6 音声認識実用化に向けた高次言語モデルの検討

花沢　健
NEC 情報・メディアプロセッシング研究所

音声認識における言語モデルの重要性

近年，音声インタフェースを持つ情報機器が実用化されている．産業用途では，工場などでの業務端末や IVR（Interactive Voice Response：自動電話応答システム）・コンタクトセンタにおける業務支援，会議議事録作成支援システムなどが挙げられる．民生品では，カーナビや家電・情報端末の制御インタフェースとして実用化の実績が多く，特に最近では携帯電話における音声認識の実用化が盛んである（図 -1）．

使いやすい音声インタフェースを実現するためには，ユーザの発するさまざまな語彙・言い回しによる発話を可能な限り認識し，コマンドやテキストとして受理できることが望ましい．現在の音声認識においては，音響モデルと単語辞書，および言語モデルを用いるのが一般的である．このうち言語モデルは，辞書に含まれる単語が発話の中でどのように出現するかをモデル化するものである．

システムが用意した単語辞書・言語モデルに適合するもののみが音声認識の対象となるため，多様な表現を受理可能な音声認識として実現するためには，許容範囲の広く，かつ強い言語制約を課すことが可能な言語モデルが必要となる．音声認識方式の種類とそこで使用される言語モデルのタイプ例，さらに代表的なアプリケーションの例を表 -1 に示す．

現在，一般に広く普及していると言える音声インタフェースは，カーナビなどの車載機器操作や IVR での電話自動応答のための音声認識であろう．このような場合の音声認識に用いられる言語モデルは，単語（地名・品名等のキーワード）のみを受理する単一単語（離散単語）モデル，あるいはあらかじめ定められた文法に従った単語列のみを受理する文法（ネットワーク）型モデルであることが多い．

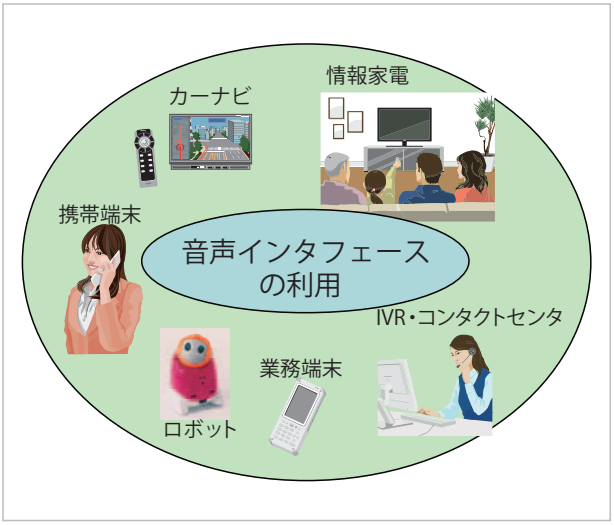

図 -1　音声インタフェースの実用化／利用例

音声認識方式	言語モデル例	アプリケーション例
離散単語	単一単語	カーナビ，IVR
文法型	ネットワーク	家電操作，ロボット
大語彙連続	統計的モデル	字幕／議事録作成支援

表 -1　音声認識方式と言語モデルタイプの例

単一単語モデルは，あらかじめ定められた単語が単独で発声されることを想定した言語モデルである．システム開発者は，単語辞書のみを用意すればよく，比較的低コストで実現が可能である．実用場面においては，ユーザは単語で発声してくださいと言われても，「えー」や「あのー」といった間投詞や，「です」「お願いします」といった文末表現などの余計な単語をつけてしまいがちである．これに対し，間投詞や文末表現をガベージモデルで表現し，これらを受理可能とするワードスポッティングという方法もあるが，ガベージモデルの作り方によっては必要なキーワードまでガベージモデルとして認識されてしまうなど，課題も残されている．

