2003－SLP－46 (10)
2003／5／27

雑音に頑健な音声認識のための韻律情報の利用

岩野公司　　関　高浩　　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
Email: {iwano, tseki, furui}@furui.cs.titech.ac.jp

本論文では，韻律情報を利用した雑音に頑健な音声認識手法について述べる．韻律特徴量として，時間–ケプストラム平面のハフ変換から得られる対数基本周波数の傾き（$\Delta \log F_0$）と最大累積投票値を利用し，通常の音声認識で用いられる音響特徴量と結合して用いる．音韻と韻律の融合モデルは，音節単位のマルチストリーム HMM で構築する．融合モデルの様々な雑音環境における頑健性を確認するため，不特定話者の連続数字発声を対象とした音声認識実験を行った．実験の結果，本手法によって様々な雑音環境において数字正解精度の改善が確認され，$\Delta \log F_0$ と最大累積投票値が相補的に認識性能の向上に貢献することがわかった．また，基本周波数情報を音声認識に用いることで，雑音環境下における数字境界の推定精度が向上し，それによって，数字正解精度の改善と，頑健な挿入ペナルティーの設定が実現されることが確認された．

Use of Prosodic Information for Noise-Robust Speech Recognition

Koji Iwano, Takahiro Seki, and Sadaoki Furui

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

This paper describes a noise robust speech recognition method using prosodic information. This method uses two prosodic features: a time derivative $\log F_0$ ($\Delta \log F_0$) and a maximum accumulated voting value computed by the Hough transform of time-cepstrum images. The prosodic features are combined with conventional cepstral parameters. Segmental and prosodic HMMs are integrated using a multi-stream technique. Speaker-independent experiments were conducted using connected digits uttered in various kinds of noise and SNR conditions. Experimental results show that both prosodic features improve the digit recognition accuracy in all noise conditions and the effects are almost additive. Furthermore, it is confirmed that improvements of digit accuracy and robust adjustments of the insertion penalty in noisy environments are attributed to precise digit boundary detection using F_0 information.

1　はじめに

近年，自動車内環境や携帯電話環境における音声認識システム導入の需要が高まっている．このような環境においては，音声認識の雑音に対する頑健性の向上が重要な課題である．

一方，人間は雑音環境下において，韻律情報を音声の認識・理解に役立てていると言われている．日本語の場合，韻律情報としては，特に基本周波数（F_0）パターンに現れる，単語のアクセント，文のイントネーションなどが重要である．

このような観点から，我々は，韻律情報として F_0 情報を利用した，雑音に頑健な音声認識手法を提案している [1, 2]．雑音環境下において正確に F_0 を抽出することは難しいため，我々の方法では，ハフ変換 [3] を利用して雑音に頑健な F_0 情報の抽出を行っている [4]．ハフ変換とは，ノイズ成分を含んだ画像中から，パラメトリックな図形情報を頑健に抽出する方法であり，提案手法では，時間–ケプストラム平面を画像とみなし，フレームごとに一定窓長で切り出した画像に対してハフ変換を施し，直線成分の抽出を行っている．このようにする

ことで，窓長に相当する時間の F_0 の連続性が考慮された，頑健な F_0 の傾き成分が抽出される．同時に，ハフ変換の過程で，直線の信頼度に相当する特徴量（最大累積投票値）が得られる．文献 [1] では，これら 2 つの特徴量を通常の音声認識で用いられる音響特徴量と組み合わせて融合特徴量を作成し，それぞれの特徴量系列を別ストリームとして受理するマルチストリーム HMM に入力することで，認識を行う手法について報告を行っており，雑音環境における連続数字音声の認識性能の改善を確認している．また，その効果が，韻律情報の利用による数字境界の検出性能の改善から得られていることを報告している．

本論文では，本手法の性能に関するさらに詳細な検討を行う．具体的には，2 つの韻律に関する特徴量が，それぞれどの程度性能改善に貢献しているかについての実験と，その結果についての報告を行う．また，韻律情報を用いることによって，最適な挿入ペナルティーを雑音に対して頑健に設定できることを示す実験結果について報告する．

2 ハフ変換による F_0 情報の抽出

時間–ケプストラム領域に現れる 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–ケプストラム領域を切り出し，その中の最も優位な直線成分をハフ変換によって取り出すことで，時間連続性が考慮され，雑音に頑健な F_0 情報を抽出することができる [4].

