解　説　音声処理技術とその応用

6. 音声識別

Speech Indentification by Masahide SUGIYAMA (Univ. of Aizu).

杉山　雅英[1]

1 会津大学コンピュータ理工学部

1. はじめに

音声には多様な情報が含まれており，我々はそれらを用いて豊かな情報交換を行っている．音声に含まれる多様な情報を自動認識することが音声認識の目的である．狭い意味での音声認識は音声に含まれる言語情報(音素系列，単語系列，文，などの発話内容，テキスト，何をいっているかに対応する)の復号化問題として捉えられ，音声に含まれるそれ以外の情報は言語情報復号化を困難にする雑音(厄介者)として扱われている．本稿では言語情報以外の音声情報の識別問題についていくつかのトピックに限定して研究現状，想定される応用例，将来の可能性について述べる．

音声に含まれる情報カテゴリーを表-1に示す．識別対象である情報カテゴリーを自動識別するためにはカテゴリーモデルの学習とそのモデルを用いた識別の2つのステップが必要である．識別段階で入力される音声と同様の性質をもつ音声(音声データベース)を用いてカテゴリーモデルの学習を行う．入力音声を限定することによってカテゴリーモデルをよりコンパクト，精密に作成し識別性能を向上させることが可能であるが，入力音声が限定され発話が制限されるので使用者にとっては使い難いものとなる．

2. 話者識別

2.1 話者識別と話者照合

音声における話者性の存在(音声の性質が話者によって異なる)は音声認識を困難にする1つの原因であるが，一方話者の特徴が含まれていることにより話者の特定を可能としている．話者を特定すれば音声認識の性能は向上し，一方発話内容を限定すれば話者認識の性能は向上する．そのような観点から話者認識と音声認識とは相補的な側面をもっており，話者認識技術は話者正規化や話者適応技術と関係している．

話者認識[1]は，話者識別(speaker-identification)と話者照合(speaker-verification)とに大分される．話者識別は，あらかじめ登録された話者集合の中から入力音声に最も近い話者を選択することであり，話者照合とは与えられた音声が該当する話者であるかどうかを判定することである．前者の計算量・認識性能は話者集合の大きさに大きく依存する．一方，後者は話者集合の大きさとは無関係である．音声を用いたドアロックの解除などへの応用は後者の技術となる．

話者認識に用いる音声内容の種類によって以下の3つに分類される．

1. テキスト従属(text-dependent)
2. テキスト独立(text-independent)
3. テキスト指定(text-prompted)

第1のテキスト従属(text-dependent)とは発話内容(テキスト)があらかじめ決められている場合であり，たとえば単語音声認識などの技術を用いることが可能である．第2のテキスト独立(text-independent)の場合には，発話内容をとくに限定しない音声を用いるので，テキスト従属の場合に比べて発話内容を覚えておかなくてよいという点で使用者への負担は少ないが発話内容が限定されないだけ認識性能は低い．たとえば，これは犯罪者の残した音声を用いて過去の犯罪者集合の中から該当者をみつける場合などへの応用となる．また音声認識のための話者適応技術として応用される．これらの2つの方法を用いた話者認識セキュリティシステムは，録音した話者音声を用いることによって簡単にシステムを破られてしまうという弱点がある．第3の方法であるテキスト指定(text-prompted)による方法は，システムが発話内容を指定し，話者認識と音声認識技術と組み合

表-1 音声に含まれる情報カテゴリーとその指定方法

情報カテゴリー	指定方法	研究例
テキスト	文字列（単語，...）	word spotting
話者	話者の発話した音声	話者認識[1]
男女性	男／女	男女性識別[2]
年齢	子供／青年／壮年／老年	−
感情	happy/angry/sad/...	−
言語	日本語／英語／露語／...	言語識別[3]
方言	東京／関西／会津／...	方言識別[23]
音声・非音声		
話題		話題識別[4]
環境	静か／雑音／...	−
⋮	⋮	⋮

