論文 / 著書情報
Article / Book Information

論題(和文)	スペクトルサブトラクションとハフ変換による基本周波数情報を用いた耐雑音音声認識
Title(English)	Noise robust speech recognition using spectral subtraction and Fo information extracted by Hough transformation
著者(和文)	安井 英巳, 岩野 公司, 篠田 浩一, 古井貞熙
Authors(English)	Hideki YASUI, Koji IWANO, Koichi SHINODA, SADAOKI FURUI
出典(和文)	日本音響学会2008年秋季講演論文集, Vol. , No. 1-1-2, p. 3-6
Citation(English)	, Vol. , No. 1-1-2, p. 3-6
発行日 / Pub. date	2008, 9

スペクトルサブトラクションとハフ変換による基本周波数情報を用いた耐雑音音声認識*

◎安井英己, 岩野公司, 篠田浩一, 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, 句や単語境界の推定や, 有声部と無声/無音部の境界推定に役立ち, 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.

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スペクトルの調波構造を韻律特徴量として利用した研究 [1] が報告されている. また, 岩野らは韻律情報として, 時間-ケプストラム平面をハフ変換することで得られる F_0 情報 [2] を利用し, 雑音環境下での連続数字音声認識において手法の有効性を確認している [3]. F_0 情報を利用する際に, 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し, このラベル情報と韻律特徴量を用いて, 韻律モデルの学習を行っている.

本稿では, 大語彙連続音声認識をタスクとしており, そのため人手によって韻律ラベルを付与するのが困難である. そこで, 韻律ラベルを用いて韻律モデルを作成する代わりに, 音素ラベルを用いて韻律モデルの学習を行う. そして, 雑音環境下での F_0 情報の抽出をさらに頑健に行うために, スペクトルサブトラクション法 [4] とハフ変換の併用を提案し, 雑音環境下での音声認識性能の向上を目指す.

2 F_0 情報の抽出

2.1 ハフ変換による F_0 情報の抽出

高木ら [1] はスペクトルの調波構造を特徴量として利用する際に, ケプストラムの高次ピークの強さを利用している. しかし, 雑音環境下では, 求めるべき音声の基本周波数に対応するピークと, 雑音によって発生するピークが混ざり合ってしまう. そのため, 1 フレームのケプストラム情報からでは F_0 情報を頑健に抽出できない場合が多い. そこで, 文献 [2] では, 音声の F_0 パターンの時間連続性を利用した, 雑音に頑健な F_0 情報の抽出法を提案している. この手法では, 適当な窓幅で時間-ケプストラム領域を切り出し, ハフ変換 [5] によりその中の最も優位な直線を取り出すことで, 時間連続性が考慮された, 雑音に頑健な F_0 抽出を行っている. 本研究でも, この手法を利用して韻律特徴量の抽出を行う.

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. ピークの探索範囲はケプストラムの 30 次元以上 (F_0 で 540Hz 以下) に限定する. さらに, 雑音の重畳した音声ケプストラムは, 低次部分ほどピーク値が大きくなる傾向がある. それを補正するため, 探索領域の低次部 (30〜140 次元) の d 次のケプストラムに次式で示す値 k_d を乗算しておく.

$$k_d = 0.6 + 0.4 \sin\left(\frac{d-30}{140-30} \times \frac{\pi}{2}\right) \qquad (1)$$

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) $(i = 1, \cdots, n)$ 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.

$$c = -x_i m + y_i \quad (i = 1, \cdots, n) \qquad (2)$$

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. ただし, 全ての画素について投票を行うことは効率的でないため, 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, 本実験では閾値は実験的に 0.1 とした. 次に, m-c 平面上で投票値の累積が最大となる点 $(\hat{m}, \hat{c})$ 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.

$$y = \hat{m}x + \hat{c} \qquad (3)$$

ハフ変換によって得られた直線の中点に相当するケプストラム次数を最終的に決定されたピークの箇所とし, F_0 を計算する.

---

*Noise robust speech recognition using spectral subtraction and F_0 information extracted by Hough transformation By Hideki Yasui, Koji Iwano, Koichi Shinoda and Sadaoki Furui (Tokyo Institute of Technology)

2.2 韻律特徴量

岩野らは韻律特徴量として以下の 2 つの特徴量 F, V を用いている.

F: F_0 パターンの変化情報を表す $\Delta \log F_0$. $\Delta \log F_0 \approx \Delta F_0 / F_0$ で計算され, 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.

V: $(\hat{m}, \hat{c})$ におけるハフ変換の累積投票値. F_0 の時間連続性の度合いを示す.

本稿では, 韻律特徴量として以下の 2 つの特徴量 D, N についても検討する.

