論文 / 著書情報
Article / Book Information

論題(和文)	音声トランスクリプションのこれまでと今後の展望
Title(English)	Voice Transcription -Past, Present and Future-
著者(和文)	古井貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	電子情報通信学会 論文誌D-II, Vol. J83-D-II, No. 11, pp. 2059-2067
Citation(English)	The Transactions of the Institute of Electronics, Information and Communication Engineers D-II, Vol. J83-D-II, No. 11, pp. 2059-2067
発行日 / Pub. date	2000, 11
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) 2000 Institute of Electronics, Information and Communication Engineers.

招待論文

音声トランスクリプションのこれまでと今後の展望

古井　貞熙*

Voice Transcription ——Past, Present and Future ——

Sadaoki FURUI†

あらまし　統計的パターン認識理論，ケプストラムとデルタケプストラム，隠れマルコフモデル，統計的言語モデルを基本とする音声トランスクリプション（音声文字変換），広くは大語彙連続音声認識の技術が近年大きく進展し，音声ワープロ，放送への自動字幕付与，情報検索への適用など，種々の応用が展開しつつある．本論文では，これまでの研究の展開と基本技術，残されている課題について解説し，今後の展望を述べる．特に，種々の変動を含む話し言葉に対応できる適応的言語モデル及び音響モデルの構築，そのための話し言葉の大規模コーパスの作成，音声から話題語を抽出し，話し手の意図を理解し，要約する機能の実現の重要性などに焦点を当てる．

キーワード　音声トランスクリプション，大語彙連続音声認識，統計的パターン認識理論，意図理解

1. まえがき

音声は，人と人との最も自然で，基本的，容易かつ効率的な情報交換手段である．人とコンピュータとの情報交換においては，コンピュータから情報を受け取るには一般にディスプレイが最も効率的であるが，コンピュータへの入力手段としては，音声が使えれば極めて効率的になると期待される．特に日本語に関しては，優れた使いやすいワープロ（ソフト）が普及しているとはいえ，仮名・漢字変換を必要とするため，英語のようにしゃべる速さに近い速度でタイプ入力するのは極めて難しい．音声認識技術を用いた日本語音声トランスクリプション（音声から文字への変換）ができれば，単に音声ワープロとしての実現形態だけでなく，放送ニュース音声への自動字幕付与，会議の議事録，講演録，放送ニュースのアーカイブ化への利用など，数多くの応用が考えられる．

音声認識の研究は，1950年代から現在まで半世紀にわたって，世界中の研究者によって進められてきた．近年，統計的パターン認識理論に基づく音声認識技術の発展と，コンピュータの高速化に支えられて，種々の言語に対応した音声ディクテーションソフトが開発され，普及しつつある．しかしまだその能力には限界があり，研究課題は山積している．今後は，話し言葉に対するトランスクリプションの基本的性能を高めるだけでなく，話し手の意図を理解し，要約したりする機能へと高度化を進めていく必要がある．本論文では，現在の音声トランスクリプションの原理，これまでの研究の流れ，今後の展望について述べる．音声認識の研究には，コンピュータシステムとの音声対話システムの研究があるが，音声トランスクリプションの研究は，このような技術の基礎としても重要である．

2. 音声生成モデルと音声認識の原理

図1に，通信理論のアプローチからの，音声生成モデルを示す[1]．メッセージ（意図）情報源で，意図したメッセージMが生成され，言語チャンネルによって単語系列Wに変換される．同じメッセージを表すには種々の方法があるので，言語チャネルは，代表的な単語列を中心として確率的に機能する．単語列Wは，調音チャネルによって，音の系列Sとして実現される．調音器官は話者によって異なり，同じ話者でも全く同じ波形を繰り返すことはできないので，調音チャネルも確率的に機能する．音の系列Sは話者の口から放射され，部屋の音響特性が畳み込まれた後，音響入力Aとしてマイクロホンに到達する．この過程をここでは音響チャネルと呼んでいる．音響入力信号Aは，マイクロホンによって電気信号に変換され，種々の雑音やひずみを受けて，音声認識系にXとして入力される．この最後の過程を，ここでは伝達チャネルと呼んでいる．以上の各チャネルには，不確実性あるいは変動があり，

* 東京工業大学大学院情報理工学研究科計算工学専攻，東京都
Department of Computer Science, Tokyo Institute of Technology, Tokyo, 152-8552 Japan

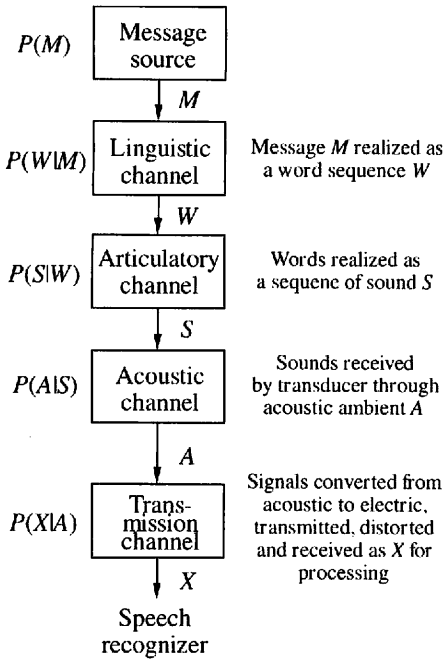

図1 通信理論のアプローチからの音声生成モデル
Fig. 1 A communication-theoretic formulation of the speech production chain.

