論文 / 著書情報
Article / Book Information

論題	話者認識
著者	古井貞熙
掲載誌/書名	昭和55年電気四学会連合大会，，，pp. 129-132
発行日	1980,
URL	http://search.ieice.org/
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright (c) Institute of Electronics, Information and Communication Engineers.

33. 音声の認識と合成

33-6

話者認識

古井 貞熙 （日本電信電話公社 武蔵野電気通信研究所）

I まえがき

音声波には，意味内容に関する情報のほかに，発声者個人の声の特徴，喜怒哀楽に関する情報等が含まれている．このうち発声者の声の特徴を自動的に抽出して，その話者を判定する話者認識の研究は，意味内容に関する情報を抽出する（狭義の）音声認識の研究に比べて地味ではあるが，国内外の研究機関で地道に進められている．このような技術はすでに，一部人間の能力を凌駕する程度にまで進んでおり，将来は，音声を印鑑や身分証明書代りに用いた，電話によるバンキング，クレジット電話，種々の場所や情報へのアクセスをコントロールするための音声による鍵などへの応用が考えられる．

話者認識の方法と形態は，次のように4つの観点から分類することができる．[1][2]

話者識別 (Speaker Identification)
　ある音声が，登録されている複数の話者のうちの誰の声であるかを判定する．

話者照合 (Speaker Verification)
　ある音声が，名乗った本人のものであるか否かを判定する．（図1参照）

発声内容依存形 (Text-dependent)
　発声する言葉をあらかじめ決めておく方法．

発声内容不依存形 (Text-independent)
　発声する内容は何でもよいとする方法．

動的特徴による
　音声スペクトル（ピッチ・パワー等も含む）の時系列を直接的に用いる．

統計的特徴による
　音声スペクトルの時系列から抽出した時間軸を持たない特徴を用いる．

先天的特徴による
　発声器官の構造を反映する特徴を用いる．

後天的特徴による
　発声器官の動かし方（発声習慣・くせ）を反映する特徴を用いる．

発声内容依存形では，動的特徴を用いるものと統計的特徴を用いるものがあるが，発声内容不依存形では主として統計的特徴が用いられる．一般に発声内容不依存形の方が難しい．これまでの研究では，後天的特徴よりも先天的特徴を用いるものの方が多い．

筆者らはこれまでに，発声内容不依存形の認識方法として，長時間平均スペクトルのケプストラムを用いる方法を提案し，この有効性を確認したが[3]，93%程度の認識率を得るために10秒程度以上の音声を必要とするという問題点があった．次に筆者らは，発声内容依存形の認識方法として，単語音声の対数断面積化および基本周波数の統計的特徴を用いた認識方法を提案し，4単語の特徴を組合せて用いれば，99%程度の認識率が得られることを示した．[4] 筆者らはさらに，動的特徴を用いた発声内容依存形の認識方法として，対数断面積化と基本周波数の時系列と，パターンマッチング時の時間正規化に伴って抽出される時間伸縮関数に含まれる情報を用いた認識方法を提案し，この有効性を示した．[5]

一般に話者認識を行う上で最も難しい問題点の一つは，同じ人が同じ言葉を発声しても，数か月程度以上の期間をおくと，そのつど音声のスペクトルが変化し，それが認識の精度に大きな影響を与える点にある．筆者らはこれまでに，このような問題に対処する方法として，一次系，または二次臨界制動系の逆フィルタによるスペクトル等化処理が有効であることを明らかにした．[6][7]

ここでは，筆者らが最近検討した発声内容依存形の新しい二つの方法を紹介する．

話者認識の研究は，国内では他に日立，科警研，早大，北大等，国外では，米国ベル研，TI，西独HHI，フィリップス研等で進められている．[1][2]