わせて，指定されたテキストを話者が即座に応答できない場合には拒否する方式である[5]．将来，任意の話者の音声をテキストから合成することができるようになればこのテキスト指定の方式でも十分とはいえないことになる．

そのほかの問題として認識の堅牢性がある．一般に話者モデルと入力音声の間の距離や確率を計算し，閾値論理を用いて判定を行う．話者モデルを学習する際に用いる音声，認識における入力音声の特性が一致していれば，高い認識率を得ることができる．しかし，入力音声に対して雑音や伝達特性の変化によって大きく認識性能が劣化する．閾値はタスクに依存して最適に設定される．しかし，雑音のない環境での入力音声に対して設定される最適閾値は，雑音環境下での入力音声に対しては最適の閾値とは異なる．堅牢性を高めるための研究が進められている[6]．

話者識別の新しい応用として長時間の音声に含まれる複数話者の音声の区間を話者の切替りに応じて区分化し分類する問題が検討されている[7)~9)]．文献7)では管制塔とパイロットとの交信記録の音声を取り扱っている．今後は従来の話者識別の枠を越えた新たな音声処理応用としてさらに研究が深められていくものと考えられる．

2.2 応用例

テキスト指定による話者照合方式に関して，10名の男性と5名の女性話者が10カ月にわたる5時期に発声した文章データの中から，ある時期に発話した10文章（40秒程度）を用いて話者モデルを学習し，学習時期とは異なる約4.2秒の文章を用いて話者照合を行った．評価実験の結果，話者およびテキスト照合誤り率1.1％と報告されている[5]．今後は雑音下における評価，電話入力音声を用いた評価，最適な閾値の自動決定などを検討する必要があろう．米国においては電話会社が話者照合機能を組み込んだカード（Sprint の VOICE ACTIVATED FONCARD）やサービス（AT&T の smart card）を開始している．

3. 言語識別

3.1 言語識別

言語識別は音声に含まれる言語の特徴情報からその言語を識別することを目的とする．従来の音声認識の目的は，与えられた音声から指定された言語の指定されたタスク（語彙や文法を制限する）における最も可能性の高い音素列を推定する問題である．したがって，言語従属，タスク従属，話者独立（不特定話者）となる．それに対して言語識別の一般形式は，タスク独立（テキスト独立），話者独立，すなわち誰が喋っても何を喋ってもその発話言語を識別することとなる．一般に独立性が大きく入力の制約が少なくなるにつれて探索問題における探索空間の規模が大きくなるので，問題の難しさは大きくなる．また言語識別用の言語モデルを学習するための音声・テキストデータベースは，上で述べたタスク独立（テキスト独立），話者独立となるために大規模化することになる．

言語識別に用いられている音声データベースとして NTT 20カ国多言語データベースと OGI TS（Oregon Graduate Institute Multu-language Telephone Speech Corpus）が知られている．前者は音声符号化評価用に集められた音声データであり，収録条件や雑音レベルが収録機関によって若干異なる．後者は多言語音声識別用データベースとして設計され，現在多くの研究機関で用いられている．このデータベースは11カ国語（英語，ファルシ語（現代ペルシャ語），フランス語，ドイツ語，ヒンズー語，日本語，朝鮮語，中国語，スペイン語，タミル語，ベトナム語）の自由発話音声および固定語彙の音声から成り立っている．電話回線を通じて各言語ごとに90名のネイティブ話者の音声を収録しており，すでに公開されLDC（the Linguistic Data Cosortium）を通じて入手可能となっている．

音声に含まれる言語識別のための情報は以下の

ような各種のレベルに含まれている．

1．音響音声レベル(phonetics)
2．音素配列レベル(phonotactics)
3．韻律レベル(prosodics)
4．語彙レベル(vocabulary)