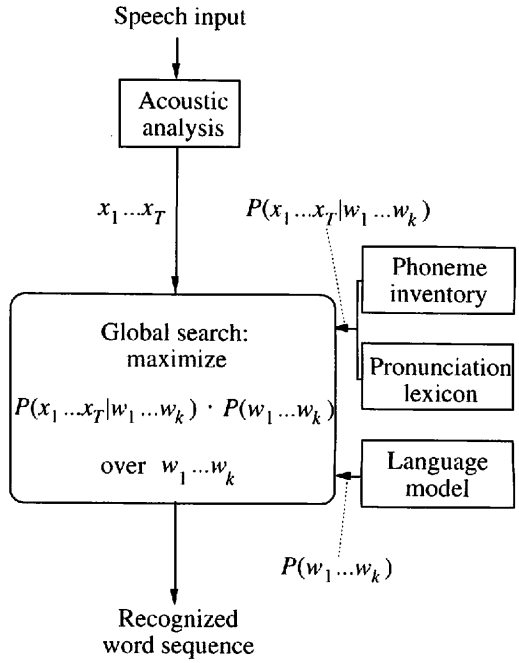

図2 現在のトランスクリプションシステムの構成
Fig. 2 Mechanism of state-of-the-art speech recognizers.

$P(W|M)$, $P(S|W)$, $P(A|S)$, $P(X|A)$の確率分布で表現される．$P(M)$はメッセージの事前分布である．究極の音声認識システムは，Xから，以上の過程を逆変換してMを回復することを目的とする．

現在のトランスクリプションシステムでは，ベイズ決定理論に基づくパターン認識理論に従って，$X=x_1, x_2, ..., x_T$が与えられたときの事後確率$P(W|X)$を最大にする単語列$W=w_1, w_2, ..., w_k$が選ばれる[2]．これは，事後確率最大化決定(maximum a posteriori (MAP) decision)と呼ばれる．このとき，確率分布が正確に与えられるなら，誤り率が統計的に最小になることが保証されている．事後確率は，ベイズの定理により，次のように展開される．

$$P(W|X) = P(X|W)P(W)/P(X) \quad (1)$$

ここで，条件付き確率$P(X|W)$は，単語列Wを音素などの基本的単位の連鎖で表した音響モデルから，特徴パラメータ列が出現する確率として計算され，単語列Wが発声される事前確率$P(W)$は，言語モデル（文法）によって計算される．通常，特徴パラメータとしては，ケプストラム及びパワーとその1次及び2次回帰係数（多項式展開係数）が用いられる．音響モデルとしてはHMM（隠れマルコフモデル）が用いられ，その学習には多数の人が発声した音声が用いられる．言語モデルとしては，多くの場合，統計的言語モデルが用いられ，その学習には大量のテキストが用いられる．$P(X)$はWを変数とする最大化とは無関係なので，図2に示すように，トランスクリプションは，次の$\hat{W}$を選ぶプロセスになる．

$$\hat{W} = \arg\max_{W} P(X|W)P(W) \quad (2)$$

3. 音声トランスクリプション研究のこれまで

3.1 米国における先行研究

近年の音声トランスクリプション研究の牽引車は，何といっても，Wall Street Journalをはじめとする北米の種々の新聞を読み上げた音声を対象として行われたDARPA(Defense Advanced Research Projects Agency)のプロジェクトである[3]．このプロジェクトは1991年に開始され，文脈依存音素HMMと単語を単位とする統計的言語モデルを基本とする，数万語規模の大語彙連続音声認識の研究が活発に行われた．メルケプストラム，状態共有HMM，言語モデルのバックオフ平滑化など，現在の音声トランスクリプション技術の基本の多

くは，このプロジェクトによって確立したといえる．それまでの人手に頼ることを前提にしていた文法を用いるよりも，認識性能が大きく向上し，65,000語の語彙を対象とした場合に，93%の単語正解精度が報告されている．同時にこのプロジェクトでは，現在の方法による認識性能が，極めて直接的に学習データ量（特に言語モデル用コーパス）に左右されることが明らかになった．