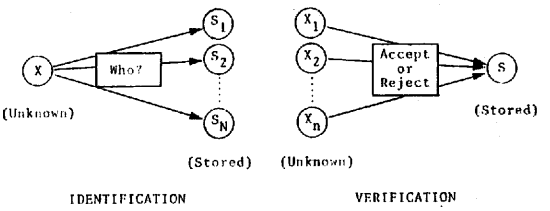

図1．話者識別と話者照合

II ケプストラムの動的特徴を用いた話者認識

2.1. 特徴抽出と認識方法

音声スペクトルの動的特徴を用いる認識方法の改良として，単に音声スペクトル時系列の非線形時間伸縮マッチングを行うだけでなく，動的特徴を積極的に抽出して用いる次のような方法を検討した．本方法は，特に電話音声に対して良好に動作することを目的として構成されて

いる.[8]~[10]

音声サンプルとしては短文章音声を用い，図2に示すように，音声の始端と終端を検出したのち，高域強調を施し，LPC分析をへて，ケプストラムの時系列に変換する．標本化周波数は6.67 kHz，フレーム長は30 ms，フレーム周期は10 ms，分析次数は10である．音声区間全体にわたって平均化したケプストラムの値を，各時点のケプストラムの値から減ずることによって，伝送系や長期間にわたるスペクトルの変動を除去する．

一方，各ケプストラムの時系列を，10 ms毎に90 msの区間を用いて直交多項式展開し，その1次と2次の係数の時系列を得る．これらの展開係数と，平均値を減じて正規化したケプストラムのうち，話者を分離する上で有効性の高い18種類の成分の時系列をとり出し，標準パターン時系列との非線形時間伸縮マッチングを行う．選択する成分は，多数話者の学習サンプルを用いた分散分析によって決定する．

時間伸縮マッチングは，始端と終端を固定しない非対称形のDPマッチングによって行い，入力音声と標準パターンのうち，全長の短い方を基準(Guide)側として用いる．図3に示すように，二つの時系列が作る平面の対角線の両側に(m_0=)20フレームずつの整合窓を設け，斜線を施した範囲内でDPマッチングを行う．このとき入力音声と標準パターンのフレーム間の距離としては，各成分毎にその安定性にもとづいた重みづけを施した距離を用いる．こうして音声区間全体にわたる平均距離を求め，この値を，あらかじめ各話者毎に設定してある閾値と比較して，話者照合の判定を行う．この閾値は，学習サンプルを用いて求めた異話者間の距離の分布にもとづいて設定しておく．この認識方法に関しては，話者照合実験のみを行い，話者識別実験は行わなかった．

2.2. 認識実験の方法

実験には，表1に示す三種類の音声サンプルセットを用いた．発声者はいずれも生まれながらの米国人で，男性は "We were away a year ago"，女性は "I know when my lawyer is due" と発声した．音声サンプルセット1と2は，カーボン送話器を含む電話機から入力して，PBXを含む電話系に通してから録音した音声で，各話者は各文章を，約2ヵ月にわたって1日2回（午前と午後），計50回発声した．音声サンプルセット3は，高品質マイクロホンから入力して録音した音声で，各話者は，同じく2ヵ月にわたって1日に1回ずつ，計26回発声した．各詐称者（登録話者以外の棄却すべき音声の話者）の発声回数は，いずれも1回である．