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．

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

2.2 F_0 情報の抽出

まず，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）の d 次のケプストラムに次式で示す値を乗算しておく．

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

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 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．このとき，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．なお，本実験では閾値は 0.05 とした．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 の直線成分が抽出される．

3 韻律情報を用いた連続数字音声認識

音声認識のタスクとして日本語連続数字発声を取り扱う．日本語の連続数字発声は 2 桁あるいは 3 桁で一つのアクセント句を形成することが多い．これを CV 音節単位で見ると，図 1 のように，それぞれのアクセント句を上昇・下降・平坦部分に分けることができる．この F_0 形状の遷移は，図 1 の点線で示される数字境界で生じていることから，F_0 の遷移情報を用いることで，認識時に数字のアライメント精度が向上し，それによって，認識性能の向上が期待される．

3.1 音韻・韻律特徴量の融合

音韻特徴量としては MFCC 12 次元・Δ MFCC 12 次元・Δ 対数パワーの計 25 次元を用いる．特徴量抽出のフレーム長は 25 ms，フレーム周期は 10 ms であり，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，1) F_0 パターンの遷移

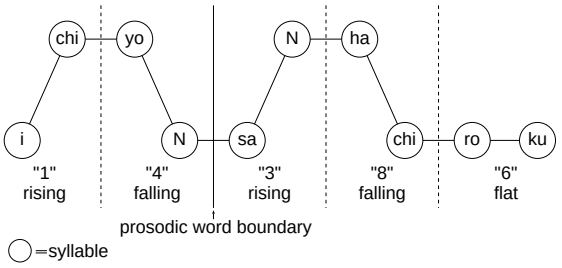


図 1: 日本語連続数字発声の F_0 パターンの例.

情報を表す $\Delta \log F_0$，2) 得られた F_0 がどの程度時間連続性を有しているかを示すハフ変換の最大累積投票値，の 2 つを考える．$\Delta \log F_0$ は,

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

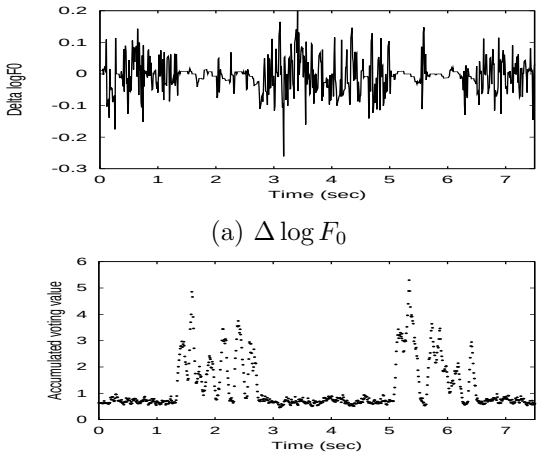
のように展開され，ハフ変換によって得られた直線の傾き ΔF_0 と，直線の中点のケプストラム次数を変換することで得られる F_0 値から計算することができる．図 2 (a) に SNR=20dB の白色雑音が重畳した男性の発声した連続数字「9053308, 3797298」における $\Delta \log F_0$ の様子を示す．この特徴量はアクセント句情報を反映することから，数字境界位置の推定に有効であり，かつ，F_0 の連続性が乏しい無声部・無音部では分散が大きくなることから，有声と無声・無音部の境界推定にも有効である．ハフ変換の最大累積投票値は得られた F_0 の信頼性を反映することから，有声部 F_0 では正の大きな値を，無声・無音部では小さな値をとる．図 2 (b) に (a) と同じ連続数字発声時の投票値の時間変化を示す．この値も $\Delta \log F_0$ とあわせて有声と無声・無音部の境界推定に有効であることがわかる．