3.2 日本における先行研究

我が国においては，単語ではなく，音素，音節，仮名・漢字などを単位とする統計的言語モデルを用いた研究[4]～[6]が，まず行われた．その主たる原因は，(1) 電子的な形で自由に使える大量の新聞テキストがなかった，(2) 日本語テキストは，英語と違って，単語を単位としてスペースで区切られていないので，大量のテキストにこのような区切りを自動的に入れることが必要であるが，そのための優れたコンピュータプログラムがなかった，(3) 日本語は英語と言語構造が違うので，英語と同じ方法を使っても，同じようにはうまくいかないのではないか，という考えがあったことなどにある．

仮名・漢字のトライグラムを用いる方法は，それまでの音素や音節を単位とする方法に比べて，後処理として仮名漢字変換をする必要がなく，更に漢字の読みのコンテクスト依存性を考慮して，漢字を読み別に分けて求めた言語モデルを用いることにより，パープレキシティと文字変換率が，ともに向上することが確認されている[6]．

3.3 日本における形態素を単位とする研究の立上げ

筆者らがNTTの研究室で，日本語音声のトランスクリプションの研究に着手したのは，1995年の3月である．その当時，米国では上記の新聞を読み上げた音声を対象とした，DARPAのプロジェクトが活発に研究を行っていたが，当時の日本では，日本語を対象としたこのような研究はまだほとんど着手されていなかった．

しかし筆者らは，英語でうまくいく方法が，日本語でうまくいかないはずはないと考え，何とか同じようなことができないか，いろいろ模索していた．ディクテーションシステム（ソフト）そのものにそれほど興味があったわけではないが，これからの大語彙音声認識技術の基礎として，あるいはベンチマークとして，研究することが必要だと考えた．そして，1995年の3月，NTTの研究所で，優れた形態素解析プログラムが開発されると同時に，その開発に使われた日経新聞の5年間にわたるテキストが，電子的に使えるようになった．Wall Street Journalと同様の経済新聞が使えるということは，英語に関する研究結果と比較するにも好都合であった．そこで直ちに，研究用のデータベースの構築にとりかかり，これを用いた音声認識実験を始めた．言語モデルとしては，単語（形態素）トライグラムを用いた．デコーダとしては複数のプログラムを並行して試したが，結局HTK（英国Cambridge大学で開発されたソフトウェア）が比較的良い結果を示し，この結果をその年の秋に発表した．予想どおり，英語の場合にほぼ匹敵する認識結果を得ることができた[7]．このような筆者らの研究と並行して，日本IBMでは米国のIBMと協力して，日本語音声のディクテーションソフトの研究開発が進められ，今日の製品へ結実している．やや遅れてNECでも同様の製品が開発されている．

これらの研究をきっかけとして，その後我が国では，新聞の読み上げ音声を対象としたトランスクリプション研究に関して，国内のボランティアグループによるデータベースとデコーダの開発が進められ，「日本語ディクテーション基本ソフトウェア」という形で成果が取りまとめられている[8]．これは，認識エンジン(Julius)，日本語音響モデル，言語モデル，及び形態素解析・読み付与ツールから構成され，優れた性能を示している．無償で一般に公開されており，今後，大語彙連続音声認識研究及び開発の共通プラットフォームとして広く使われると予想される．

3.4 放送音声のトランスクリプション

1996年早春のDARPAワークショップでは，新聞記事の読み上げタスクから，Hub4と呼ばれる放送音声のトランスクリプションに焦点が移った．新聞読み上げタスクとの基本的違いは，対象とする音声が"real-world speech data"であることである．アナウンサーが原稿をもとに静かな環境で発声した音声だけでなく，音楽・雑音・音声が重畳したり連続している音を対象としている．米国を中心に，欧州を含む大学及び研究機関がプロジェクトに参加している．語彙数は65,000語が標準的で，標準の言語モデルはないが，CMUで開発されたツールが広く使われている．

このDARPAの動きに刺激され，我々も何とか日本語の放送音声のトランスクリプション実験ができないかと考えた．そこで直ちにNHK放送技術研究所に話をもちかけたところ，国内のいくつかの大学を含むグループで，共同で研究を始めることになった．NHK研究所から過去の放送ニューステキストデータベースと，実

際の放送音声のデータを提供して頂き，研究が開始された．

我々のニュース音声の認識システム[9]では，各音素のHMMを，男女別に約50名が発声した約13,000文の音声を用いて学習している．統計的言語モデルとしては，過去約5年間のニュース原稿に含まれる比較的頻度の高い20,000種類の単語について，トライグラムを計算して用いている．これまでの日経新聞の記事から作成したトライグラムを組み合わせて用いることを試みたが，これはうまくいかなかったため，ニュース原稿のみから言語モデルを作成している．実験の結果，スタジオでアナウンサーが原稿を見ながら発声した音声であれば，90%近い認識率を得ることができた．

