音声言語情報処理 25−16
（1999. 2. 6）

基本周波数とケプストラムによる多言語音声の分類

木内俊一　山本幹雄　板橋秀一

筑波大学

〒305-8573 茨城県つくば市天王台 1-1-1
TEL：0298-53-5382
E-mail：{kiuchi, myama, itahashi}@milab.is.tsukuba.ac.jp

あらまし　2種類の方法を用いて、多言語音声の識別を行った。音声データには約45秒の自発的発話、各言語50人の10言語を用いた。第1の方法は韻律情報として基本周波数を用い、その時間的変化パターンを指数関数と直線で近似し、そのパラメータを用いる方法であり、第2は音韻情報としてケプストラムをパラメータとして用いたHMMによる方法である。閉じた実験では音韻，韻律それぞれ96.7%，36.7%の識別率がこれら2つの方法を併合することにより、97.3%になった。また、開いた実験ではそれぞれ55.5%，25.5%であったが、併合することで60.0%に向上した。

Classification of Multi-language Speech based on Fundamental Frequency and Cepstrum

Toshikazu KIUCHI　Mikio YAMAMOTO　Shuichi ITAHASHI

University of Tsukuba

1-1-1 Tennodai, Tsukuba-shi, Ibaraki 305-8573 Japan

Abstract　Multi-language speech classification was conducted with two methods. Speech data of 45 seconds spontaneous utterances, spoken by 50 speakers for each of ten languages, were used. In the first method, fundamental frequency contour was used as prosodic information and its trajectory approximated by exponential functions and lines and their parameters were used for discrimination. The second method is based on HMM using cepstral parameters as segmental information. It is shown that better discrimination is obtained by combining the two methods.

1 はじめに

数多くの言語が世界中で使われており、それぞれの言語間の円滑なコミュニケーションをはかるために、自動言語識別は重要な課題の一つとなっている。現在の多くの音声認識システムでは単一言語のみを受け入れるように設計されているが、将来は多言語を識別する音声認識システムが必要とされる。電話の場合を例にとると、言語識別システムは、通話者を言語ごとに分類し、その話者に適した音声認識システムまたはオペレータへ接続するために用いることができる。旅行情報・言語の翻訳・電話情報・テレフォンショッピング・銀行取引・株式取引など、いずれの場面でも短い発話で高性能の識別力を発揮する必要がある。

音声の特徴は大きく音韻的特徴と韻律的特徴の2つに分けられる。通常、言語識別では音韻情報が用いられるが、本研究では音韻情報による言語識別だけでなく、韻律情報を用いた識別も行った。さらに、両手法を併合することにより、単独の場合よりもよい結果が得られることを示す。

2 音声データ

音声データとしては、"Multi-language Telephone Speech Corpus"[1] を用いた。これは Oregon Graduate Institute において、自動言語識別と多言語音声認識を目的に作成されたものである。商用電話回線を用いて 11 言語が収集された。内容は語彙固定、話題指定、自発的発話に分かれている。音声データはサンプリング周波数 8kHz、量子化精度 14bit でディジタル化されている。

以下では、このデータベースから英語、イラン語、フランス語、ドイツ語、日本語、韓国語、中国語、スペイン語、タミール語、ベトナム語、の 10 言語を選択し、男性話者から 1 話者当たり約 45 秒の自発的発話を利用した。言語識別するための学習用として各言語 30 人、テスト用として各 20 人で、計 500 話者を使用した。

3 音声の分析

最初に、音韻情報を用いた言語識別と韻律情報を用いた識別をそれぞれ独立に行い、その後、2 手法を併合して識別を行った。

3.1 音韻情報を用いた分析方法[4]

3.1.1 音声の学習

2節で示した、各言語 30 人の音声話者を学習として用い、各言語に 1 つずつのエルゴディック HMM を作成した。今回の実験では識別率の比較のため、状態数を 16,32,64 の 3 通りで行った。なお尤度を高くするためにモデルの学習を 20 回繰り返した。また、学習前の初期モデルとしては音素領域全体をカバーするために、ベクトル量子化で得られた値を用いた。

音声データの分析条件はサンプリング周波数 8kHz、フレーム周期は 10ms、ハミング窓は 25ms で、特徴パラメータは 12 次のメルケプストラムと 12 次の回帰係数 (Δメルケプストラム)、1 次の Δパワーである。また、HMM のパラメータは Baum-Welch アルゴリズムを用いて推定される。

