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

2008－SLP－72 (11)
2008／7／19

Δ-logF0に基づく韻律特徴量を用いた耐雑音性の高い声調認識

木田 祐介　酒井 優　益子 貴史　河村 聡典

東芝 研究開発センター
〒212-8582 川崎市幸区小向東芝町1
e-mail: yusuke.kida@toshiba.co.jp

あらまし　中国語やタイ語などの声調言語では，声調を認識するために，音声の基本周波数に基づく韻律特徴量が用いられている．しかし，実環境雑音下においては，基本周波数の推定精度の劣化が，声調認識精度に悪影響を与える問題がある．また，話者やイントネーションによる基本周波数の変動を正規化する処理によって，韻律特徴抽出の実時間性が損なわれる問題もある．これらの問題を解決するため，我々は基本周波数の対数の時間変化量 (Δ-logF0) に基づく韻律特徴量に着目する．Δ-logF0 は，基本周波数やその対数に比べて話者性やイントネーションの影響を受けにくく，遅延を伴う正規化処理を行う必要がない．そのため，遅延時間の小さい韻律特徴量の抽出が可能である．さらに我々は，耐雑音性の高い Δ-logF0 推定手法として，対数周波数軸上の周波数スペクトルのシフト量推定に基づく手法を提案する．マンダリンの認識実験の結果，提案手法により推定した Δ-logF0 に基づく韻律特徴量が，従来の logF0 に基づく韻律特徴量に対して高い耐雑音性を示すことを確認した．

Noise-Robust Tone Recognition using Δ-logF0 based Prosodic Feature

Yusuke Kida　Masaru Sakai　Takashi Masuko　Akinori Kawamura

Toshiba R&D Center
1, Komukai-Toshiba-cho, Saiwai-ku, Kawasaki, 212-8582, Japan

Abstract　This paper presents a noise-robust prosodic feature based on Δ-logF0. In conventional ASR systems of tonal languages such as Mandarin and Thai, fundamental frequency (F0) is used as a prosodic feature. However, noise-robust F0 estimation has still been a difficult task, and sufficient accuracy has not been obtained in noisy environment. Another problem is that the normalization of F0 generates delay of feature extraction. Normalization of F0 is essential because F0 is highly affected by speaker's personality and intonation. To solve these problems, we focus on Δ-logF0 based prosodic feature. Δ-logF0 does not need to be normalized because it is not affected by above factors. In this paper, we propose a new algorithm which estimates Δ-logF0 directly based on shift estimation on log-scale frequency spectrum. The experimental results show that the proposed method improves the noise-robustness of Mandarin tone recognition in comparison with conventional logF0 based prosodic feature.

1 はじめに

中国語やタイ語などの声調言語では，同じ音をもつ語句であっても，音の高低のパターン (声調) に応じてその意味が区別されるため，音に加えて声調も認識する必要がある．例えば，近年注目されている携帯電話やカーナビへの音声認識技術の応用では，人名や地名などの固有名詞を認識するために，固有名詞に含まれる同音かつ異なる声調をもつ語句を区別するための声調認識が必須である．

声調認識が可能な音声認識システムとして，MFCC に代表されるスペクトル包絡の特徴量と，音の高低やその変化に関する特徴量 (本稿では韻律特徴量と呼ぶ) を併用し，音と声調を同時に認識するシステムが提案されている [1][2]．これらのシステムでは，自己相関法 [3] などを用いて推定した音声の基本周波数 (F0) やその対数 (logF0)，及びそれらの一次微分 (Δ)· 二次微分 (ΔΔ) を韻律特徴量として用いており，主にクリーンな音声を対象とした連続音声認識において一定の成果を達成している．

しかし，携帯電話やカーナビへの応用を考えた場合，基本周波数に基づく韻律特徴量には耐雑音性の点で問題がある．雑音の影響による基本周波数の推定精度の劣化は，声調認識精度に直接影響を与える．しかし，実環境雑音下における基本周波数の高精度な推定は非常に難しい問題であり，現在でも決定的な解決には至っていない．また，話者 (特に性別) やイントネーションによる基本周波数の変動を正規化する処理によって，韻律特徴抽出処理の実時間性が損なわれる問題もある．例えば文献 [2] では，実時間処理に適した正規化手法として，基本周波数の移動平均による正規化を提案している．しかし，ある時刻の基本周波数を正規化するための移動窓を当該時刻の近傍に設定する場合，窓長に応じた遅延時間の発生が避けられない．