第1の音響音声レベル[10),11)]は，たとえば英語に存在する/r，v，f，th/などの音素が日本語にはないなどに対応する．第2の音素配列レベル[13),14),16)]は音素に基づく処理をより深めその配列を考慮する．たとえば，英語には/str/などの子音列があるが日本語には存在しないことに対応する．第3の韻律レベルは音声の基本周波数や音声パワーなどの時間的な変化に着目する[21)]．第4の語彙レベルは，日本語の“mashita”や朝鮮語の“imnida”などの特徴的な単語，その言語のもつ語彙，さらに文法などの言語情報を積極的に利用することとなる．電話音声への応用など，言語に依存する単語(“もしもし”，“hello”，など)の発話が期待される場合にはそれらの単語をワードスポッティング技術を用いて検出することにより言語識別を行うことが可能となる場合がある．また一般化して言語ごとに設計した大語彙音声認識システムを用いて音声認識を行うことによって言語を識別することも検討されている[18),19)]．また言語識別の前処理として必要となる言語音声区間の自動検出の研究も行われている[24)]．

米国の8つの研究機関(AT&T，ITT，Lockheed-Sanders，MIT，MIT Lincoln Lab.，Natural Speech Technologies，OGI，RPI)が参加してNIST(The National Institute of Standards and Technology)が1993年6月に言語識別技術の第1回の評価を行った[12)]．

OGI TSの11カ国データベースに対し，学習用音声に対する認識率は45秒の入力音声に対して89％，10秒の入力音声に対して79％であり，2言語の識別に対しては98％(45秒入力音声)，95％(10秒入力音声)と報告されている[16)]．一般に入力音声の継続長が大きくなるほど認識性能は向上する．OGI TSの10カ国語に対する人間の言語識別能力を評価した結果，ある程度の訓練の後，約6秒の音声に対して平均69.4％の識別率であった．人間は言語を識別するために音素の情報や単語の情報など各種のレベルの情報を組み合わせて行っていることが報告されている[15)]．

OGI TSは応用を想定して電話入力音声を用いているので，音声入力系の特性は統一されているわけではない．固定のマイクロホン入力を用いたよりSNRの高い音声に対してはより高い認識性能が期待される．

3.2 方言識別

方言を自動識別・分類するための研究も言語識別と同様に行われている[17)]．同一国内でも各地域での方言のため単一の音響モデル(音素モデル)の音声認識システムでは高い認識性能を得ることが難しい．そこで，方言認識(分類)を前処理として用いて認識性能を向上させる研究が行われている[22)]．また外国人の発話による言語の違い(native／日本人英語／中国人英語など)に関する研究も行われている[20)]．国内でも方言の自動認識は研究されている[23)]が，音声認識システムと接続した評価研究は少ない．方言の違いは，言語における違いと同様に音響レベルから言語(語彙)レベルまであるが，現在は韻律情報(声の高さ)が用いられている．国内の方言識別研究を進めるためには，今後方言識別用データベースの整備が必要であろう．

3.3 応用例

言語識別技術は多言語を取り扱う各種の応用が期待される．有力な応用は電話における多言語サービスのための言語振り分け☆，音声翻訳システムにおける言語選択などに応用が想定される．平成3年に東京地区の電話番号が1桁追加された時に，従来の電話番号で発信する使用者は新しい番号でかけなおすよう音声による案内を受けた．しかしながら日本語を理解できない外国人使用者はその案内メッセージを理解できないため混乱を生じた．この場合電話で話される音声を用いて言語識別が可能であれば識別結果に基づいた各種の言語による案内文を流すことも可能となる．また近年国際化が進み国内においても多くの外国人が居住している．彼らからの緊急電話を受ける際，その言語を理解できるオペレータに短時間で接続するためには言語識別が必要となる．AT&Tの140言語をカバーするLanguage Line Serviceにおいてタミル語でそのサービスを受けるまでに3分かかった経験が文献12)に報告されている．

2カ国語の識別に限定すれば高い精度で言語識

☆ その言語を理解できるオペレータに接続すること．

別が可能であるので，音声翻訳への応用は有力であると考えられる．たとえば，日本語／英語の音声翻訳では日本語入力音声を認識し翻訳し英語音声を出力するが簡単な文であれば日本人使用者は英語を話すことも可能である．全文音声翻訳ではなく音声翻訳支援を想定すると英語文入力の場合には翻訳処理をする必要はない．入力音声が英語であると判定された場合のみ翻訳処理を行うこととなる．