D: V の変化情報. 第 t フレームにおける韻律特徴量 D_t は以下の式で与えられる.

$$D_t = \frac{\sum_{i=1}^{2} i(V_{t+i} - V_{t-i})}{10} \tag{4}$$

ここで, V_t は t 番目フレームにおける韻律特徴量 V である.

N: V を正規化した特徴量. 以下の式で与えられる.

$$N = \frac{V - V_{\min}}{V_{\max} - V_{\min}} \tag{5}$$

ここで, $V_{\max}, V_{\min}$ はそれぞれ V の各発声における最大値, 最小値である.

2.3 スペクトルサブトラクション法の併用

ハフ変換を画像処理で用いる際に, 一般的には事前に雑音低減を行うことが多い. そこで, 本稿ではより雑音に頑健な特徴量を抽出するため, 時間-ケプストラム画像にハフ変換を行う前に, スペクトルサブトラクション法によって雑音低減を行う.

$$|\hat{S}(f)|^2 = \max\{|X(f)|^2 - \alpha|\hat{N}(f)|^2, \beta|\hat{N}(f)|^2\} \tag{6}$$

ここで, $|\hat{S}(f)|^2$ は推定するパワースペクトル, $|X(f)|^2$ は観測信号のパワースペクトル, $|\hat{N}(f)|^2$ は無声音区間から推定される雑音のパワースペクトル, α, β はサブトラクション係数である. 本実験では $\alpha = 1, \beta = 0.24$ とした.

提案手法の手順として, まずパワースペクトル領域においてスペクトルサブトラクションを行う. 次に, スペクトルサブトラクション後のスペクトルをケプストラムに変換し, 時間-ケプストラム画像を取り出し, ハフ変換を行う.

3 音韻・韻律情報の融合

3.1 音韻・韻律特徴量の融合

音韻特徴量は, MFCC12 次元・ΔMFCC12 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 である. 入力音声ごとにケプストラム正規化 (CMN) を行っている.

韻律特徴量は, 効果の比較のため (V, F), (V, D), (N, D) の 3 通りについて検討する.

音韻特徴量と韻律特徴量は同じフレーム周期であり, 両者をフレーム毎に結合して, 合計 27 次元の融合特徴ベクトルを作成する.

3.2 音韻・韻律モデルの融合

音素を単位とした音韻・韻律の融合モデル (SP-HMM: Segmental-Prosodic HMM) を構築する. 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. 音韻特徴量 $\boldsymbol{O}_S$ と韻律特徴量 $\boldsymbol{O}_P$ を 2 つのストリームに分け, それぞれから得られる 尤度を重み付けし, 合わせることで, 融合特徴量 $\boldsymbol{O}_{SP}$ の尤度を得る. 状態 j における尤度 $b_j(\boldsymbol{O}_{SP})$ は以下の式で与えられる.

$$b_j(\boldsymbol{O}_{SP}) = b_j(\boldsymbol{O}_S)^{\lambda_S} \cdot b_j(\boldsymbol{O}_P)^{\lambda_P} \tag{7}$$

ここで, $b_j(\boldsymbol{O}_S), b_j(\boldsymbol{O}_P)$ はそれぞれ状態 j での音韻特徴量 $\boldsymbol{O}_S, \boldsymbol{O}_P$ の尤度である. 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.

具体的には, 以下のような手順で融合モデルを構築する.

(1) 音韻特徴量のみを用いて音韻モデル (S-HMM: Segmental HMM) を学習する. 音韻モデルには状態共有化された HMM を使用する.

(2) 作成した音韻モデルを用いて, 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. このラベルには, 音素の各状態ごとに時間情報が付与されている.

(3) 時間ラベルを用いて, 学習データから抽出した韻律特徴量により, 各状態ごとに韻律モデル (P-GMM: Prosodic GMM) の学習を行う.