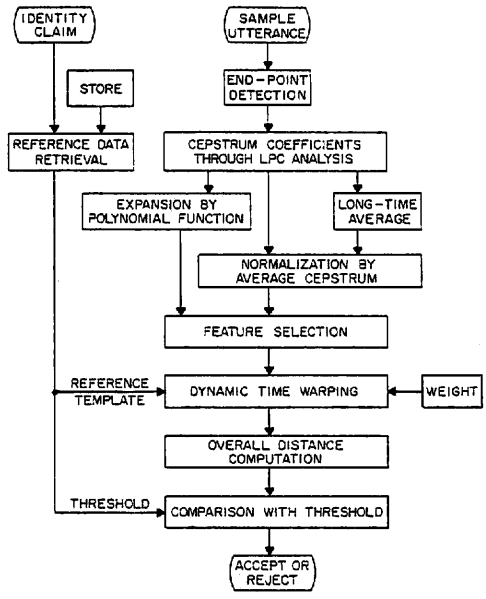

図2. ケプストラムの動的特徴を用いた話者照合のブロック図

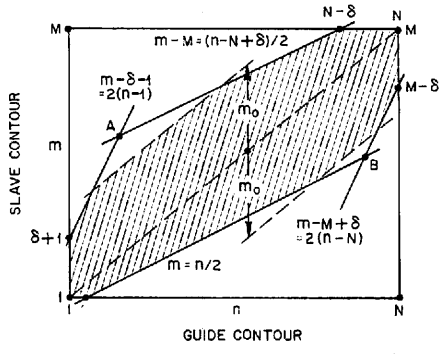

図3. 時間伸縮マッチングにおけるDPパスの許容範囲

表1. 三種類の音声サンプルセットと認識誤り率

UTTERANCE SET		1	2	3
CUSTOMERS & IMPOSTORS		10 MALE & 40 MALE	10 FEMALE & 40 FEMALE	21 MALE & 55 MALE
RECORDING CONDITION		TELEPHONE	TELEPHONE	HIGH QUALITY MICROPHONE
AVERAGE ERROR RATES	A PRIORI FIXED	0.30 %	0.42 %	1.03 %
	A PRIORI ESTIMATED	0.19 %	0.36 %	0.77 %
	A POSTERIORI	0 %	0.06 %	0.64 %

各話者の標準パターンは，図4に示すように，まず初めの5サンプルを相互に時間伸縮マッチ

ングしたのち平均化して作成し，その後7回発声する毎に，その直前の5回の音声を用いて更新した．話者照合の判定の閾値も，同時に更新した．

2.3. 認識実験の結果

認識実験の結果は表1に示す通りで，比較のために，閾値を事後的に最適値に設定する場合や，すべての話者に共通の値にあらかじめ固定した場合についても実験を行った．この結果から，三種類の音声サンプルセットに共通して，閾値を話者毎に推定して事前に設定することが可能で，1%以下の低い誤り率を得ることができることがわかる．閾値をすべての話者に共通の値に固定すると，誤り率がやや増加する．

参考のために，音声サンプルセット1を用いて，学習サンプルと入力音声の時期差を，これまでの条件（平均6日）から6週間まで延長して認識実験を行ったところ，図5に示す結果が得られた．認識誤りを生ずるサンプルの絶対数が少ないので，結果にややばらつきが見られるが，時期差による誤り率の増加傾向は見られない．音声サンプルセット2についても，同様の結果が得られた．時期差をさらに延長した場合については，実験の必要があろう．

次に上記の電話音声を，24 kbit ADPCM およびLPCボコーダの伝送系に通し，元の電話音声を含む三種類のサンプルセットの組を作って，学習サンプルと入力音声が異なる伝送系を通った場合について認識実験を行ったところ，この場合でも，誤り率は1%以下となることがわかった．