これらの問題点を解決するためには，実環境雑音下でも高い認識精度を発揮し，かつ，できるだけ小さい遅延時間で抽出可能な韻律特徴量を開発する必要がある．そこで我々は，基本周波数の対数の時間変化量 (Δ-logF0) に基づく韻律特徴量に着目する．Δ-logF0 は，基本周波数やその対数に比べて話者性やイントネーションの影響を受けにくく，遅延を伴う正規化処理を行う必要がない．さらに我々は，耐雑音性の高い Δ-logF0 推定手法として，対数周波数軸上の周波数スペクトルのシフト量推定に基づく手法を提案し，声調認識の耐雑音性の向上を試みる．

我々は，中国語普通話 (マンダリン) の声調認識を対象として，本稿で提案する韻律特徴量の声調認識精度を評価する．マンダリンには四声と軽声の 5 種類の声調が存在するが，軽声を除く四声には，音の高低だけでなくその時間変化パターンにも差異があるため，Δ-logF0 に基づく韻律特徴量でも声調認識が可能だと考えられる．

以下，本稿ではまず，実験に用いる 2 つの韻律特徴量の抽出手順について述べる．次に，本稿で提案する耐雑音性の高い Δ-logF0 推定手法の原理とアルゴリズムを説明する．最後に，提案手法により推定した Δ-logF0 に基づく韻律特徴量の耐雑音性を評価するための，マンダリンの音声認識実験結果を報告する．

2 韻律特徴量の抽出手順

本章では，本稿で実験に用いる 2 つの韻律特徴量の抽出手順を述べる．1 つは，正規化 logF0 にその 1 次微分 (Δ)·2 次微分 (ΔΔ) を加えた 3 次元の特徴量であり，本稿ではこれを正規化 logF0 特徴と呼ぶ．もう 1 つは，Δ-logF0 に ΔΔ を加えた 2 次元の特徴量であり，本稿ではこれを Δ-logF0 特徴と呼ぶ．なお，本稿の音声認識システムでは，これらの韻律特徴量をスペクトル特徴量と結合した特徴量を用いて，HMM による音響モデルの学習 · 認識を行う．

2.1 正規化 logF0 特徴の抽出手順

正規化 logF0 特徴の抽出手順を図 1(a) に示す．まず，logF0 推定処理と有声／無声判定処理により，有声フレームにおける logF0 の推定値を求める．次に，話者性やイントネーションの影響を排除するために，logF0 の正規化処理を行う．ここでは，着目フレームを中心とする移動窓で計算した logF0 の移動平均値により正規化を行う．さらに，Δ·ΔΔ の計算を行う．最後に，無声フレームの補間処理を行う．本稿では，logF0·Δ·ΔΔ のそれぞれの値域と同程度の値域を持つ乱数で無声部を補間する．

2.2 Δ-logF0 特徴の抽出手順

Δ-logF0 特徴の抽出手順を図 1(b) に示す．正規化 logF0 特徴の抽出手順との相違点は以下の 2 点である．1 つは，logF0 推定処理を Δ-logF0 推定処理に置き換えた点である．もう 1 つは，Δ-logF0 は話者性やイントネーションの影響を受けにくいため，正規化処理を不要とした点である．これにより，正規化による遅延時間のない特徴抽出が可能となる．

図 1(b) の Δ-logF0 推定処理は，logF0 の推定とその 1 次微分の計算と実施してもよい．しかしその場合は，雑音の影響により logF0 の推定精度が劣化すると，それに応じて Δ-logF0 の推定精度も劣化してしまう．そこで本稿では，logF0 推定処理を介さず