なお、今回の音韻情報を用いた実験における、音声の分析、HMM の学習、認識、評価には HTK(HMM Toolkit)[3] を使用した。

3.1.2 言語の識別

ここでは、学習用に用いた音声データを識別する「閉じた実験」とそうでないデータ (テストデータ) を識別する「開いた実験」の 2 種類を行う。

入力音声に対する尤度は、各 HMM を用いてフレームごとに計算され、累積される。この尤度はすべての言語モデルに対して計算され、尤度が最も大きいモデルの言語を入力の言語であると判定する。

3.2 韻律情報を用いた分析方法[5],[6]

音声データを判別分析で分類するために以下に示す方法を用いた。

3.2.1 韻律情報の抽出

基本周波数の抽出

基本周波数 (以下 F_0) の抽出方法としては種々提案されているが、波形の位相歪みに強く、簡単に実現できる相関処理法が最も広く用いられている。本研究ではこの相関処理法の 1 つである平均振幅差関数 (AMDF) を用いる。

有声区間の検出

音声パワーと F_0 の二階階差を各フレームについて求め、それぞれ閾値を用いて有声 / 無声を決定する。音声パワーの閾値 P_{th} は、各音声データにおける全音声区間中のパワーの最大値 P_{max} と最小値 P_{min} から式 (1) のように算出する。

$$P_{th} = P_{min} + (P_{max} - P_{min}) \times 0.15 \qquad (1)$$

音声パワーが P_{th} より大きいフレームを有声区間候補とする。また、F_0 の二階階差の閾値 F_{0th} は、予備実験より 6.0 ($\log_2$Hz) とし、F_0 の二階階差が F_{0th} より小さいフレームを有声区間候補とする。この両者の条件を共に見たす区間を有声区間として決定する。

3.2.2 F_0 パターンの近似

F_0 は 10msec 毎に抽出されるため、データ量が多く、扱い難い。そこでデータ量を減少させ、F_0 の概形を捉えやすくするように、F_0 の時間的変化を関数で近似する。この近似には、以下に示す 2 通りの方法を用いた。

折れ線による近似

抽出した F_0 パターンの時系列 $f_0(t)$ の正規化対数値が、

$$F_0(t) = \log_2 f_0(t) - \overline{(\log_2 f_0)} \qquad t = 1, 2, \cdots, T \qquad (2)$$

で与えられるとする。$F_0(t)$ を式 (3) で表わされる K 本の直線によって近似する。

$$y_k(t) = a_k(t - t_{k-1}) + b_k \qquad k = 1, 2, \cdots, K \quad (3)$$

ただし、a は直線の傾き、b は直線の切片、t_{k-1} は隣接した 2 つの直線の境界である。

a, b は各 k について、$y(t)$ と $F_0(t)$ の二乗誤差を最小にすることで求める。$F_0(t)$ から時間的な変化の特徴を捉えるため、互いにつながった直線 (折れ線) を用いる。

折れ線の境界の決定には、DP(動的計画法) を用いる。このとき、近似する折れ線の本数が多くなりすぎないように、近似誤差の閾値により本数を決定する。

指数関数を取り入れたモデルによる近似

折れ線近似では正確に近似できる反面、折れ線の本数が多くなる傾向がある。このため、パラメータ数が多くなり、韻律的特徴を捉えるのが難しい。そこで、関数の個数を少なくできるような近似関数を考える。

一般に、F_0 の概形には「へ」の字型のパターンが多く観察される。音声処理におけるパラメータの近似に有用な関数のうち、「へ」の字型に近い、線形 2 次系インパルス応答関数を選んだ。

近似式の導出

まず近似関数として、次式のような線形 2 次系のインパルス応答関数を用いる。

$$g(t) = \frac{e}{\tau} t e^{-\frac{t}{\tau}} \qquad (4)$$

ただし、e/τ は $g(\tau)$ の最大振幅を 1 にするための正規化係数である。この関数は、このままでも「へ」の字型の F_0 パターンに近く、振幅や時定数 τ を変化させることでその概形を比較的自由に変形できる関数となっている。