その後，本システムを用いて，男女計120名の実際の電話音声によるオンラインの実験を約6か月にわたって継続し，良好な結果を得ている．

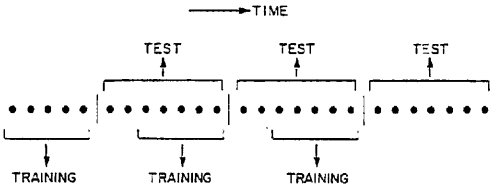

図4．話者照合の閾値と標準パターンの更新方法

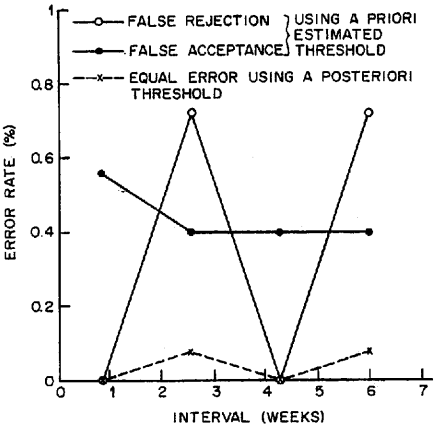

図5．時期差と認識誤り率の関係（音声サンプルセット1）

III ケプストラムの統計的特徴を用いた話者認識

3.1. 特徴抽出と認識方法

音声スペクトルの統計的特徴を用いる認識方法の改良として，次のような方法を検討した．図6にシステムのブロック図を示す．[2] 音声サンプルとしては，種々の単語音声を用い，音声の始端と終端を検出したのち，10ms毎に30msの区間を切り出して，各区間毎に有声／無声の判定を行い，有声音区間のみの相関係数の時系列に変換する．この相関係数を，単語の全有声音区間にわたって平均化し，この値を用いて，一次系の逆フィルタを構成する．もとの相関係数に畳み込みを行ってスペクトル等化処理を行い，こののちに基本周波数 f_{ot} と10次までのケプストラム $\{c_{it}\}_{i=1}^{10}$ からなる11次元ベクトル X_t の時系列に変換する．

$$X_t = (c_{1t}, c_{2t}, \cdots, c_{10t}, f_{ot}) \qquad (1)$$

この時系列から，各単語の全有声音区間における平均値ベクトル μ，標準偏差ベクトル σ，共分散行列 Σ，相互相関行列 A，およびフーリエ（コサイン）展開係数行列 F を計算する．

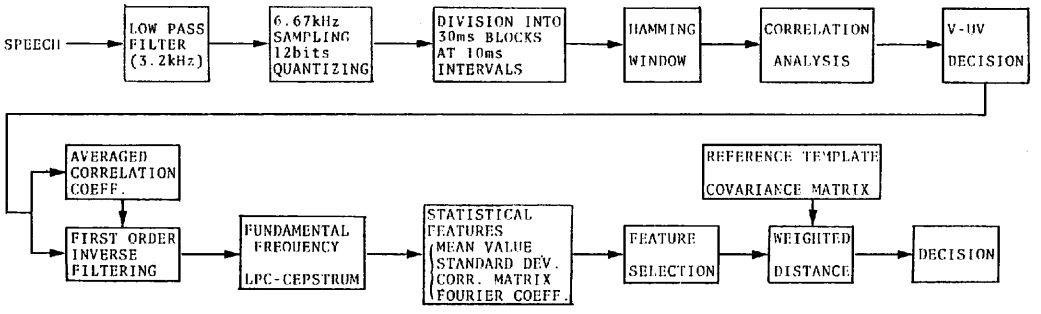

図6．ケプストラムの統計的特徴を用いた話者認識のブロック図

$$\mu_\mu = \left(\sum_{t=1}^{N} x_t\right)/N \quad (2)$$

$$\Sigma = \left(\sigma_{ij}\right) = \left\{\sum_{t=1}^{N}(x_t-\mu_\mu)'(x_t-\mu_\mu)\right\}/(N-1) \quad (3)$$

$$\sigma = (\sqrt{\sigma_{11}}, \sqrt{\sigma_{22}}, \cdots \sqrt{\sigma_{\mu\mu}}) \quad (4)$$

$$\Lambda = \left(\lambda_{ij}\right), \quad \lambda_{ij} = \sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}} \quad (5)$$

$$F = \left(F_{il}\right), \quad F_{il} = \left\{\sum_{t=1}^{N} x_{it}\cos\frac{\pi(t-1)l}{N}\right\}/N \quad : 1 \le l \le 6 \quad (6)$$

ここでNは，その単語音声における有声音区間のフレーム数である．

これらの統計量のうち，標準偏差ベクトルσ，相互相関行列Λ，およびフーリエ展開係数Fから，話者の分離に有効な60成分を選択し，平均値ベクトルμの全成分と組合せて，71次元の特徴ベクトルとする．選択する成分は，多数話者の学習サンプルを用いた分散分析によって決定する．話者認識の判定は，入力音声の特徴ベクトルと各話者の標準パターンの，各成分の安定性を考慮した重みつき距離によって行う．