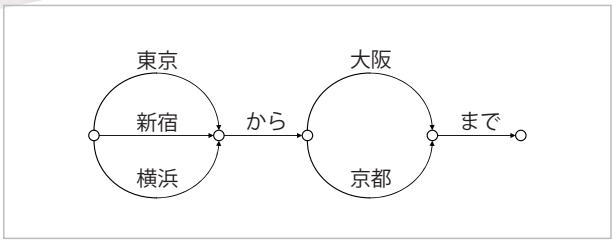

図 -2　ネットワークモデルの例

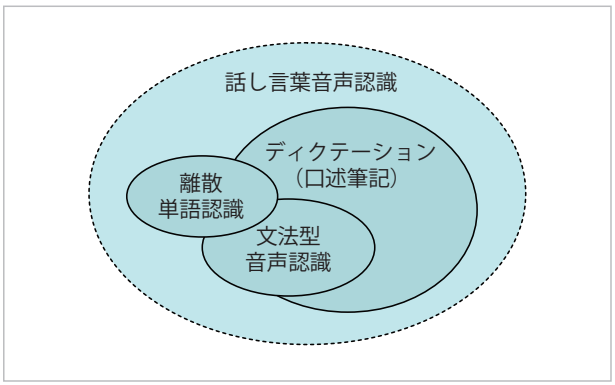

図 -3　音声認識の適用範囲イメージ

文法型モデルでは，図 -2 のようにあらかじめ受理すべきパターンをネットワークあるいは文法という形で言語モデルとして記述しておく．このため，ネットワークで記述された通りの発声に対しては精度が高いというメリットを持つ反面，さまざまなユーザによるさまざまな言い回しを広くカバーするような大規模なネットワークを記述することは現実的でない．そこで，IVR のように特定の言い回しに限定することが可能な用途において実用化がされてきた．ほかにも，家電機器操作やロボットとの対話システムなどの用途では，文法型モデルが用いられている例がある．

一方で，音声のテキスト化／書き起こしといった用途を想定した場合には，文法型モデルでは対応が困難であり，統計的言語モデルを用いた大語彙連続音声認識が用いられることが多い．統計的言語モデルでは，大量のテキストコーパスがあればそれを元に機械学習手法を用いて自動的に言語モデルを構築することができる．2000 年頃にコンシューマ向けに実用化されてきたパソコン上のディクテーションソフトなどは，電子化された新聞記事や WWW のテキストなどの大量の書き言葉コーパスの普及を背景に，このようなコーパスを活用した大語彙連続音声認識を用いていることが多い．ほかにも，音声コンテンツのインデクシングといった用途にも適用可能である．また大語彙連続音声認識は，TV 番組の字幕作成支援や会議における議事録作成支援といった，より非定型な話し言葉的要素を含む分野への応用が進んでいる．しかし，統計的言語モデルを用いる性質上，対象分野（ドメイン）・対象の文体（スタイル）に対する大量の学習データが必要であり，逆に言うと学習した対象でのみ効果を発揮する．たとえば，新聞記事などの大量の書き言葉を用いて学習したディクテーションソフトを会議や講演のような話し言葉の音声認識に適用しようとしても，ドメインやスタイルの違いから満足な精度が得られないといったことが起こる．そこで，特にニーズの高かった日本語の話し言葉においては，大規模な話し言葉コーパスを構築する試みが国立国語研究所・情報通信研究機構・東京工業大学の共同開発により進められた [1]．その成果は CSJ コーパスとして利用可能である [2]．大語彙連続音声認識の高精度化のためには，対象となるドメインやスタイルに特化したモデルを用いるなど言語制約を強めることによる基本性能の向上はもちろん，たとえば間投詞や言い直しといった話し言葉現象へも対応していくことで，受理可能な言語表現を広げることも必要と言える．

このように，音声認識を実用化し広く普及させる上では，高い精度を保ちつつその受理可能な表現を大規模化・多様化することで，適用範囲を広くしていくことが重要である（図 -3 参照）．これによりユーザは，音声入力の際にシステム側が規定する制約を気にすることなく，自由に発声することが可能になる．以下では，この課題を主に言語モデルの側面から議論する．