日本語では同じ単語に複数の読みがあり，正しい読みは前後関係（コンテクスト）から決定される性質がある．ベースラインの言語モデルでは，読みが違っていても，表記や品詞が同じであれば同じエントリとして扱い，異なる読みに対して同じ確率を与えていた．しかしこれでは，読みに対するコンテクストの影響を考慮することができず，まれな読みにも他と同じ確率が与えられてしまうため，奇異な認識誤りを生ずる問題があった．この問題を解決して，音声認識性能を向上させるため，同じ文字が使われていても読みの違う単語は別の単位とし，統計的言語モデルを作成した．更に，話し言葉には必ず挿入される，「え」「えー」などの間投詞が，文頭や読点の位置に生じやすい性質に着目して，大量の言語テキストのこれらの位置に間投詞の発音を確率的に与えた．この両者の対策により，放送ニュース音声の認識性能を向上させることができた．

3.5 放送音声への字幕付与

NHKでは，音声認識を用いてニュース音声に字幕を付与するシステムの運用を，2000年3月から開始した[10]．音声認識では，誤認識を完全に回避することはできないため，認識結果を人が即座にチェックして，誤認識を素早く修正してから放送するシステムになっている．修正が素早く実行できるレベルとして，95%の単語正解精度が必要で，アナウンサーが原稿を読んでいる音声に関してはこれが達成されているが，それ以外のインタビューなどへの音声認識の利用は今後の課題である．ニュースでは日々新しい語彙（システムにとっては未知語）が生ずるため，これに対処するための方法や，文末にデコーダの処理が到達するまで認識結果が確定しない従来の方法では具合が悪いので，文中でも常に一定の遅れ以内で認識結果を確定しながら処理を進めるデコーダなどが開発されている．

3.6 ディクテーションソフト（音声ワープロ）

商品としてのディクテーションソフト（音声ワープロ）に関しては，1990年に米国のDragon Systemsが単語区切りで発声した音声を入力とするPC用プログラムを発売し，その後，IBM, L&H, Philipsなど種々の会社が参入した．単語区切りでないと入力できないのはユーザへの負担が大きいので，1996年ごろから単語を続けて発声できる製品が開発されるようになった．これらの製品の多くは，英語版だけでなく，ドイツ語，フランス語，イタリア語，スペイン語，中国語，日本語版などに発展している[11]．主な利用者は，障害者，X線医師（医療所見の入力），弁護士などである．

3.7 情報検索への応用

米国のDARPAのプロジェクトを中心に，音声認識と情報検索の技術を組み合わせる研究が，最近活発に行われている[12]．その一つは，放送ニュースの音声を音声認識によってトランスクリプションした結果としての単語列から，その話題（トピックス）を表す重要語やフレーズを自動的に検出し，それを用いてインデクシングを行うことである．これにより，大量のニュースを話題に応じて分類するなど，自動的にアーカイブ（データベース）化することができる．このために，放送やビデオの中から，音声の部分だけを取り出す研究，更にある特定の話題に関する部分だけを取り出す研究も行われている．

4. 音声トランスクリプション技術の今後の課題

4.1 話し言葉認識の難しさ

現在の音声認識技術で，放送音声をトランスクリプションすると，上述のように，スタジオでアナウンサーが原稿を見ながら発声した音声であれば，90%あるいはそれ以上の認識率が得られるが，現場からの中継などの音声や，対談の音声に対しては大幅に認識率が低下する[9]．この原因は，発声が必ずしも明確でないこと，雑音が重畳していること，原稿を読んでいない自発的な音声は，言語モデルの学習に用いたニュース原稿の文とはかなり言語的構造が異なることなどである．

話し言葉特有の難しさには，次のようなものがある．

(a) 助詞などの言葉の脱落・省略，間投詞などの不要語の付加，倒置への対処

(b) 言い間違い，言いよどみ，言い直し，繰返しへ

の対処

(c) 言語モデルには含まれていないが，意味のある未知語（新しい言葉）の発声への対処

これらの問題への対処法としては，

(i) 単語やフレーズを単位としたスポッティング(spotting)あるいは検出(detection)と，発声確認(utterance verification)のアプローチ

(ii) 余計な言葉をスキップしながら構文解析する方法

(iii) 未知語検出アルゴリズム

などが研究されている．ただし，現在の中心的方法である統計的言語モデルとマッチする効率的な方法はまだない．

4.2 話題語抽出

究極の音声認識は，**2.**で述べたように，発声者の意図Mを抽出することである．その簡便な方法としては，話題語（キーワード）あるいはキーフレーズの抽出がある．我々は，ニュース音声を自動的にトランスクリプションし，その認識結果としての単語列から話題語を自動抽出する試みを行っている[9]．