本研究では，これら 2 つの特徴量それぞれの効果を確認するため，以下の 3 種類の韻律特徴量（**P-D**, **P-V**, **P-DV**）を作成し，それぞれを音韻特徴量と融合して用いた場合について評価実験を行った．**P-DV** のみ 2 次元ベクトルであり，他の 2 つはスカラーとなる．

P-D: $\Delta \log F_0$

P-V: 最大累積投票値

P-DV: $\Delta \log F_0$ + 最大累積投票値



(a) $\Delta \log F_0$

(b) 最大累積投票値

図 2: 日本語連続数字発声の $\Delta \log F_0$ と ハフ変換の最大累積投票値の例．SNR=20dB の白色雑音が重畳した男性の発声「9053308, 3797298」.

3.2 音韻・韻律モデルの融合

音韻と韻律情報を融合した HMM（SP-HMM: Segmental-Prosodic HMM）を CV 音節を単位としたマルチストリーム HMM で構築する．F_0 情報の遷移は CV 音節を単位として表現されていることから，音節を単位としたモデリングを行うことで，韻律情報の融合が容易となる．なお，音韻特徴量のみを用いた予備実験を本タスクにおいて行った結果，音節モデルは状態共有した triphone モデルとほぼ同等の認識性能を有していることが確認されている．

3.2.1 音節モデリング

音韻・韻律の融合モデル（SP-HMM）は，数字内部のみの音韻環境と，F_0 遷移を考慮してモデル化される．

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）．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」と表す．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．

HMM は left-to-right 型であり，状態数は各音節の持つ音素数 × 3 とした．連続数字間の長い無音区間を表現する sil モデルは 3 状態，数字間の短い無音区間を表現する sp モデルは 1 状態とした．

3.2.2 マルチストリーム HMM

融合モデル（SP-HMM）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．

$$b_j(\boldsymbol{O}_{SP}) = b_j(\boldsymbol{O}_S)^{\lambda_S} \cdot b_j(\boldsymbol{O}_P)^{\lambda_P} \qquad (5)$$

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，$\lambda_S + \lambda_P = 1.0$ とした．

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,*」「ni+i,*」「ni-i,*」のようにワイルド・カード記号「*」を用いて表記する．sil, sp モデルをあわせて合計 20 のモデルを作成する．状態数は前述した融合モデルの状態数と同じである．

(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を元に，各音節に人手によって上昇・下降・平坦の韻律ラベルを付与した時間ラベルを作成する．このラベル情報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」と表記する．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 とした．

(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．この時，韻律モデルの状態数は 1 であるので，融合モデルの全ての状態はこの 1 つの状態の混合分布と共有を行う．日本語の場合，ほとんどのモーラ（CV 音節）の $\log F_0$ のパターンが直線で表現できることが先行研究 [5] よりわかっていることから，音節モデルの各状態で同じ $\Delta \log F_0$ の分布の共有を行っている．モデル融合の様子を図 3 に示す．

4 認識実験

4.1 実験条件

使用する音声データは 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 名の数字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 [6] 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．

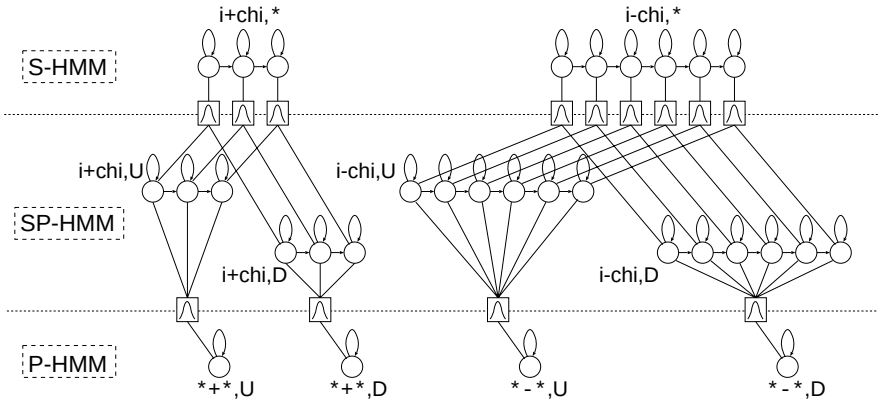


図 3: 混合分布の共有による融合モデル（SP-HMM）の構築．音韻モデル（S-HMM）は音韻特徴量のみから，韻律モデル（P-HMM）は韻律特徴量のみから学習される．