しかし、より自由に近似ができるよう、式 (4) に傾きを考慮し、式 (5) で表わされる直線を取り入れたモデル (以下指数モデル) を用いる。

$$y(t) = ag(t) + bt + c \qquad (5)$$

ここで、a は振幅、b は傾き、c は切片を表す。図 1に、式 (5) で表わされる指数モデルの概形を示す。

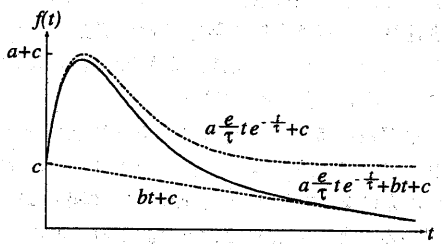


図 1: 指数関数を取り入れたモデルの概形

$t = 1, \cdots, T$ を分析フレームとして観測される F_0 を $f_0(t)$ とし、対数振幅で表わした $F_0(t) = \log_2 f_0(t)$ を用いると、$y(t)$ と $F_0(t)$ の二乗誤差 $\varphi(t)$ は式 (6) で表わされる。

$$\varphi(t) = \sum_{t=1}^{T} \{ag(t) + bt + c - F_0(t)\}^2 w(t) \quad (6)$$

誤差は有声と判定された区間に対してのみ考慮するので、有声ならば1、無声ならば0をとる関数 $w(t)$ を掛ける。

a, b, c それぞれについての偏微分をとり、これを0とすることで $\varphi(t)$ を最小にする a, b, c を求めることができる。

得られた a, b, c の値を用いて、指数関数の時定数 τ を $1 < \tau < T$ の範囲で変化させる。$\varphi(t)$ が最小となるときの τ を求めることで、最適な近似曲線 $y(t)$ を得ることができる。

近似の方法について

求まる近似関数は「へ」の字型以外にも自由に変形できてしまうため、近似関数の変数の取る値に制限を設ける。

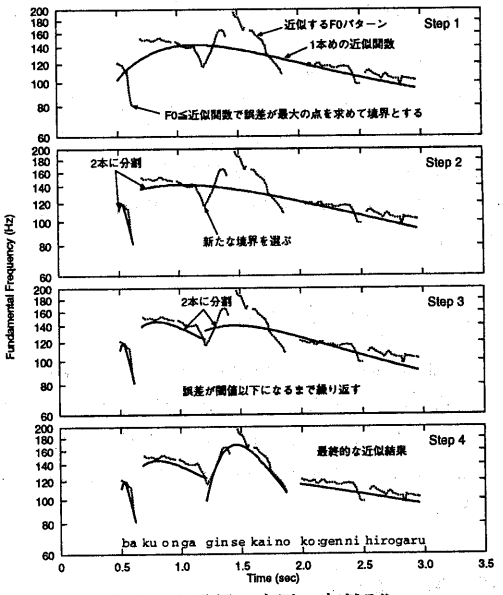
振幅 a は正と負の両方の値を取るが、$a < 0$ のときは近似関数は谷型となってしまう。$a \geq 0$ に制限することで近似関数の頂点を正方向に統一することができる。

$b > 0$ のとき、近似関数の概形は左上がりとなるが、F_0 がこのようなパターンとなるのは極めて少ないと考えて、ここでは $b \leq 0$ とする。

折れ線近似ではそれぞれの直線が境界上でつながるように直線の変数を求めていたが、この指数モデルではそれぞれの近似関数が連続する必要はないという方針をとった。また、折れ線近似では有声区間毎に近似を行っていたが、指数モデルではその範囲を拡大し、短時間の有声 / 無声にとらわれず、ある程度連続した音声区間を数本の近似関数で近似することにした。しかし、無声区間が500msec以上続いた場合には別の区間として扱う。

近似関数の境界の決定には、折れ線近似同様DPを利用することが望ましいが、DPのための計算時間が膨大になってしまう[7]。そこで、本稿では簡易法を用いることで高速化を計った。そのアルゴリズムを以下に示す。また、実際の F_0 パターンを簡易法によって近似をする様子を図2に示す。図2は日本人男性話者による“爆音が銀世界の高原に広がる”という発話である(OGIのデータではない[2])。近似関数の本数は最大誤差の閾値で決定する。

1. 最初に各有声区間を1本の関数で近似する。
2. 最大誤差を閾値と比較、閾値以下ならば終了。
3. F_0≦近似関数を満たす区間で近似関数と F_0 との誤差が最大となる点を探し、その点を境界として近似区間を2分割し、再度近似する。
4. 上記step2,3を繰り返す。