言語モデルの課題

大語彙でさまざまな言い回しを含む大規模タスクに対して効果的な言語制約を比較的容易に与えられ

る方法として，単語のつながりやすさを大量のテキストコーパスから自動で学習する統計的言語モデルがあることは前章で述べた．中でも，隣接したN個の単語の単語連鎖をモデル化した単語N-gramが，その利便性と精度の両面から現在は主流であると言える．単語N-gramは，Nの値が大きくなると，すなわち4単語や5単語の連鎖を扱おうとすると種類数が膨大になり，有効に学習するためには膨大なテキストコーパスが必要となり現実的でない．そこで，単語（あるいは品詞など）の2連鎖を扱うbigram，3連鎖を扱うtrigramが用いられていることが多い．ところが，bigramやtrigramでは，局所的な制約は与えられるが大局的な情報は持ち得ないという欠点がある．

さらに，これまでの言語モデルは日本語として正しい言葉，たとえば文法や発音の点で正しい日本語を受理することを想定してきた．しかし，実用場面においては必ずしも正しい言葉が話されているわけではなく，これが音声認識をより困難にしていることが分かっている．たとえば，人と人との会話を分析すると，不完全／部分的な文の発声や，フィラーと呼ばれる間投詞が多く挿入されること，また言い直し・言いよどみといった現象が発生することが分かっている．このような困難な話し言葉現象は，統計的に学習するとしても現状利用可能な学習コーパスが不足しているため有効な学習ができず，さらにbigramやtrigramといった低次のN-gramではコンテクストとしてカバーしきれない．このため，話し言葉に対しては十分な音声認識性能を獲得できていないのが実情と言える．

また，統計的言語モデルは学習したテキストコーパスの性質を反映する．すなわち，テキストコーパスのカテゴリ（単語や文が表現する話題や分野などの意味的なラベル）を表すことになる．話題別の複数の言語モデルを構築し，それらを選択・混合することで，話題の変化に応じた精度の高い言語モデルを作る方法も知られている．しかし，これまでの方法はコーパス全体の性質を暗黙的に反映・活用するものであり，より詳細かつ高精度なモデルの構築を可能にするためには，キーワードといったより詳細なレベルでの明示的なモデル化・活用が求められる．

音声認識性能向上のための高次言語知識の利用

●高次言語知識としての大局的な情報

我々は，音声認識の適用範囲を広げるため，言い直しなどの困難な話し言葉現象を含むような発話においても，学習データ不足の問題を低減しつつ特定のドメインへ特化することで認識精度を確保することを主たる目的とし，従来のbigramやtrigramより高次な言語知識を利用した手法を検討している．これにより，ユーザへの制約を緩和し，自然な言い回しを許容した音声情報検索等を可能にすることを目標とする．なお，本研究の一部は経済産業省における「音声認識基盤技術の開発」プロジェクトの支援を受けたものである．

情報検索におけるキーワード認識を対象とし，例としてTV番組を検索する模擬システムを構築し，ユーザによる情報検索のための発話を実際に収集した．収集された発話データをもとに，キーワードがどのように発話されるか，trigram言語モデルを用いた音声認識を行うとキーワードがどのように誤認識されるか，を観察すると，次のような現象が見られた．

- 人名（出演者名など）や番組名といった種類数の多いキーワードについて誤認識が多い．
- キーワード周辺の言い回しの部分は比較的正しく認識できる．
- 同一発話内でキーワードと共起する単語の頻度分布を観察すると，キーワードの種類（人名，番組名，放送局名）によって共起する頻度の大きい単語が異なる．すなわちキーワードの種類ごとに固有な言い回し表現がある．

このような観察結果に基づき，特定のカテゴリに属するキーワードに固有の表現を手がかりとして，話し言葉音声認識においてキーワードの認識精度を向上させる方法を検討した．これは，発話全体の特

