためのリフタをかける．そして，特徴量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 [3]．

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) $(i = 1, \cdots, n)$ 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．

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

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．

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

韻律特徴量としては以下の 2 つの特徴量（P_D, P_V）を考える。

(1) F_0 パタンの変化情報を表す $\Delta \log F_0$ (P_D)

(2) F_0 の時間連続性の度合いを示すハフ変換の累積投票値 (P_V)

なお，$\Delta \log F_0$ は，$\Delta F_0 / F_0$ と展開されることから，ハフ変換によって得られる直線の傾きを直接特徴量の算出に利用することができる．

それぞれの特徴量は，単語（句）境界の推定や，有声部と無声/無音部の境界推定に有効である．

3　音韻・韻律情報の融合

3.1　音韻・韻律特徴量の融合

音響特徴量は，MFCC 12 次元・Δ MFCC 12 次元・Δ パワーの計 25 次元を用いる．特徴量抽出のフレーム長は 25ms，フレーム周期は 10ms であり，入力音声ごとに CMS を行っている．

韻律特徴量は，P_D, P_V の両方を用いる場合（2 次元），P_D のみを用いる場合（1 次元），P_V のみを用いる場合（1 次元）の 3 通りについて検討する。

韻律特徴量と音韻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 27 または 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，連続数字音声認識をタスクとし，数字発声に出現する音節と韻律（F_0）のパタンを利用する．このため，音節を単位とした音韻・韻律の融合モデル（SP-HMM: Segmantal-Prosodic HMM）を構築する [4]．

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）「SYL+RC,PM」をモデル化する．ここで「PM」は F_0 パタンの遷移を示し，上昇（U）・下降（D）・平坦（F）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 によってモデル化する．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．

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．連続数字間の無音を表す sil, 数字間に短い無音が入った場合にそれを吸収するための sp モデルをあわせて合計 20 のモデルを作成する．状態数は，音素数 × 3 とする．

(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・平坦の韻律ラベルを付与する．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

∗ Noise robust speech recognition using F_0 information extracted by Hough transformation

By Koji Iwano, Takahiro Seki and Sadaoki Furui (Tokyo Institute of Technology)

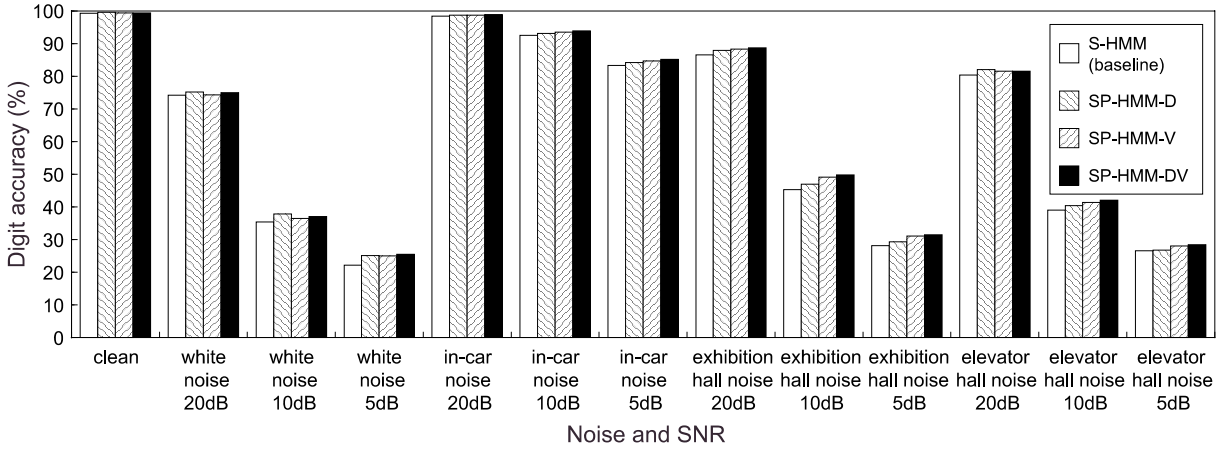

図 1. 各雑音条件における融合モデル（SP-HMM）と音韻モデル（S-HMM）の数字正解精度

二音節」は「`*-*,U`」と表される．`sil`, `sp` を含め合計 8 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，融合モデルの状態数は音韻モデルと同じ（音素数 $\times 3$ ）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．