この際，ニュースの話題語はほとんどが名詞であるので，トランスクリプション結果から名詞のみを抽出し，その中から，重要度の尺度によって話題語を抽出する方法を検討した．重要度の尺度は，大量の学習用ニュース記事に出現する各名詞の出現頻度に基づく情報量（一種のtf-idf尺度）によって定義した．すなわち，話題語を抽出すべきニュース音声（1ないし複数の文から構成される）の認識結果から，情報量の大きい単語を上位から順に抽出した．人がニュース文を見て抽出した話題語を正解として評価した結果，話題語を上位5個(recall: 13.0%)抽出すると，84.1%のprecisionが得られることがわかった．

話題語に関する重要度をどのように与えるか，単語や句が担う意味的情報をどのように用いるかが，今後の重要な課題である．

4.3 発声者の意図理解の枠組み

これまでのトランスクリプションの枠組みから，発声者の意図を抽出する枠組みに変更するには，これまでのように$P(W|X)$を最大化するのではなく，図1に示したモデルで，$P(M|X)$を最大化するMを選ぶことが必要である[1]．これにより，単純に$P(W|X)$を最大化するよりも，Wの正解精度を上げることができる可能性もある．筆者らは最近，このための新しい定式化を提案し，実験を進めている[9]．

この定式化は，これまでに試みられてきた種々の方法を包含し，かつ新しい方法を示唆する．すなわち，条件付き確率$P(W|M)$の表現法には，これまでに研究された話題別の言語モデル，cacheモデルなど種々の方法が考えられる．筆者らは，Mとして連続量を扱うことができ，しかもそれを明示的に表現する必要がない方法として，Mは単語の共起によって表されると考え，これによる定式化を行った．更にこの方法を放送ニュース音声に適用し，これによって結果的に単語列の正解精度を向上させる試みを行っている．

意図理解の結果をデコーダにフィードバックし，その結果に基づいて意図理解をやり直し，この結果を再びデコーダにフィードバックするという方法で，MとWの両方の精度を上げることができる可能性もある[1]．

4.4 発声検出に基づく音声理解のアプローチ

不要語や間投詞を多数含む話し言葉から，意味のある部分だけを抽出（意味抽出，意味検出）する方法として，単語及びフレーズの検出に基づく音声理解のアプローチが研究されている[13], [14]．この原理を図3に示す[1]．各検出器(detector)では，あらかじめ規定された各イベント（音素，単語，フレーズなど）と，その対立モデルに対するゆう度比に基づき，Neymann-Pearson lemmaに従って検出判定が行われる．モデルの学習には，通常の最ゆう法あるいは識別学習が用いられる．各検出器の出力は，その信頼度とともに探索部

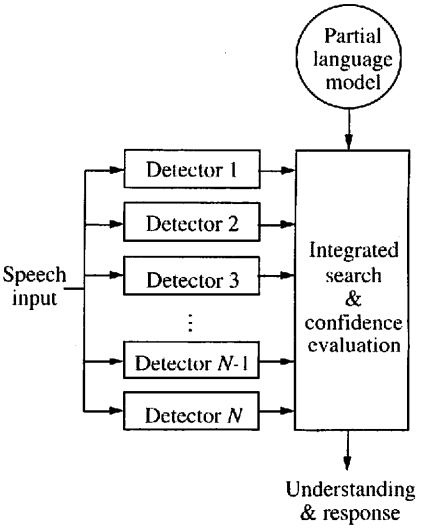

図3 音声検出・確認に基づく音声理解システム

Fig. 3 An architecture of a detection-based speech understanding system.

へ送られ，言語モデルとの組合せによって音声理解の判定が行われる．

不要語を吸収する音響モデルや，その単語やフレーズの受理を判定するための尺度を適切に構成することによって，よい性能を得ることができることが実験的に示されている．ただしこれまでの実験は，中規模程度の語彙数のシステムに対してであり，大語彙のシステムにどのように適用し，適度な処理量で，これまでの方法をどれだけ上回れるかは今後の課題である．

この研究は，音声認識結果からの話題語（キーワード）抽出とともに，次で述べる音声要約のための一つのアプローチとなろう．

4.5 音声要約

発声者の意図Mの表現形態の一つとして，音声の要約作成がある．筆者らは音声中から，話題を担っている重要語（話題語）をできるだけ文法的制約に従うように最適に選んで要約文を作成する方法として，各単語の重要性スコアと要約文の統計的言語スコアをもとに，動的計画法を用いて単語を選ぶ方法を提案した[15]．重要性スコアには，**4.2**で述べた尺度を用いた．統計的言語スコアは，音声よりも比較的簡潔な文体となっていると考えられる新聞（毎日新聞）の1年分の記事から計算した単語トライグラムを用いた．