4.2 実験結果

予備実験として，音韻モデル（S-HMM）のみを用いたクリーン環境での認識実験を行ったところ，混合数 4 で最も高い数字正解精度を得た．このときの性能を baseline とする．融合モデル（SP-HMM）において，音韻ストリームの混合数を 4 で固定し，韻律ストリームの混合数を変化させて正解精度を調べたところ，混合数 4 で最高となった．そこで，以降ではそれぞれのストリームの混合数を 4 としたときの結果を示す．

表 1 に，それぞれの SNR における S-HMM と SP-HMM の数字正解精度の比較を示す．“**SP-HMM-X**” は，韻律特徴量 “**P-X**” を使ったときの融合モデルの結果を表している．音韻・韻律ストリーム重み，挿入ペナルティーは，各雑音条件ごとに最適化してあり，表 1 中の数字正解精度は 4 種類の雑音（白色・走行車内・展示場・エレベータホール）における結果を各 SNR ごとに平均したものである．全ての SNR 条件で，SP-HMM を用いることによる数字正解精度の改善が確認できる．実際には，全ての雑音種・SNR 条件で性能の改善が得られている [2]．総合的には **SP-HMM-DV** が最も良い性能を示していることから，韻律特徴量 **P-DV** の効果が最も大きいことがわかる．これは，$\Delta \log F_0$ と 最大累積投票値が相補的な役割を果たしていることを意味している．なお，最も認識性能の改善がみられたのは，10dB の展示場雑音条件で，絶対値で 4.5%（45.3 → 49.8%）数字正解精度が改善した．また，別の実験結果からは，認識性能の改善が全ての話者で観測されることがわかった．これは，提案手法が不特定話者の音声認識に対して有効であることを示している．

表 1: 音韻モデルと融合モデルの数字正解精度の比較．

SNR	**S-HMM** (baseline)	**SP-HMM -D**	**SP-HMM -V**	**SP-HMM -DV**
clean	99.3	99.6	99.4	99.4
20 dB	84.9	86.0	85.7	86.1
10 dB	53.1	54.6	55.1	55.7
5 dB	40.1	41.4	42.2	42.7

図 4 に数字正解精度の改善と韻律ストリーム重み（λ_P）の関係を示す．数字正解精度の改善は SNR 条件ごとに 4 種類の雑音における結果を平均している．また，韻律特徴量としては **P-DV** を用い，挿入ペナルティーは雑音環境ごとに最適化している．SP-HMM を用いたことによる正解精度の改善は，全ての SNR 条件で 0.0 ∼ 0.7 という広範囲に渡って観測されており，最高の改善を与える韻律ストリーム重みは，SNR に関わらず約 0.6 となった．

図 5 は最適な挿入ペナルティーと韻律ストリーム重み（λ_P）の関係を示している．白色雑音条件の各 SNR における結果を示しており，韻律特徴量には **P-DV** を用いている．音韻特徴量のみで認識を行う場合（$\lambda_P = 0$）には，SNR が小さい雑音環境になればなるほど，音韻特徴量の信頼性が乏しくなり挿入誤りが多く発生する．そのため，挿入ペナルティーを大きく設定し，認識性能を最適化する必要がある．しかし，韻律情報を用いることで，各 SNR 条件における最適な挿入ペナルティーが収束していく様子がわかる．最適な韻律ストリーム重み 0.6 付近では，韻律情報を使用しない場合に比べ，最適な挿入ペナルティーの範囲が半分以下

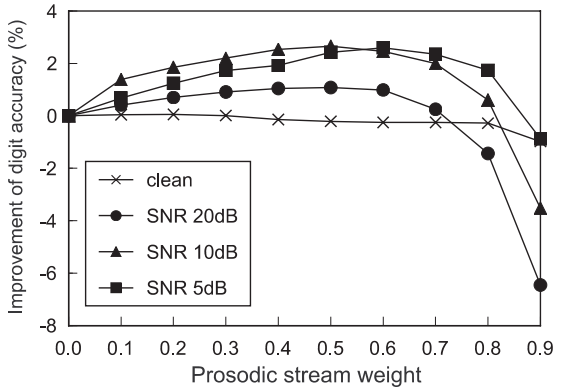


図 4: 数字正解精度の改善と韻律ストリーム重み（λ_P）の関係.