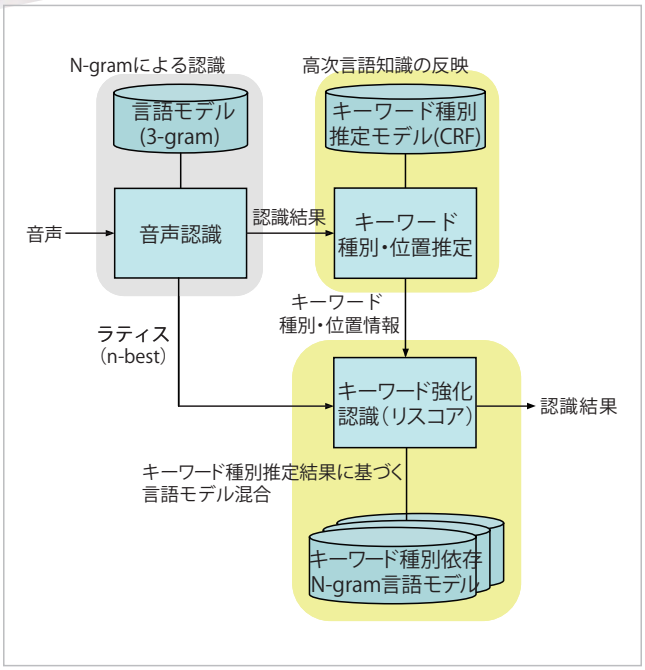

図 -4　全体構成

徴として言い回し表現という情報を参照し，それにより，その発話に含まれるキーワードの種類（カテゴリ）と区間を推定し，その結果を言語的な制約として再度キーワード認識を行う方法である．

今回，その考え方に基づき，CRF（Conditional Random Field：条件付き確率場）[3] によるカテゴリ推定結果を利用した言語モデルの時間依存線形補間方式を開発したので紹介する．

本方式の全体構成を図 -4 に示す．まず，最初に一度音声認識を行う．得られた認識結果に対し，言い回し表現をモデル化したキーワード種別推定モデルを適用し，認識結果の単語単位に，本来当該部分に出現すると考えられるキーワードの種類をラベリングする．これにより，発話中のキーワードの種類・位置を推定する．キーワード種別推定モデルとしては CRF を用い，複数のキーワードの種類のラベル付けやさまざまな言い回し表現を含む大量のテキストデータから学習する．CRF は，複数の特徴量を用いることで，連続したデータに対する大局的なラベル付けを精度良く学習できるという特性を持つ．

その後，推定されたキーワードの種類・位置の情報を用いて再認識を行う．この際，キーワードの種類ごとに用意された言語モデルを，キーワード種類・位置の推定結果に基づいて適切に混合し，再認識に用いることで，キーワード認識精度を向上させる．

各段階での処理結果のイメージを図 -5 に示す．まず従来と同様の方式で音声認識を行う．図 -5 の例を用いると，キーワードは単語の種類数が多く相対頻度が低いため，「稲垣吾郎」が「田舎紀行ろう」に誤るなど誤認識が多く，それと比較して言い回し部分である「の」「出演している」「番組」は誤認識が少ない．得られた認識結果に対しキーワード種別推定モデルを適用し，言い回し表現である「の」「出演している」を手掛かりとして，「の」の前には＜人名＞が来るといった発話中のキーワード種類・位置を推