4. 音声キーに基づく検索

ここで述べる音声キーによる情報検索は，検索のキーワード(単語などの言語情報)を音声で指定する，もしくは検索システムを音声で制御する，などの従来から研究されている応用とは異なっている．音声に含まれる多様な情報を検索の対象とする必要性は今後のマルチメディア・コンピュータネットワーク環境における多様な情報検索のためにより高まっている．インターネットWWWで各種の検索エンジンが作成されているが，たとえば，アクセス可能なサイトの多様な情報の中から特定の話者の音声のみを検索するなどの応用が考えられる[27]．また従来の音声認識技術は音声検索という観点からの研究ではなかった．たとえば，FD，CD，MDさらに大容量化されるDVDなどのランダムアクセスが可能な媒体や留守番電話などのような音声蓄積サービスが普及するにつれて，音声特徴をキーとする情報検索は，今後の付加価値の高い音声によるサービスを提供することが可能であろう．そのような観点から音声特徴を検索キーとする音声検索について検討されている[25]．話者検索実験の結果，10単語を用いて検索キーを作成した時，音声長が5単語程度の長さがあれば，誤り率4％程度で検索できることが報告されている．

米国ではカーネギーメロン大学(CMU)のDLプロジェクトにおいて，ビデオの画像および音声を解析し，音声認識，画像認識，ビデオ画像のシーン切り出し，自然言語理解を用いた自動索引づけを行い内容の情報検索を可能とするInformediaが研究されている．これらは本学会誌の特集「ディジタル図書館」に詳しく報告されている[26]．今後は国内においても，電子図書館やインターネットにおける検索などの音声特徴による検索の必要性が高まり，それを支える研究が活発化するものと考えられる．

5. むすび

言語情報(発話内容)に着目する従来の音声認識研究において，雑音成分として取り扱われてきた個人性や言語の違いなどの，多様な情報カテゴリーを自動識別する研究例を取り上げ，その研究現状，応用例，将来の可能性について述べた．音声識別技術により新たな情報処理サービスが作り出されることを期待している．

参考文献

1) Shaughnessy, D. O': Speaker Recognition, IEEE ASSP Magazine, pp.4-17 (1986).
2) 杉山，村上：男女性音声特徴に基づく音声セグメンテーションとクラスタリング，音学講論，1-4-9，pp.17-18 (Mar. 1993).
3) 杉山：音響特徴量を用いた多言語音声識別の検討，音学講論，3-3-6，pp.81-82 (Mar. 1990).
4) McDonough, J. et al.: ICASSP94, 42.5 (Apr. 1994).
5) 松井，古井：テキスト指定形話者認識，電子情報通信学会論文誌，Vol.J79-D-II，No.5，pp.647-656 (1996).
6) Mammone, R. J., Zhang, X. and Ramachandran, R.P.: Robust Speaker Recognition, IEEE Signal Processing, Vol.13, No.5, pp.58-71 (Sep. 1996).
7) Yu, G., Siu, M. and Gish, H.: An Unsupervised, Sequential, Learning Algorithm for the Segmentation of Speech Waveform with Multiple Speakers, ICASSP92, 86.10 (1992).
8) Gish, H. and Schmidt, M.: Text-Independent Speaker Identification, IEEE Signal Processing, Vol.11, No.4, pp.18-32 (Oct. 1994).
9) 村上，杉山，渡辺：Ergodic HMMを用いた未知・複数信号源クラスタリング問題の検討，電子情報通信学会論文誌，DII，Vol.J78-D-II，No.2，pp.197-204 (Feb.1995).
10) Sugiyama, M.: Automatic Language Recognition Using Acoustic Features, ICASSP91, 16.S12.8, pp.813-816 (May 1991).
11) Nakagawa, S., Ueda, Y. and Seino, T.: Speaker-Independent, Text-Independent Language Identification by HMM, Proc. of ICSLP (Oct. 1992).
12) Muthusamy, Y. K., Barnard, E. and Cole, R.A.: Reviewing Automatic Language Identification, IEEE Signal Processing, Vol.11, No.4, pp.33-41 (Oct. 1994).
13) Berkling, K. M., Arai, T. and Barnard, E.: Analysis of Phoneme-Based Features for Language Identification, ICASSP94, 1-293 (1994).