4 認識実験

4.1 実験条件

本実験で使用した音声データは clean な環境で録音した男性話者 11 名による連続数字音声である．全ての話者は 2 桁から 8 桁の連続数字をそれぞれ 30 回発声しており，話者 1 名あたり 210 連続数字（1,050 数字）を発声している．なお，連続数字間の無音数は 1 名あたり約 229 であった．

実験には leave-one-out 法を用いる．これは，ある話者の発声を認識する際に，残りの 10 名の話者の音声を学習データとするもので，これを全話者の認識に対して行い，最終的に 11 名の正解精度の平均で評価を行う．

モデルの学習は clean な音声を用いる．認識実験では，clean な音声に加え，白色雑音と，電子協騒音データベースの走行車内・展示場・エレベータホール雑音の計 4 種類の雑音を重畳した音声を用いる．重畳する雑音の SNR は 5, 10, 20dB とした．

認識に用いる文法には，「連続数字 → 無音 → 連続数字 ···」というような繰り返しを定義しており，連続数字に桁数制限はない．韻律のパタンについては，数字内の音節遷移では変化せず，数字から数字への遷移では変化は任意としている．

4.2 実験結果

各雑音条件における融合モデル（SP-HMM）と音韻モデル（S-HMM）の数字正解精度を図 1 に示す．図中 SP-HMM-DV は P_D, P_V 両方ともに韻律特徴量として利用していることを示しており，SP-HMM-D は P_D のみ，SP-HMM-V は P_V のみを利用していることを示している．HMM の混合数は，予備的な実験から S-HMM, P-HMM ともに 4 とした．挿入ペナルティ，音韻・韻律ストリーム重みについては，各雑音条件ごとに事後的に最適値を定めた．

全ての雑音条件において，どのような韻律特徴量の組み合わせを用いた場合でも，融合モデルの性能が上回っていることがわかる．また，どちらか一方の韻律特徴量を使った場合よりも，P_D, R_V 両方の韻律特徴量を用いた方が，認識性能が向上する傾向にある．これは，両韻律特徴量が相補的な役割を果たしていることを意味している．

最も韻律特徴量による認識性能の改善が得られたのは，10dB の展示場雑音が重畳した音声を P_D, P_V の両方を特徴量として用いて認識したときであり，絶対値で約 4.5% の数字正解精度の改善がみられた．

5 まとめ

音声認識の耐雑音性の向上を目的として，音韻と韻律の特徴量・モデルの融合手法を提案し，連続数字音声認識において本手法の有効性を確認した．また，本稿では示さなかったが，1) 本手法の有効性が話者に依存しないこと，2) 韻律特徴量の導入により，数字境界の推定精度が向上していること，3) 韻律情報を用いない場合は，最適な挿入ペナルティーが各雑音条件ごとに大きくばらつくが，韻律情報を導入することで，各雑音条件における最適な挿入ペナルティーが収束するため，雑音環境下において頑健なパラメータ設定が可能になる，といったことが実験的に確認されている [4]．

今後の課題としては，MLLR などの雑音適応手法と組み合わせたときの有効性の検証などが挙げられる．

参 考 文 献

[1] A. Stolcke, et al., *Proc. Eurospeech'99*, vol.1, pp.311–314 (1999-9).

[2] P.V.C. Hough, U.S. Patent #3069654 (1962).

[3] 関 他, 情処研報, 2001-SLP-38-2, pp.9-14 (2001-10).

[4] 岩野 他, 信学技報, SP2002-13, pp.37-42 (2002-4).