(4) 得られた 2 つのモデル (音韻モデル, 韻律モデル) を状態毎に融合し, 音韻-韻律マルチス

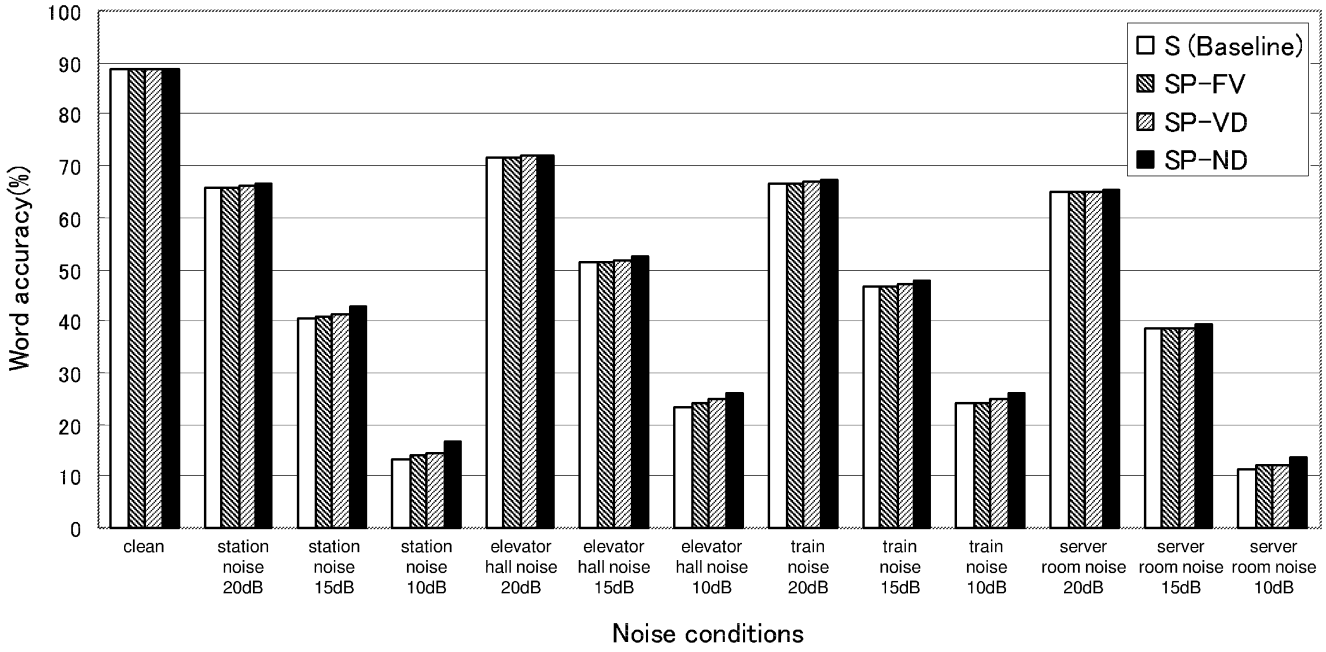

Fig. 1　3 種類の融合モデルの性能 (スペクトルサブトラクションなし)

トリーム HMM を構築する．なお, 融合されたマルチストリーム HMM の状態遷移行列には音韻モデルのものを用いる．

4　認識実験

4.1　実験条件

データベースとして, 新聞記事読み上げコーパス (JNAS) を使用した．実験には JNAS の 260 話者 (男女各 130 名) を学習データとして用い, JNAS の 46 話者 (男女各 22 名) を認識データとして用いた．話者 1 名あたり約 100 文を読み上げている．

モデルの学習は静かな環境下の (clean な) 音声を用いる．認識実験では, clean な音声に加え, 電子協騒音データベースの駅・エレベータホール・列車・計算機室雑音の計 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 10, 15, 20dB とした．

m-c 平面への投票を行う際は $-20 < m < 20$, $-300 < c < 300$ の範囲に m が 0.2, c が 0.5 刻みで量子化されたビンを 200×1200 個用意して投票した．

言語重み, 挿入ペナルティーについては clean な音声における最適値に定め, 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．

4.2　実験結果

4.2.1　韻律特徴量の違いによる認識性能の比較

まず, 韻律特徴量の違いによる融合モデルの認識性能の比較実験を行った．各雑音条件における融合モデル (SP) と音韻モデル (S) の単語正解精度を Fig. 1 に示す．図中の SP-FV, SP-VD, SP-ND はそれぞれ韻律特徴量として, (F, V), (V, D), (N, D) を利用して融合モデルを作成していることを示している．

今回の実験において, SP-FV は耐雑音性の向上に対してあまり効果がなかった．原因としては韻律ラベルを用いた韻律モデルの学習を行わなかったためと考える．SP-VD については全ての雑音条件において効果があった．SP-ND についても全ての雑音条件において効果があった．また, SP-VD との比較においても全ての雑音条件下で性能が上回っている．正規化を行ったことにより, 雑音の影響による特徴量のミスマッチが軽減されたためと考えられる．

4.2.2　スペクトルサブトラクション法の併用による認識性能の比較

次に, スペクトルサブトラクション法と併用による認識性能の比較実験を行った．この際に, 音韻特徴量, 韻律特徴量ともにスペクトルサブトラクション法を行っている．先の実験の結果より (N, D) を韻律特徴量として利用する．各雑音条件における融合モデルと音韻モデルの単語正解