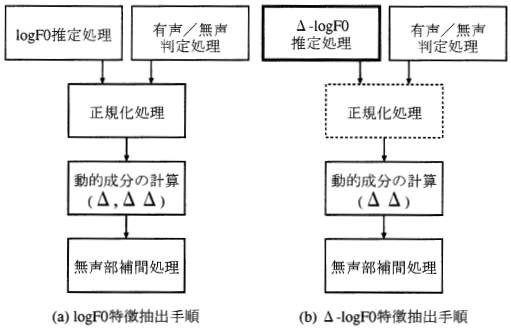

図 1: 韻律特徴量の抽出手順

に，音声から Δ-logF0 を直接かつ高精度に推定する方法を提案する．

3 Δ-logF0 推定の提案手法

本章では，耐雑音性の高い Δ-logF0 推定に関する提案手法の原理とアルゴリズムを説明する．

3.1 原理

音声の調波構造は，基本周波数とその倍音周波数から構成される．図 2 に例示するように，各倍音周波数の対数周波数軸上での時間変化量は，基本周波数の対数周波数軸上の時間変化量，すなわち Δ-logF0 に等しくなる．このとき，logF0 の時間変化は，対数周波数軸上での調波構造の平行移動として観測される．このことから，Δ-logF0 を，調波構造の対数周波数軸上での移動量として推定できる．このように，各調波成分の時間変化量の一貫性を評価することで，雑音の影響を受けて調波構造の一部が不明瞭な場合でも，高い推定精度を得ることが期待される．

そこで提案手法では，対数周波数軸上の周波数スペクトル (対数周波数スペクトル) に対して，異なる 2 つのフレーム間の相互相関関数を算出し，対数周波数軸上でのスペクトルの移動量 (シフト量) を推定することで Δ-logF0 を推定する．本稿では，この提案手法を SELF (Shift Estimation on Log-Frequency domain) と呼ぶ．次節では，その具体的なアルゴリズムを説明する．

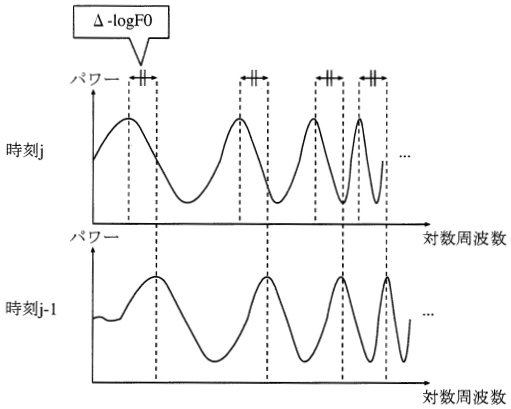

図 2: 対数周波数軸上における周波数スペクトル

3.2 アルゴリズム

3.2.1 対数周波数スペクトルの計算

まず，音声信号の周波数分析により周波数スペクトルを求めた後，周波数軸を対数化し，対数周波数軸上で周波数点幅を等間隔にとるリサンプリング処理により対数周波数スペクトルを得る．なお本稿では，音声の調波構造に着目するために，LPC 分析によって包絡成分を除去した周波数スペクトルを用いた．

3.2.2 相互相関関数

次に，フレームごとに，隣接する 2 つのフレームの対数周波数スペクトルに対して，以下の式 (1) を用いて相互相関関数を計算する．

$$C_t(n) = \sum_f S_t(f) \cdot S_{t-1}(f+n) \qquad 0 \le f < F \tag{1}$$

ここで，$S_t(f)$ はフレーム t の対数周波数スペクトルにおける周波数点番号 f の周波数成分のパワーである．また，$C_t(n)$ は周波数点番号を単位とするシフト量 n に対する相互相関値である．なお，F は対数周波数スペクトルの解像度であり，本稿では 2048 とした．

次に，以下の式 (2) を用いて，着目フレーム t の周辺にあるフレーム $t'(t-a+1 \le t' \le t+a)$ で算出した相互相関値の和を計算する．なお，本稿では $a = 2$ とした．

$$C'_t(n) = \sum_{t'=t-a+1}^{t+a} C_{t'}(n) \tag{2}$$

ある短時間区間内で Δ-logF0 が一定と仮定すると，区間内の各フレームで算出した相互相関値は，ある共通のシフト量に対して高い値を得ると考えられる．そのため，Δ-logF0 がある短時間区間内で定常である場合には，複数のフレームで算出した相互相関値の和を用いることで，さらに推定精度を高められると期待される．