“爆音が銀世界の高原に広がる”

図 2: 簡易法による近似の様子 (日本語 男性)

3.2.3 分析に用いる統計量

F_0、音声パワー、F_0 パターンを近似した関数の係数などから特徴量を話者ごとに求める。表1に折れ線近似から求める統計量、表2に指数(直線)モデル近似から求める統計量を示す。

ただし、これらの統計量は単位が統一されていないので、統計量の平均が0、分散が1となるように正規化してから、判別分析を行う。

3.3 両手法の統合

3.1節と3.2節により、テスト用データと各言語の学習用データとの、それぞれ尤度と距離が求められる。この2つの数値を合わせることでより高い認識率を得ることを考える。3.1節で求められた、10言語の尤度の絶対値を平均0、分散1となるように正規化した。それを $d_1(i)$(ただし i は各言語) とする。

表 1: F_0 と折れ線から求める統計量

No.	PARAMETERS
	F_0 と音声パワーについて
1,2,3	F_0 の標準偏差・歪度・尖度
4,5,6	音声パワーの標準偏差・歪度・尖度
7	F_0 と音声パワーの相関係数
	近似した折れ線について
8	傾き正と負の区間長の比 (正 / 負)
9,10	単位時間当たりの本数 (正, 負)
11,12	傾きの平均値 (正, 負)
13,14	傾きの標準偏差 (正, 負)
15,16	相対開始周波数 (正, 負)

表 2: F_0 と指数モデルから求める統計量

No.	PARAMETERS
	F_0 と音声パワーについて
1 ～ 7	表1と同じ
	近似した指数モデルについて
8	近似区間長の平均値
9	単位時間当たりの本数
10,11	振幅 a の平均値, 標準偏差
12,13	傾き b の平均値, 標準偏差
14,15	時定数 τ の平均値, 標準偏差
16	相対開始周波数

また、3.2節で求められたテスト用データと各言語の学習用データの距離も、同様に正規化を行う。この値を $d_2(i)$ で表す。

これらをある割合で足し合わせたものをテスト用データと各言語の学習用データとの値 $v(i)$ とし、この値が最も小さいものをテスト用データの言語であると分類する。すなわち、

$$v(i) = c \cdot d_1(i) + d_2(i) \qquad (7)$$

ただし、c は最適な分類を行うための重みである。

4 分析結果

4.1 音韻情報を用いた分類結果

表3に音韻情報を用いた分類の結果を示す。識別率は train(閉じた実験) では 300 人、test(開いた実験) では 200 人中での割合である。

表 3: 音韻情報を用いた実験結果

データ	train		
状態数	**16**	**32**	**64**
識別率 (%)	74.0	89.7	96.7

データ	test		
状態数	**16**	**32**	**64**
識別率 (%)	46.0	55.5	56.0

4.2 韻律情報を用いた分類結果

図3に、F_0 抽出結果、および F_0 パターンの近似結果を示す。近似結果は同一図中に折れ線と指数モデルの両方について示した。横軸は時間 (sec) を表し、縦軸は基本周波数 (Hz) を表す。

図3は日本人男性話者による "今わたしはシアトルに住んでいますけれども、え、シアトルは大変住みやすく" という発話である。

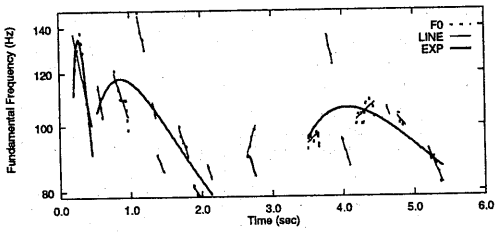


図 3: F_0 パターンの近似結果

また、判別分析の結果を表4に示す。音韻情報を用いた結果と比べると、認識率は低いが、いずれもベースライン (10%) を越え、4.3節に示す統合結果に大きな影響を与えている。

表 4: 韻律情報を用いた実験結果

データ	train		test	
近似法	**line**	**exp**	**line**	**exp**
識別率 (%)	36.7	37.0	25.5	28.0

4.3 統合方法による分類結果

3.3節で示した方法による結果を表5と表6に示す。この表の最上段の 16,32,64 は学習の HMM の状態数を表し、**line**,**exp** は近似の方法を表す。