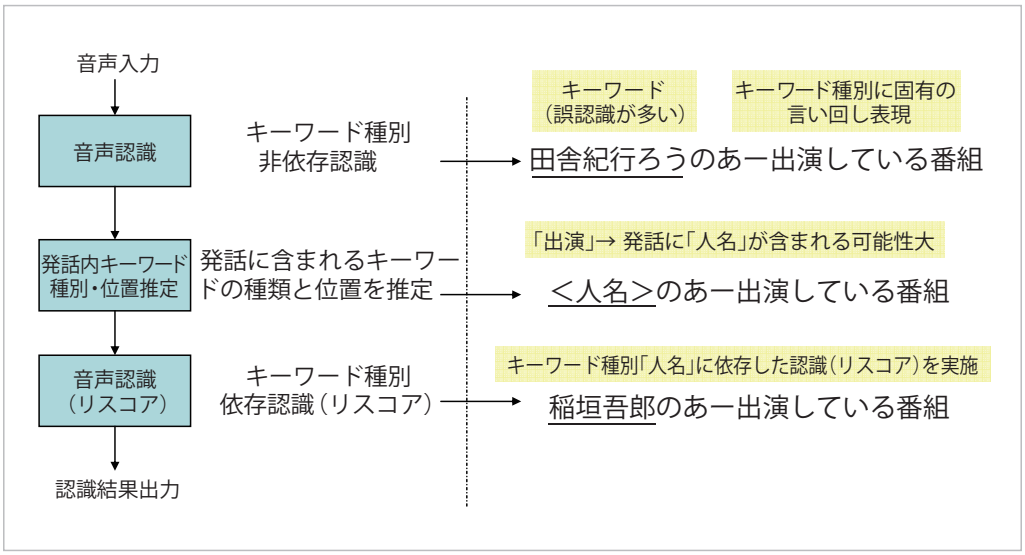

図 -5　処理結果のイメージ

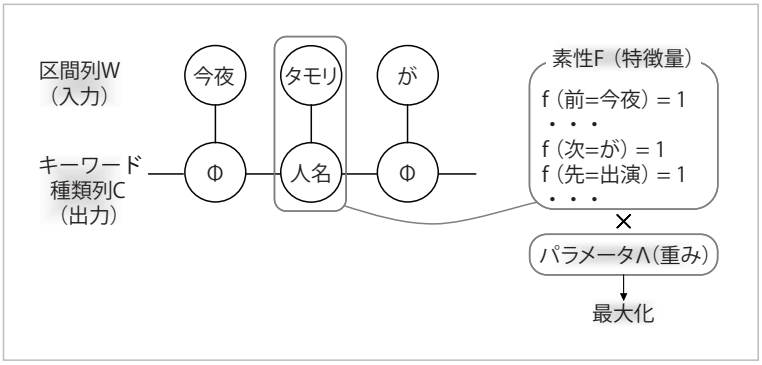

図 -6　ラベリングの例

定する．最後に，発話中のどの辺り（ここでは「の」の前）にどのような種類（ここでは＜人名＞）のキーワードが含まれるかという推定結果に基づいてキーワード認識を強化した言語モデルを動的に生成し，適用することで，最終的な認識結果を出力する．

● CRF による単語種別と位置の推定

《CRF を用いる高次言語処理》

多様な言語情報を組み合わせて扱える枠組みとして，識別モデルの一種である CRF が知られている．CRF は，観測される系列データに対して多種多様の素性（特徴）に基づく識別を行い，最適なラベルを付与する方法であり，自然言語処理分野の諸問題に適用されている．たとえば，単語列に対して固有表現かどうかを示すラベルを付与する問題（固有表現抽出）において，また音声認識分野においては認識仮説の正答・誤答を識別する問題における有効性が報告されている．このように，CRF による単語列処理は，音声認識に複数の言語情報を用いる際に適用する手法として適していると考えられる．

《キーワード種類・位置の推定方法》

前項で述べた通り，キーワード種類・位置を用いる音声認識方法では，対象発話内のキーワードの種類・位置の推定処理と，その推定結果を言語制約とする認識処理を行う．ここでは，その 1 つ目の課題である，発話内のキーワード種類と位置を推定する問題を，CRF による単語列へのラベリング問題として解く方法について述べる．

発話内のキーワード種類と位置を推定するための手がかりとして，前述の通り，キーワード種類に固有の特徴的な言い回し表現に着目する．そこで，提案方法では，対象発話の認識結果として得られる単語列に対して，単語列に含まれる単語ごとに，前後の共起単語を主な素性として，キーワード種類を識別する．単語列内の単語の共起関係を，発話全体に関する大局的な特徴として用いることにより，キーワードと言い回し表現のような必ずしも連接はしないが発話内で頻繁に共起する関係を，モデルの素性に取り込むことができる．このため，従来の 2 単語あるいは 3 単語の連接モデル（bigram，trigram）では困難であった長距離の依存関係も扱うことができる．また，単語列の単語ごとに識別処理を行うことにより，キーワードの種類と位置を同時に推定できる．このため，1 つの発話に複数のキーワード（検索条件）が含まれる場合であっても，それぞれ位置を含めて検出することが可能である．