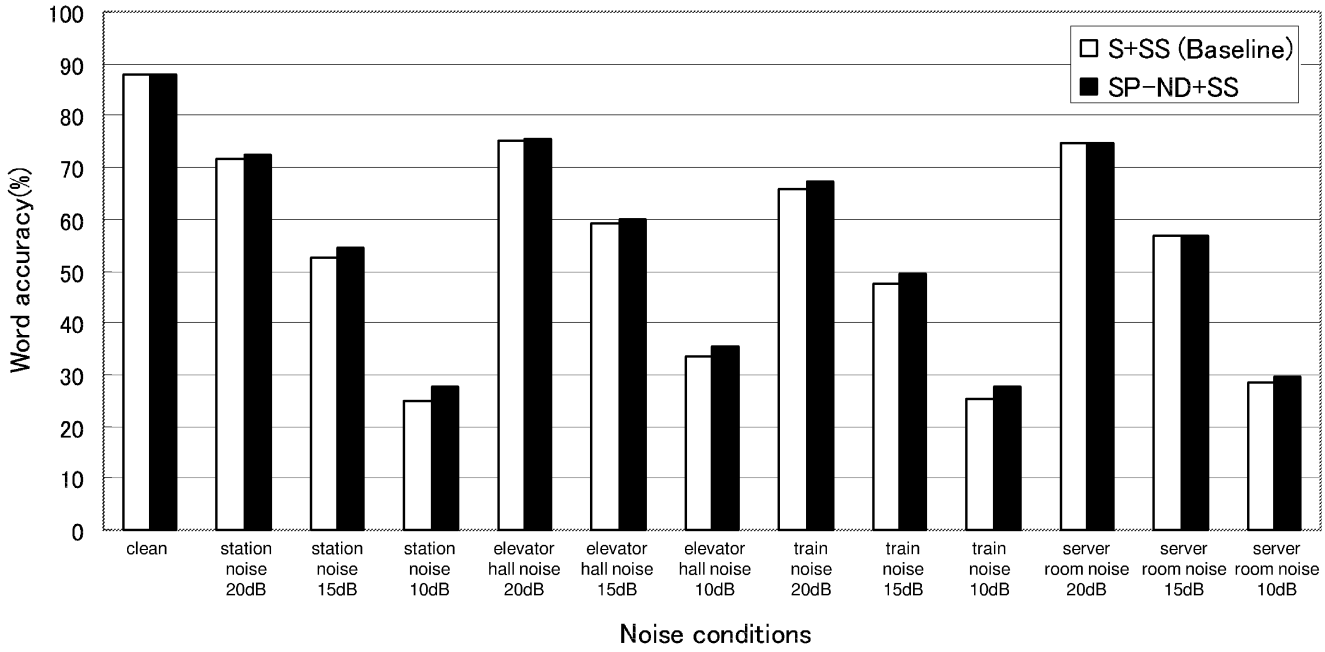

Fig. 2　韻律特徴量とスペクトルサブトラクションを併用する手法の性能

Table 1　1 発声 (12.6 秒) に対するハフ変換の計算時間. CPU2.4GHz の計算機を使用.

ビンの数	計算時間 (秒)
200×1200	192
100×600	42

精度を Fig. 2 に示す. 図中の SP-ND+SS はスペクトルサブトラクション法を併用した韻律特徴量を利用して融合モデルを作成していることを示している.

提案手法では全ての雑音条件において効果があった. 提案手法による認識性能の改善が最も得られたのは, 10dB の駅雑音が重畳した音声を認識したときであり, 2.6 ポイントの改善がみられた.

4.2.3　ハフ変換の計算量削減

先の実験条件ではハフ変換により F_0 情報を抽出する際の計算時間が非常に大きくなってしまう. そこで, 投票平面におけるビンの数を減らすことによって計算時間の削減を行った.

Table 1 に 1 発声の音声ケプストラムからハフ変換により F_0 情報を抽出する際のビンの数と計算時間を示す. ビンの数を減らしたことにより, 計算時間の削減を行うことができたが, 未だ計算時間が大きい. また, 認識性能の劣化が多少みられた. 15dB の駅雑音が重畳した音声を認識したときに単語正解精度の劣化が最大 0.15 ポイントとなった.

5　おわりに

本稿では, 音声認識の耐雑音性の向上を目的として, ハフ変換とスペクトルサブトラクション法の併用を提案し, 連続単語音声認識において提案手法の有効性を確認した.

今後の課題としては, さらなるハフ変換の計算量削減やスペクトルサブトラクション法以外の耐雑音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] 高木 他, “音声認識のためのスペクトルの調波構造の利用,” 秋季音講論, pp. 3-4 (1997).
[2] 関 他, “ハフ変換による雑音に頑健な基本周波数抽出法,” 情報処理学会研究報告, vol. 2001, no. 100, pp. 9-14 (2001).
[3] 岩野 他, “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,” 秋季音講論, pp. 23-24 (2002).
[4] S.F.Boll, “Suppression of Acoustic Noise in Speech Using Spectral Subtraction,” IEEE Trans.ASSP, vol. ASSP-27, no. 2, pp. 113-120 (1979).
[5] P.V.C.Hough, U.S. Patent #3069654(1962).