3.2.3 シフト量の推定

次に，着目フレーム t の対数周波数スペクトルのシフト量 d_t を，以下に示す式 (3) により推定する．

$$d_t = \underset{n}{\operatorname{argmax}}\, C'_t(n) \tag{3}$$

以上の処理によって求めたシフト量を Δ-logF0 の推定値として，2.2 節で述べた処理を適用することによって得られる Δ-logF0 特徴を，本稿における提案特徴量とする．

4 音声認識実験

提案特徴量の耐雑音性を評価するため，マンダリンの音声認識実験を行った．本章では音声分析条件と音響モデルについて述べた後，最適音節列認識実験と人名認識実験の 2 種類の実験について述べる．

4.1 音声分析条件

本実験では，スペクトル特徴量として，CMN を適用した 12 次元の MFCC と 1 次元の正規化対数エネルギー，及びそれぞれの Δ·ΔΔ の計 39 次元の特徴量を使用した．なお，特徴抽出時のフレーム長は 25 ミリ秒，フレーム周期は 10 ミリ秒とした．

また，ベースラインとなる正規化 logF0 特徴では，Ghulam らの手法 [4] により推定した F0 を対数化することで logF0 を得た．この手法では，まず，音声信号を帯域分割し，それぞれの帯域信号に対して波形処理を行った後で，自己相関関数を計算する．さらに，帯域ごとに求められた自己相関関数の和関数の最大値を与える時間遅れに基づき F0 を推定する．本稿では，この F0 推定手法を SACF (Summary Auto-Correlation Function) と呼ぶ．なお，[4] では F0 推定後に推定誤りを訂正するトラッキング処理が行われているが，本実験では省いた．

SACF では，各帯域の自己相関関数の和関数の最大値に対するしきい値処理によって有声／無声判定処理を行った．一方，SELF では，数式 (3) で得られる相互相関関数の最大値に対するしきい値処理によって有声／無声判定処理を行った．SACF·SELF ともに，学習 · 評価データには含まれないクリーンな発話に対して，誤受理率と誤棄却率が等しくなる値をしきい値として選択した．

4.2 音響モデル

音響モデルには，3 状態で left-to-right 型の状態共有トライフォン HMM を用いた．状態数は 4000，混合数は 32 とした．本実験では，異なる SNR で実走行雑音を重畳した音声により学習したマルチ SNR モデルを用いて認識を行った．ここで，重畳を行う際の SNR は 5，10，20dB の 3 種とした．

本実験では，音節を声母 (Initial) と韻母 (Final) に分割し，それぞれを単位として音響モデルを作成した．ここで，Final のみを声調ごとに分割した [5]．

音響モデルの学習に用いた音声データはマンダリンの短文発声データで，話者数は 160(男女各 80)，総発話数は 81969 発話である．

4.3 最適音節列認識実験

声調言語に対する音声認識システムは，音と声調の両方を認識する必要がある．そこで，まず最適音節列認識実験により，音と声調の両方が正しく認識された場合を正解とした認識精度を評価した．評価データは 4~6 音節発声で，学習データに重畳した実走行雑音をクリーン音声に重畳することで作成した．話者数は 12(男女各 6)，総発話数は 888 とした．なお，本実験では音節数を既知として認識を行った．

4.3.1 従来の韻律特徴量のベースライン評価

表 1 に実験結果を示す．表は，スペクトル特徴に各韻律特徴を加えた特徴量の音節正解精度を，評価データの SNR ごとに示している．表の"MFCC" は韻律特徴量を用いない 39 次元のスペクトル特徴量を示す．また，"MFCC+logF0+Δ+ΔΔ" は MFCC に正規化 logF0 特徴を加えた特徴量を，"MFCC+Δ-logF0+ΔΔ" は MFCC に Δ-logF0 特徴を加えた特徴量を示す．正規化 logF0 特徴については，正規化に用いる移動窓長を 110~1610ms の 5 パターン用意してそれぞれ実験を行った．