認識結果の単語列に対してキーワード種類をラベリングする例を図 -6 に示す．たとえば，TV 番組検索のある発話について「今夜タモリさんが 8 チャンで出演する番組」という認識結果の単語列が得られたとする（図 -6 はその一部）．図の「タモリ」という単語に着目するとき，単語列内の共起関係を表す素性（特徴）として，『直前の「今夜」と共起』，『直後の「さん」と共起』，『数単語先の「出演」と共起』などが得られる．これらの素性の出現頻度と，あらかじめ学習した CRF のパラメータを重みとして，出現

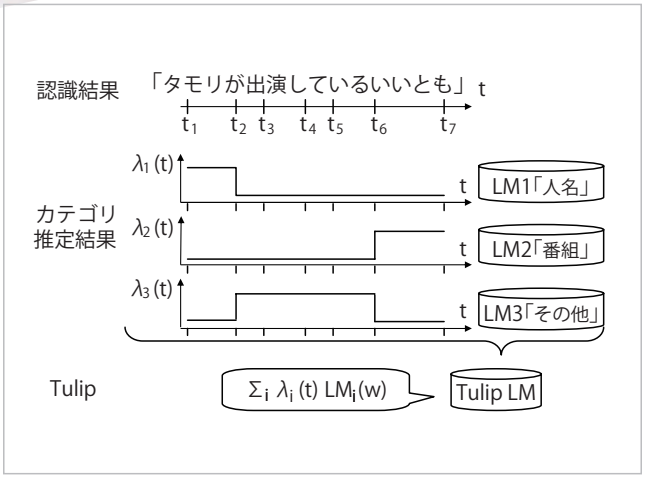

図 -7　Tulip の動作例

確率を最大化するキーワード種類を出力する．この例の場合，「人名」キーワードと「さん」や「出演」との共起関係が大きいことから，認識結果の「タモリ」に相当する部分は「人名」キーワードであると推定できる．さらに，単語連接モデルとは異なり，たとえば「えー，タモリさんが，しゅ，出演している」のように言いよどみやフィラーといった話し言葉現象が存在した場合でも頑健な推定が期待できる．

識別のための素性には，前述の通り，単語列内の共起単語の情報を用いる．これらは，言い回し表現の位置や語順を反映するため，識別対象単語との位置関係で分類し別々の素性とする．これまで述べたように，本来キーワードが発話された部分は誤認識する可能性が高いため，本手法では識別対象自身の表記は素性から除く．すなわち，上記の例では認識結果「タモリ」のキーワード種類を推定するための素性として「タモリ」という認識結果の表記は使用しない．一方で，手がかりとする言い回し表現は認識精度が高いと見込まれるため，これらとの共起関係を用いることで，誤認識に頑健なキーワード検出が期待できる．

このようにして推定されたキーワードの種類・位置を，後段の音声認識において新たな言語制約として使用する．推定結果は，認識結果の各単語について得られたキーワード種類の 1 位の識別結果を用いる．推定結果を用いて，次に述べるようにキーワードの種類・位置に応じた出現確率を用いて言語モデルを重み付けする．