話者識別の場合は，入力音声との距離が最も小さい標準パターンの話者による音声と判定し，話者照合の場合は，前章と同様に，距離と閾値との大小関係によって判定を行う．

3.2. 認識実験の方法

実験には，男性9名が3年間の長期間にわたって繰返し発声した/namae/,/baNgo:/,/ko:geN/,/bakuoN/の日本語4単語の音声サンプルを用い，4単語の特徴を組合せて認識の判定を行った．

[実験1] マイクロホンで収録した音声を用い，各話者毎に，3か月おきの4時期に収録した12サンプルを学習サンプルとして，その3か月後の音声を認識する実験を，多時期の音声サンプルを用いて行った．全登録話者の入力音声の総数は216サンプルである．話者照合の詐称者の音声としては，103名の男性話者が1回ずつ発声したものを用いた．

[実験2] カーボン送話器を含む電話機及び擬似加入者線（0.4mmφ-7dB）を通った音声と，同時に録音したマイクロホン音声を用いて，実験を行った．学習は6か月おきの3時期（9サンプル），入力音声との時期差は6か月とし，詐称者の音声として，11名が3回ずつ発声したものを用いた．全登録話者の入力音声の総数は54サンプルである．

3.3. 認識実験の結果

実験1の結果を表2に，実験2の結果を表3に示す．話者照合に関しては，ここでは閾値を，すべての話者に共通の値にあらかじめ固定した場合と，事後的に最適値に設定する場合についてのみ実験を行い，話者毎に推定して設定する場合については実験を行わなかった．

表2．実験1の認識誤り率

Identification		0 %
Verification	A Priori Fixed Threshold	0.10 %
	A Posteriori Threshold	0.02 %

表3．実験2の認識誤り率

Tel.	Identification		0 %
	Verification	A Priori Fixed Threshold	2.44 %
		A Posteriori Threshold	0.49 %
Mic.	Identification		0 %
	Verification	A Priori Fixed Threshold	2.44 %
		A Posteriori Threshold	1.36 %

表2の結果から，ここで検討したケプストラムの統計的特徴による方法が極めて有効性が高く，極めて低い認識誤り率が得られることがわかる．実験2は実験1よりも実験条件が厳しいので，マイクロホン音声の場合でも，やや誤り率が大きくなっているが，本方法が，電話音声に対しても，マイクロホン音声とほぼ同等の性能で動作することがわかる．

市外回線や雑音の影響等に関しては，さらに実験を行う必要があろう．

Ⅳ むすび

筆者らが最近検討した二つの話者認識方法と実験結果を紹介した．いずれもLPC-ケプストラム分析法にもとづき，スペクトル等化処理を含んでおり，一方は動的特徴を，他方は統計的特徴を用いている．両者の実験では共通のデータベースを用いていないので厳密な比較はできないが，いずれもほぼ同程度の高い認識率を示しており，処理量は後者の方が1/10程度に少ない．今後実際の電話系を用いた両者の比較や，両者の長所を組合せた方法の検討を進めて行きたい．

[謝辞] 日頃御指導頂く当所・野田直躍研究部長，小池武四研究室長，ならびに第2章の研究に関し御指導頂いた米国ベル研究所 J.L. Flanagan, A.E. Rosenberg 両博士に感謝する．

[文献] (1) 古井：音響音声学講論 1-7-2 ('80), (2) 古井：音響音声研資，6月 ('80), (3) 古井他：信学論 55-A, p.549 ('72), (4) 古井他：信学論 56-A, p.717 ('73), (5) S. Saito et al.: 4th IJCPR, p.1014 ('78), (6) 古井：信学論，57-A, p.880 ('74), (7) S. Furui: 96th ASA, NNN 28 ('78), (8) S. Furui: 98th ASA, R4 ('79), (9) S. Furui et al.: ICASSP, p.1060 ('80) (10) 古井：音響音声学講論 1-7-1 ('80), (11) B.S. Atal: Proc. IEEE, 64, p.460 ('76), (12) A.E. Rosenberg: ibid., p.475.