表中の"MFCC" と"MFCC+logF0+Δ+ΔΔ" を比較すると，MFCC に正規化 logF0 特徴を加えることによる認識精度の向上が見られた．これは，韻律特徴量の導入により，声調が認識できなかった音節が正しく認識されたためだと考えられる．また，正規化に用いる移動窓長により正規化 logF0 特徴を比較すると，窓長に対する認識精度の変化は小さかったものの，410ms の窓長において平均的な精度が最も

高かった．410ms の窓長において認識精度が高かった理由は，話者の個人性とイントネーションの影響を最もうまくキャンセルできたためだと考えられる．

4.3.2 logF0 特徴と Δ-logF0 特徴の比較

次に，SACF で推定した logF0 の一次微分によって求めた Δ-logF0 に対して 2.2 節で述べた処理を適用することで得た Δ-logF0 特徴と，正規化 logF0 特徴を比較した．その結果，Δ-logF0 特徴は最も認識精度の高かった 410ms 窓の正規化 logF0 特徴に対してはやや精度の劣化が見られたものの，その精度差は小さかった．このことから，Δ-logF0 特徴により正規化 logF0 特徴と同程度にマンダリンの声調認識が可能であることが示された．Δ-logF0 特徴の認識精度が正規化 logF0 特徴に比べて低かった理由は，音高に関する情報を失ったためだと考えられる．また，Δ-logF0 特徴は認識精度においては正規化 logF0 特徴に及ばなかったものの，正規化に伴う遅延時間を必要としない点で正規化 logF0 特徴に対して優位である．

4.3.3 提案特徴量による耐雑音性の向上

最後に，SACF に基づく Δ-logF0 特徴と SELF に基づく提案特徴量を比較すると，提案特徴量はいずれの SNR においても SACF に基づく特徴量より高い認識精度を示した．特に，5dB で提案特徴量による精度の向上が見られた．また，正規化 logF0 特徴と比較すると，提案特徴量は SNR が 10dB 以上では 410ms 窓の正規化 logF0 特徴に比べてやや認識精度が劣化したものの，5dB で高い精度を示した．以上の結果から，提案特徴量は低 SNR に対する頑健性が高いのではないかと考えられる．

4.4 人名認識実験

提案特徴量による声調認識の耐雑音性の向上をさらに検証するため，同音異声調の単語を含む人名の認識実験を行い，同音異声調の単語間の認識精度を評価した，評価データはマンダリンの 2 音節人名 (first name) の発声で，アイドリング時と高速道路走行時の車内でそれぞれ収録を行った．話者数は 40，総発話数は約 2000 とした．発声内容は 50 単語とした．なお，本実験では正規化に用いる移動窓長を 410ms として正規化 logF0 特徴を求めた．これは，最適音節列認識実験の結果，410ms 窓の logF0 特徴が最が高い精度を示したためである．

4.4.1 音の認識に与える影響

まず，提案特徴量の導入が音の認識に悪影響を及ぼさないことを検証するため，同音異声調の人名を含まない 50 単語を認識語彙として実験を行った．

アイドリング時と高速道路走行時の評価データに対する 50 単語認識実験の結果を，それぞれ表 2，3 の”50 単語” に示す．表の”WER” が単語誤り率を示している．実験結果から，提案特徴量を含めて MFCC に韻律特徴量を加えたいずれの特徴量も，MFCC に対する認識精度の劣化が見られなかった．このことから，提案特徴量の導入が音の認識に悪影響を及ぼさないことが確認された．

4.4.2 同音異声調の単語間の認識誤りの改善

次に，提案特徴量による声調認識の耐雑音性の向上を検証するため，前述の 50 単語に同音異声調の人名を加えた 107 単語を認識語彙として実験を行った．表 2，3 の”107 単語” に 107 単語認識実験の結果を示す．表の WER に対して，”TCER” は同音異声調の単語間の認識誤り率を，”OER” はその他の認識誤り率を表しており，$WER = TCER + OER$ である．本実験では，この 3 つの評価尺度のうち TCER に着目することで，声調認識の評価を行った．

表 2 より，韻律特徴を含む特徴量の TCER は，いずれもアイドリング時において MFCC より低かった．このことは，韻律特徴量の導入により，同音異声調の単語間の認識誤りが削減されたことを示している．また，提案特徴量と正規化 logF0 特徴を比較すると，提案特徴量の TCER は正規化 logF0 特徴と同程度であった．