●推定結果に基づく言語モデル混合方式の提案

言語モデルの時間依存線形補間方式（Time Utilized Linear Interpolation：以下，Tulip）は，音声認識において，言語モデルをその発話内の適用位置において動的に変化・適応させる我々独自の方式である．カテゴリとその発話内での時間位置情報を基に，各カテゴリに特化して作られたカテゴリ依存言語モデルを，その重みを変えながら線形補間方式により適用する．発話内の位置に応じて最適なカテゴリを表現するカテゴリ依存言語モデルを用いることで，単一の言語モデルを用いる場合と比較して音声認識の精度を向上させる効果が期待できる．今回，言語モデルとしては統計的言語モデル N-gram を用い，カテゴリごとに用意されたテキストコーパスによってカテゴリ別言語モデルを学習することとする．実装としては，従来から広く用いられている大語彙連続音声認識に組み込むことが可能である．入力音声中のある仮説単語の言語モデルスコアを求める際に，その単語に与える言語モデルスコアの重みを，時刻情報を媒介としてカテゴリ推定結果から求め，各カテゴリ別言語モデルのスコアを線形補間し，当該単語の言語モデルスコアとする．たとえば図 -7 の例では，時刻 t_1 から t_2 まではカテゴリ推定結果に基づき「人名」言語モデル LM1 の重み λ_1 が大きく，単語タモリの言語モデルスコアは LM1 が支配的な状態で計算される．その後 t_2 から t_6 までは「その他」言語モデル LM3 の重み λ_3 が，t_6 から t_7 までは「番組」言語モデル LM2 の重み λ_2 が支配的になっている．このように，時間位置に応じて動的に重みを変更した線形補間が行われることになる．

高次言語知識を用いた音声認識性能の評価

●評価用データの構築

以下では，高次言語知識を用いた音声認識方式の効果について述べる．

図 -8　音声検索試作システムの画面例

	再現率	適合率
レストラン	64.1%（33.4%）	69.4%
TV 番組	55.9%（38.2%）	31.7%

表 -2　カテゴリ種別・位置の推定精度

効果を検証するための評価対象のタスクとしては，情報家電インタフェース開発のために検討を進めている，レストラン検索タスクと TV 番組検索タスクとを用いる．複数のタスクが設定されているのは，開発する手法が特定のタスクに依存したものにならないため，すなわち異なるタスクでも効果があることを示すためである．TV 番組検索タスクは Wizard of Oz 方式（WoZ 方式：システムになりすました人と被験者が対話する方式）によって収録した評価データを，レストラン検索タスクは音声検索が動作する試作システム（図 -8）を用いて被験者に実際に使ってもらいながら収録した評価データを，それぞれ用いる．いずれも，可能な限り実利用場面に近い環境での収録を目指したものである．このため，フィラーや言い直しといった話し言葉現象も，実利用場面と同様に入っていることを確認している．

カテゴリの定義は，レストラン検索タスクでは「地名（駅名含む）」「ジャンル名」「店名」および「その他」の 4 種類，TV 番組検索タスクでは，「人名（出演者名）」「放送局名」「番組名」および「その他」の 4 種類とする．

● CRF を用いた単語種別推定の評価

CRF を用いたカテゴリ種別・位置推定の効果について述べる．

まず評価データに対して大語彙連続音声認識で認識し，次に得られた評価データの音声認識結果に対して，あらかじめ学習した CRF を用いて，単語列に含まれる各単語に対しカテゴリ種別を推定し，推定結果とそのスコア（事後確率）を出力する．推定結果のラベルは，タスクごとに設定した 3 つのカテゴリ種別と，これらのカテゴリ種別ではないことを示す「その他」を加えた 4 種類とする．CRF の学習データには，評価データの話者とは異なる話者の発話で，カテゴリに属するキーワードを含むものを用いる．学習用の音声データの書き起こし単語列にカテゴリ種別情報を付与し，その書き起こしと認識結果の単語列とを DP マッチングにより対応付けて，教師データとする．CRF の学習には CRF＋＋[4] を用いる．

識別のための素性には，単語列内の共起単語などの音声認識結果から得られる情報を使用する．識別対象単語の周辺の情報としては，発話内の前後 7 単語までの共起単語の表層と単語事後確率の組を用いる．さらに，これらを識別対象単語との出現位置の前後関係と単語間の距離に区別する．また，識別対象単語自身の情報として，単語事後確率，音節数，先行無音の有無の 3 種類を用いる．これらのうち，単語事後確率と音節数はその単語の正誤と関連のある特徴である．また，先行無音の有無は，発話中でのキーワードの出現しやすさと関連のある特徴である．

評価尺度にはカテゴリ種別それぞれに対する識別結果の再現率（recall）と適合率（precision）を用いる．再現率は，発話されたキーワードの総数に対する正しく識別したものの割合を示す．また適合率は，識別結果のうち正しかったものの割合を示す．