14) Zissman, M. A. and Singer, E.: Automatic Language Identification of Telephone Speech Messages Using Phoneme Recognition And N-Gram Modeling, ICASSP94, 1-305 (1994).
15) Muthusamy, Y. K., Jain, N. and Cole, R.A.: Perceptual Benchmarks for Automatic Language Identification, ICASSP94, 1-333 (1994).
16) Zissman, M. A.: Language Identification Using Phoneme Recognition and Phonotactic Language Modeling, ICASSP95, Vol.5, pp.3503 (1995).
17) Zissman, M., Gleason, T., Rekart, D. and Losiewicz, B.: Automatic Dialect Identification of Extemporaneous, Conversational, Latin American Spanish Speech, ICASSP96, SP21A, II-777 (1996).
18) Mendoza, S., Gillick, L., Ito, Y., Lowe, S. and Newman, M.: Automatic Language Identification Using Large Vocabulary Continuous Speech Recognition, ICASSP96, SP21A, II-785 (1996).
19) Hieronymus, J. and Kadambe, S.: Robust Spoken Language Identification Using Large Vocabulary Speech Recognition, ICASSP97, Vol.2, pp.1111 (1997).
20) Arslan, L. and Hansen, J.: Frequency Characteristics of Foreign Accented Speech, ICASSP97, Vol.2, pp.1123 (1997).
21) 笹沼，板橋：韻律情報を用いた多言語音声の分類，音響学会講演論文集，3-6-14，pp.115-116 (Mar. 1997).
22) Brousseau, J. and Fox, S. A.: Dialect-Dependent Speech Recognizer for Canadian and European French, ICSLP92, Vol.2, pp.1003-1006 (1992).
23) 田中，板橋：基本周波数を利用した日本語方言音声の識別，音響学会講演論文集，2-4-7，pp.19-20 (Mar. 1995).
24) 水野，高橋，嵯峨山：スペクトルの動的および静的特徴量を用いた言語音声の検出，音響学会講演論文集，3-2-1，pp.107-108 (Sep. 1995).
25) Sugiyama, M.: Speech Database Retrieval Using Speech Key, Proc. of ICSPAT95, pp.1916-1919 (Oct. 1995).
26) 特集：ディジタル図書館，情報処理，Vol.37, No.9, pp.813-864 (Sep. 1996).
27) 大室他：インターネットと音声・オーディオ処理技術，日本音響学会誌，Vol.52, No.8, pp.631-636 (1996).

(平成9年9月8日受付)

杉山 雅英 (正会員)

1954年生．1977年東北大学理学部数学科卒業．1979年同大学院理学研究科数学専攻修士課程修了．同年日本電信電話公社武蔵野電気通信研究所(現在NTT武蔵野研究センター)入所．1985年東北大学より工学博士号を取得．1986年米国AT&T Bell研究所滞在研究員，1987年NTT基礎研究所主任研究員，1990年ATR自動翻訳電話研究所主幹研究員の後，1993年会津大学コンピュータ理工学部ヒューマンインタフェース学講座教授．現在まで，LPCスペクトル距離尺度(歪み尺度)，ベクトル量子化による音声認識，特徴ベースによる音声認識，教師なし話者適応，テキスト独立話者認識，音声スペクトル推定，情報幾何学(微分幾何学)による音声分析，音響特徴量による言語識別，音声特徴キーによる音声検索などの音声認識処理の研究に従事．日本音響学会，電子情報通信学会，人工知能学会，IEEE各会員．e-mail:sugiyama@u-aizu.ac.jp