放送ニュース音声の音声認識結果を用いて，文字数にして60～70%に要約する条件で実験を行った．8名の被験者による評価の結果，もとの音声中の重要単語のうち86%が要約文に保持され，72%の要約文が原文の意味を保存していることが確認された．これにより，提案方法の有効性が示された．本方法は，ニュースの自動字幕作成への適用のほか，議事録作成など，今後の種々の応用分野への展開が期待できる．ニュース音声，講演，講義などを音声認識技術を用いて自動的に要約してデータベース化することができれば，情報検索に限らず種々の用途に極めて有用であろう．音声メールを自動的に要約して文字化できれば，内容を一覧するために役に立つであろう．

4.6 講義や講演の記録作成

コンピュータシステムなどへの入力手段として，音声には多くのメリットがあるが，誤認識が完全には回避できない，騒音を発生するので周囲に迷惑をかける，他人に聴かれてしまう，などの欠点もある．このような意味では，講演，講義，会議などでは音声は永久的に不可欠であるから，このような場面での音声認識技術の利用が，これから重要な分野の一つとなってくるであろう．トランスクリプション技術によって議事録や講演録ができ，更に自動的に内容を要約することができれば，極めて有用であろう[16]．

4.7 音声の音響的変動への対処

4.7.1 音声変動の影響

音声認識で誤認識を生ずるのは，音響モデルや言語モデルの学習に用いたデータと，実際の認識すべき音声の間に，何らかのずれ(mismatch)があるためである．音響的変動の原因には，音声（声質や話し方）の個人差（方言や年齢によるものも含む）が極めて大きいこと，同じ人でも体調や精神状態（ストレス）などによって発声速度や話し方が変化すること，多くの場合に，部屋の反響を含む種々の雑音が音声に加わっており，しかもそれが時間的に変化すること，更にその上にマイクロホンや電話機，伝送系などのひずみが加わることなどがある[2]．

これらの種々の音声の変動をカバーするような大量の音声データを用いて，統計的方法によりモデルを作成すれば，ある程度高い認識精度を達成することができる．しかし，集められるデータ量には自ら限界があり，すべての変動条件を尽くすことは不可能である．更に，データに含まれる変動があまりに大きいと，モデルがぼけて，異なる音素や単語の物理的特徴の重なりが大きくなるため，この方法には限界がある．このため，「Sheep and goats 現象」といわれるように，大多数の話者に対してはうまく動作しても，少数ではあるが認識精度が極めて悪い話者や条件を生ずることになることが多い．

4.7.2 音声変動への適応化

これらの種々の変動に対処するため，音声認識システムに，少量の音声サンプルに基づいて，変動に自動的に追随してくれるような，適応機能を持たせる研究が広く行われている．一般に適応化アルゴリズムには，その規則を学習するためにあらかじめ決められた言葉を発声してもらう「教師付き(supervised)学習」と，任意の音声を用いて規則を学習する「教師なし(unsupervised)学習」がある．前者の方が規則の獲得は比較的容易であるが，ユーザにとっては，認識装置を使う前にわざわざ特定の言葉を発声しなければならないのは煩わしい．後者の方が，認識すべき音声をそのまま使って適応化できるので，ユーザにとっては全く意識せずに使えて便利である．極めて個人差やひずみの大きい音声に対しても，この機能が実現できるアルゴリズムを追及していくべきであろう．更に加えて，認識装置

を使っているうちに，自動的に適応化が進んでいくオンライン逐次型適応が望ましい．

筆者らはこのような動機から，話者の交代を自動的に検出しながら，音素モデル（隠れマルコフモデル：HMM）をオンラインで，逐次的に教師なしで，入力音声に自動的に適応させる方法を提案した．話者適応には，ゆう度最大化線形回帰法(MLLR)に，MAP推定とベクトル空間平滑法(VFS)を組み合わせて用いた．放送ニュース音声を用いた実験の結果，本方法によって音声認識性能が向上することを確認した[17]．

音声は通常，スペクトルあるいは対数スペクトル領域のパラメータで表現されるが，上で述べた種々の音声の変動は，これらの領域での線形あるいは非線形の変化として表される．更に，その変化が，もとの音声パラメータに対して，加算的な場合と乗算的な場合がある．また，その影響が種々の音素に共通に現れるものと，音素に依存して異なるものとがある．雑音やマイクロホンなどによるひずみは音素に共通であるが，個人差は音素によって異なる．この違いによって，適応化技術も異なってくる．また，これらの種々の変動は，単独にではなく複合して音声に加わるのが普通である．したがって，これらに同時に対応できる適応化法を構築することが必要である．

適応化法は更に，音声の波形や特徴パラメータの領域で適用する方法と，音響モデルの領域で適用する方法がある．前者は発声内容に無関係に適用できるので容易であるが，音素に依存した適応化ができない限界がある．両者の特徴を生かして，うまく組み合わせることが必要であろう．