また、比較のために音韻のみ、韻律のみの結果も示した。式7の c の値を $0.1, 0.2, 0.3, \ldots, 10.0$ と 0.1 ずつ変化させた。例として $4.0, 3.0, 2.0, 1.0, 0.5$ の重

表 5: 統合した実験結果 — 閉じた実験

状態数		16		32		64	
近似		line	exp	line	exp	line	exp
音韻のみ		74.0		89.7		96.7	
韻律のみ		36.7	37.0	36.7	37.0	36.7	37.0
4.0		79.0	75.3	91.7	91.0	96.7	96.7
3.0		79.0	75.3	92.0	90.3	97.0	95.7
2.0		79.7	76.3	90.7	89.0	96.3	95.3
1.0		77.3	73.0	86.0	80.3	90.7	87.0
0.5		68.3	59.7	74.3	65.0	80.3	69.3
最大	重み	2.4	1.4	2.4	3.8	7.3	3.9
	識別率	80.3	77.0	92.0	91.0	97.3	96.7

表 6: 統合した実験結果 — 開いた実験

状態数		16		32		64	
近似		line	exp	line	exp	line	exp
音韻のみ		46.0		55.5		56.0	
韻律のみ		25.5	28.0	25.5	28.0	25.5	28.0
4.0		51.5	48.5	59.0	55.0	55.5	54.5
3.0		52.5	50.0	58.5	54.5	55.5	54.5
2.0		54.0	50.5	58.0	53.0	58.5	53.5
1.0		51.5	46.5	53.5	50.0	53.0	50.5
0.5		41.0	38.5	46.0	39.5	42.5	41.0
最大	重み	1.7	2.0	2.4	4.4	1.8	6.8
	識別率	54.0	50.5	60.0	57.0	59.0	57.5

みで得られた識別率と、最大となった時の重みと識別率を示す。ただし最大の重みが複数ある場合は最小のものを示す。

5 考察

音韻情報のみによる手法や韻律情報のみによる手法よりも、これら2つの手法を統合することで、ほとんどの場合、より良い認識ができることが示された。また、音韻情報の重みを韻律情報の重みの2～4倍程度にすると、さらに良くなることが示された。

6 今後の課題

音韻情報を用いた実験では、HMM の状態数 64 ではまだ認識率が収束していない。そのため、さらに多くの状態数での学習が必要であると考えられる。

韻律情報を用いた実験の認識率はそれほど高くない。これは判別分析に用いた特徴量に適切でないものが含まれているためと考えられる。検定を行うなどで、最適な特徴量を探す必要がある。

韻律情報のみでは指数関数近似を用いた認識の方が結果が良いのに対し、統合結果では折れ線近似の方が良くなっていることが多い。これは統合の方法が最適でないためと考えられる。HMM の学習で得られた結果を判別分析のパラメータに加える、といった方法なども検討してみたい。

参考文献

[1] Y.K.Muthusamy, R.A.Cole and B.T.Oshika, "The OGI Multi-language Telephone Speech Corpus", *Proc. ICSLP92*, Banff, Canada (1992).

[2] 音声データベース (CD&CD-ROM), "Multi-Lingual Speech Database for Telephonometry 1994", NTT アドバンステクノロジ (1994)

[3] Cambridge University Engineering Department Speech Group *HTK : Hidden Markov Model Toolkit ver.2.1* (1997).

[4] Allan A.Reyes, 中川 聖一, "エルゴディック HMM と最適状態シーケンスによる言語識別法の OGI データベースによる評価", 日本音響学会講演論文集, 1-Q-26, pp. 147-148 (Mar. 1995).

[5] S.Itahashi, J.Zhou and K.Tanaka, "Spoken Language Discrimination using Speech Fundamental Frequency", *Proc. ICSLP94*, S31-17, pp.1899-1902, Yokohama (Sep. 1994).

[6] ITAHASHI Shuichi, DU Liang, "Language Identification Based On Speech Fundamental Frequency", *Proc. EUROSPEECH95*, pp.1359-1362, Madrid (Sep. 1995).

[7] 笹沼 満, 板橋 秀一, "基本周波数による多言語音声の分類", 信学技報, SP96-57, pp.13-20 (Oct. 1996).