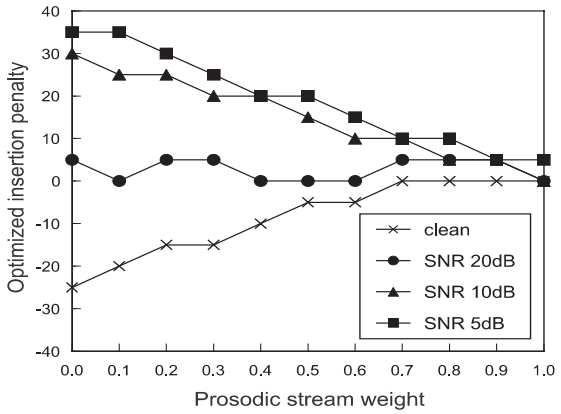


図 5: 最適な挿入ペナルティーと韻律ストリーム重み（λ_P）の関係.

となっている．この傾向は，白色雑音以外の雑音条件でも同様に観測された．これは，F_0 情報を用いたことにより，数字境界の推定精度が向上し，雑音環境下における挿入誤りの発生が，ある程度未然に防がれ，SNR の変化に対して頑健に挿入ペナルティーの最適化が可能になったものと考えられる．

そこで，S-HMM と SP-HMM の雑音環境下での数字境界の検出能力を比較するための追加実験を行った．テストセット中の，クリーンの音声とそれに雑音を重畳させた音声を，S-HMM, SP-HMM を用いて強制切り出しを行い，クリーンの音声における数字境界位置を正解として，雑音重畳音声の数字境界の検出位置の誤差（ms）を求める．4 種類の雑音全てについて誤差を計算し，SNR ごとに平均誤差を算出する．その結果，**SP-HMM-DV** を用いることで，境界検出誤差が 10dB 条件で 23.2%，5dB 条件で 52.2% 削減されることを確認した．この結果から，数字境界の検出精度の向上に韻律情報が役立っていることが確認された．

5 まとめ

本論文では，雑音に頑健な音声認識の手法として，韻律情報を利用した音声認識手法について述べ，雑音を重畳した連続数字音声の認識性能の改善について報告した．利用する韻律特徴量としては，時間–ケプストラム平面のハフ変換によって得られる $\Delta \log F_0$ と最大累積投票値を用い，それぞれの特徴量が相補的に認識性能の改善に貢献していることを確認した．また，韻律情報を用いることによって，雑音環境下における数字境界の推定精度が向上し，それによって 1) 数字正解精度が改善すること，2) 挿入ペナルティーが頑健に設定できること，を示した．

今後の課題としては，他の雑音適応化手法（例えば MLLR）などと組み合わせたときの手法の効果の確認や，最適なストリーム重みの自動設定手法の検討などが挙げられる．

参考文献

[1] 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).

[2] 岩野公司，関　高浩，古井貞熙，“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，” 音講論，vol.I, pp.23–24 (2002-9).

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

[4] 関　高浩，岩野公司，古井貞熙，“ハフ変換による雑音に頑健な基本周波数抽出法，” 情処研報，vol.2001, no.100, pp.9–14 (2001-10).

[5] 岩野公司，広瀬啓吉，“モーラを単位とした基本周波数パターンの確率モデル化とそれによるアクセント句境界の検出，” 情処学論，vol.40, no.4, pp.1356–1364 (1999-4).

[6] http://www.sunrisemusic.co.jp/dataBase/fl/noisedata01_fl.html