これまでの話者適応化では，不特定話者の音素モデルを話者に適応させるケースが多い．ある特定の話者のモデルを適応化させるよりは，この方が良い結果が得られるが，ぼけたモデルの適応化では自ずから性能に限界がある．適応化前の初期モデルをどのように用意するかも，これからの重要な課題の一つとなるであろう．更に基本的には，音声の個人差とは何なのかを少しでも明らかにし，そのモデルに基づいた適応化法を構築すべきであろう．

4.8 話し言葉コーパスの構築

統計的言語モデルを構築するためには，使われる対象に対応した大量のテキストが必要である．上でも述べたように，NHKのニュース音声を高い精度で認識するためには，新聞記事から作成した言語モデルでは不十分である．新聞やニュースのように比較的書き言葉に近い音声と，通常の話し言葉の間には大きなギャップがある．このため，話し言葉の統計的言語モデルを作るためには，実際の話し言葉を書き起こした大量のテキストが必要である．米国では，DARPAのプロジェクトの中で，“Switchboard”，“Call Home”などと呼ばれる，自然な会話音声を書き起こしたテキストデータベースが作られ，これを用いた音声認識研究が行われている．

我が国でも最近，科学技術振興調整費開放的融合研究推進制度による「話し言葉の言語的・パラ言語的構造の解明に基づく『話し言葉工学』の構築」プロジェクトが開始され，音声認識の高度化のための音声言語データベース（コーパス）の構築，解析，認識，要約などの研究が推進されている．時期を同じくして，文部省中核的研究拠点(COE)「名古屋大学統合音響情報研究拠点CIAIR」も開始され，空間音響信号処理，音声分析合成，音声認識，自然言語による対話，音情報の認知といった研究を有機的に統合した研究が開始された．これらの大型プロジェクトが，話し言葉を音声認識するための着実なステップを提供することが期待されている[18]．

4.9 言語モデルのタスク適応

音声認識の対象タスクが変わるごとに，言語モデルの学習のために大量の（書き起こし）テキストを集めるのは決して容易ではない．このため，共通の基本的な言語モデルを，限られた量のテキストを用いて，異なる対象に適応化させる研究も行われている．言語モデルの適応化のためには，テキストや言語モデルの中で，タスクに依存する部分と，依存しない一般的な部分を分ける必要があり，このような研究も始まっている[19]．

新しいタスクへの適応と関連して重要な課題に，新しい言葉（システムにとっての未知語）をいかにして言語モデルに組み込むかという問題がある．最も一般的な方法は，あらかじめ単語を品詞や意味を用いてクラスに分類しておき，未知語をその中の一つのクラスに対応させてクラス言語モデルを用いる方法である．単語クラスの作り方としては，細分類された品詞，ゆう度最大化基準，潜在的意味解析（latent semantic analysis），エントロピー最大化基準，相互情報量，シソーラスなどを用いる方法や，これらを組み合わせる方法が試みられているが，まだ決定的な方法はない．

5. むすび

これまで述べたように，トランスクリプションシステムには，極めて多様な応用が期待できるが，その実現のためには，解決しなければならない課題も多い．主たる課題は，話し言葉特有の言語的，音響的変動に対するロバストなモデルとアルゴリズムの構築である．その基礎となる話し言葉の大規模コーパスの構築も極めて重要である．本論文では触れなかったが，音声対話システムでも，音声認識部は，最近ではトランスクリプションシステムと同様に，大量のコーパスを作成し，それから作成した統計的言語モデルを用いる方法が主流になりつつある．もちろん極めて多様な言い回しに対処する種々の工夫が必要であるが，トランスクリプションとの基本的違いは小さくなりつつあり，この意味でもトランスクリプション技術は音声認識の基本といえる．

今後，音声トランスクリプションの研究は，単に音声を単語列に変換するだけでなく，発声者の意図の理解，話題語（フレーズ）抽出，内容の自動要約へと展開し，更に幅広い応用が開けてくると期待される．

コンピュータの小型化，高性能化，低価格化により，21世紀のはじめにはubiquitous computing（遍在計算）とwearable computing（ウェアラブル計算）の時代がくるといわれている[16]．このような環境下では，キーボード入力はなじまないので，音声認識がヒューマンインタフェースの基本的手段の一つとして広く使われるようになると期待される．この際，音声認識の多くは，個人が身に付け，個人に特化したマイクロホン及びコンピュータで行われるようになり，そのシステム構成は，これまでとは大きく異なった形になると予想される．雑音などの環境情報は常時モニタでき，タスクに依存した知識は個人のコンピュータへ即座にダウンロードできるので，タスクへの適応性とロバスト性の向上が期待できる．ただし，それらの実現のためには，まだ解決しなければならない課題も多く，今後の多角的な研究の展開が期待される．

文　献

[1] B.-H. Juang and S. Furui, "Automatic recognition and understanding of spoken language — A first step towards natural human-machine communication," IEEE Proc. (to be published).