一方，高速道路走行時には，MFCC に正規化 logF0 特徴を導入することによる TCER の改善は小さかった．このことから，従来の韻律特徴量の耐雑音性が十分でないことがわかる．それに対し，提案特徴量の導入により，正規化 logF0 特徴に対して TCER を 15.8%から 10.9%まで削減できた．

以上の結果から，提案特徴量により耐雑音性が高く，かつ，正規化に伴う遅延時間のない韻律特徴量が実現できたと言える．

5 おわりに

本稿では，耐雑音性に優れ，かつ，遅延時間を必要としない韻律特徴量の実現を目的として Δ-logF0 に基づく韻律特徴量に着目した．さらに，対数周波数軸上の周波数スペクトルのシフト量推定に基づく Δ-logF0 推定手法を提案した．最適音節列認識と人名認識の 2 種類のマンダリン認識実験を行った結果，

表 1: 最適音節列認識における音節正解精度 (%)

特徴量	推定手法	窓長	5dB	10dB	20dB
MFCC	-	-	38.7	42.4	43.5
MFCC+logF0+Δ+ΔΔ	SACF	1610ms	48.7	53.2	55.5
		810ms	49.3	54.1	55.5
		410ms	49.1	**54.6**	**56.1**
		210ms	48.7	53.9	55.5
		110ms	47.6	52.5	54.3
MFCC+Δ-logF0+ΔΔ	SACF	-	47.5	52.6	54.9
	SELF	-	**51.0**	53.7	55.3

表 2: 人名認識における単語誤り率：アイドリング時 (%)

特徴量	推定手法	窓長	50 単語	107 単語		
			WER	WER	TCER	OER
MFCC	-	-	6.6	29.4	22.4	7.0
MFCC+logF0+Δ+ΔΔ	SACF	410ms	4.1	**17.6**	13.1	4.5
MFCC+Δ-logF0+ΔΔ	SACF	-	**3.8**	19.7	15.1	4.6
MFCC+Δ-logF0+ΔΔ	SELF	-	4.6	17.9	**12.7**	5.2

表 3: 人名認識における単語誤り率：高速道路走行時 (%)

特徴量	推定手法	窓長	50 単語	107 単語		
			WER	WER	TCER	OER
MFCC	-	-	31.7	51.8	18.4	33.4
MFCC+logF0+Δ+ΔΔ	SACF	410ms	29.4	47.9	15.8	32.1
MFCC+Δ-logF0+ΔΔ	SACF	-	29.8	50.9	17.0	33.9
MFCC+Δ-logF0+ΔΔ	SELF	-	**29.3**	**43.8**	**10.9**	32.9

提案手法により推定した Δ-logF0 に基づく韻律特徴量が，従来の logF0 に基づく韻律特徴量に対して高い耐雑音性を示すことを確認した．

今後は，さらなる耐雑音性の向上を目指すとともに，広東語やタイ語のように，音高の動的な変動パターンが等しく，音高そのものにより識別される声調の組を有する声調言語に対しても頑健な認識を実現するための方法を検討する．

謝辞

本研究は，経済産業省「情報家電センサー・ヒューマンインターフェースデバイス活用技術開発 < 音声認識基盤技術の開発 >」プロジェクトの一環として実施されたものである．

参考文献

[1] C.J. Chen, et al., "New methods in continuous Mandarin speech recognition," Proc. INTERSPEECH, pp.1543-1546, 1997.

[2] HC Huang, et al., "Pitch tracking and tone features for Mandarin speech recognition," Proc. ICASSP, pp.1523-1526, 2000.

[3] L.R. Rabiner, "On the use of autocorrelation analysis for pitch detection," IEEE Trans. Acoustics, Speech, and Signal Processing, Vol.25, No.1, pp.24-33, 1977.

[4] M. Ghulam, et al.,"A noise-robust feature extraction method based on pitch-synchlonous ZCPA for ASR," Proc. INTERSPEECH, pp.133-136, 2004.

[5] X. Wang, et al.,"Low complexity Mandarin speaker-independent isolated word recognition," Proc. INTERSPEECH, pp.1589-1592, 2002.