[2] 古井貞熙，音声情報処理，森北出版，1998.

[3] 古井貞熙，"大語彙連続音声認識の現状と展望，"日本音響学会春季研究発表会講演論文集，1-6-10，March 1998.

[4] T. Kawabata et al., "Japanese phonetic typewriter using HMM phone recognition and stochastic phone-sequence modeling," IEICE Trans., vol.E74, no.7, pp.1783-1787, July 1991.

[5] S. Makino, A. Ito, M. Endo, and K. Kido, "A Japanese text dictation system based on phoneme recognition and a dependency grammar," IEICE Trans., vol.E74, no.7, pp.1773-1782, July 1991.

[6] 山田智一，松永昭一，川端　豪，鹿野清宏，"音声認識における仮名・漢字文字連鎖確率に基づく統計的言語モデルの利用，"信学論(A)，vol.J77-A，no.2，pp.198-205，Feb. 1994.

[7] 松岡達雄，大附克年，森　雅至，古井貞熙，白井克彦，"新聞記事データベースを用いた大語い連続音声認識，"信学論(D-II)，vol.J79-D-II，no.12，pp.2125-2131，Dec. 1996.

[8] T. Kawahara, T. Kobayashi, K. Takeda, N. Minematsu, K. Itou, M. Yamamoto, A. Yamada, T. Utsuro, and K. Shikano, "Sharable software repository for Japanese large vocabulary continuous speech recognition," Proc. ICSLP, Fr4A1, Nov. 1998.

[9] K. Ohtsuki, S. Furui, N. Sakurai, A. Iwasaki, and Z.-P. Zhang, "Improvements in Japanese broadcast news transcription," Proc. DARPA Broadcast News Workshop, pp.231-236, Feb. 1999.

[10] 今井　亨，小林彰夫，佐藤庄衛，安藤彰男，"逐次2パスデコーダを用いたニュース音声認識システム，"信学技報，SP99-129，Dec. 1999.

[11] 西村雅史，"日本語ディクテーションシステムの現状と今後の課題，"信学技報，SP99-94，Dec. 1999.

[12] S. Furui, "Automatic speech recognition and its application to information extraction," Proc. ACL '99, pp.11-20, June 1999.

[13] T. Kawahara, C.-H. Lee, and B.-H. Juang, "Combining key-phrase detection and subword-based verification for flexible speech understanding," Proc. ICASSP 97, pp.1159-1162, 1997.

[14] B.-H. Juang, "From speech recognition to understanding: Shifting paradigm to achieve natural human-machine communication," Proc. 16th ICA and 135th Meeting ASA, pp.617-618, 1998.

[15] 堀　智織，古井貞熙，"話題語と言語モデルを用いた音声自動要約法の検討，"信学技報，SP99-110，Dec. 1999.

[16] 古井貞熙，"Ubiquitous/Wearable Computing環境における音声認識の展望，"信学技報，SP99-114，Dec. 1999.

[17] S. Furui, Z.-P. Zhang, and K. Ohtsuki, "On-line incremental speaker adaptation for broadacst news transcription," IEEE Workshop on Automatic Speech Recognition and Understanding, pp.165-168, Dec. 1999.

[18] 板橋秀一，藤崎博也，山本誠一，板倉文忠，古井貞熙，広瀬啓吉，田中穂積，市川　熹，"パネル討論：音声言語関連大型プロジェクトの現状と将来，"信学技報，SP99-130，Dec. 1999.

[19] 河原達也，石塚健太郎，堂下修司，"会話スタイル依存の言語モデルを用いたキーフレーズの検出・検証，"信学技報，SP97-78，Dec. 1997.

（平成12年2月2日受付）

古井　貞熙（正員）

昭43東大・工・計数卒．昭45同大大学院修士課程了．同年NTT電気通信研究所入社．以後，同研究所において，音声認識，話者認識，音声知覚などの研究に従事．昭53～54ベル研究所客員研究員．昭61NTT基礎研究所第四研究室長．平1 NTTヒューマンインタフェース研究所音声情報研究部長．平3同研究所古井特別研究室長．平6東京工業大学客員教授．平9東京工業大学大学院情報理工学研究科計算工学専攻教授．工博．平1科学技術庁長官賞，IEEE ASSP Society Senior Award各受賞．本会より，昭50米沢賞，昭63，平5論文賞，平2著述賞各受賞．日本音響学会より，昭60，62佐藤論文賞等各受賞．平8 IEEE Signal Processing Society Distinguished Lecturer．著書「ディジタル音声処理」，「Digital Speech Processing, Synthesis, and Recognition」，「音響・音声工学」，「音声情報処理」，編著「Advances in Speech Signal Processing」など．IEEE及び米国音響学会(ASA)Fellow．Journal of Speech CommunicationのChief Editor．ISCA理事．日本音響学会理事・技術委員長．