上記の条件で行った評価実験の結果を表 -2 に示す．再現率のカッコ内は，前段の音声認識において誤認識されたキーワードの再現率である．この結果から，提案するカテゴリ種別・位置の推定方法によ

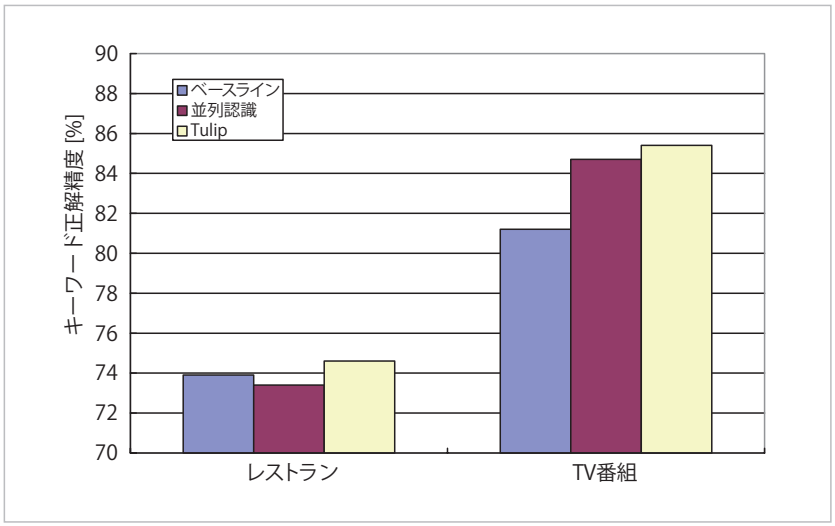

図 -9　評価結果

り，話し言葉現象によって仮にキーワードが誤認識された場合であっても，そのキーワードが属するカテゴリ種別と位置を検出可能であることが確認された．前段の音声認識でキーワードが誤認識された場合でも，その 3 割以上でカテゴリ種別が正しく検出されたことから，前後の言い回し表現が識別の素性として有効に働いたと言える．

●単語種別・位置推定を用いた音声認識の評価

音声認識の評価は，レストラン検索タスク，TV 番組検索タスクとも，収録した音声データを用いてシミュレーションによる認識実験を行う．大語彙連続音声認識を使用する．評価はすべての認識結果単語ではなく，キーワードの認識率によって評価する．

認識実験を行って効果を検証した結果を述べる．いずれのタスクにおいても，ベースラインと並列認識とを比較対象とする．ベースラインとは，カテゴリ非依存の言語モデルを用いた場合，すなわち「その他」カテゴリの言語モデルを用いた場合である．並列認識とは，各カテゴリ別言語モデルを用いた音声認識をそれぞれ並列に動作させ，最尤の認識結果を選択した場合であり，カテゴリ別言語モデルを使用する場合の従来法の 1 つと考えることができる．並列認識では 1 発話全体に 1 つの(最適な)カテゴリ別言語モデルを適用するのに対し，Tulip では 1 発話の中で最適なカテゴリ別言語モデルを切り替えながら適用するため，より緻密な適用が可能になっている．

評価結果を図 -9 に示す．レストラン検索タスクでは，ベースラインおよび並列認識と比較して，提案法である Tulip が良い精度を得られていることが分かる．このとき，カテゴリ推定精度は表 -2 の通りすべてのカテゴリ平均で再現率 64.1%・適合率 69.4% であった．カテゴリ推定精度は必ずしも高いわけではないが，そのカテゴリ推定結果を用いることで音声認識には精度向上の効果があったと言える．

また，レストラン検索タスクでは「＜地名＞にある＜店名＞」のように 1 発話中に複数のカテゴリの単語が混在することが多く，単一のカテゴリ依存言語モデルでは悪影響が大きかったと考えられる．このために，ベースラインよりも並列認識の方がやや精度が低くなっている．この点でも，1 発話中に言語モデルを切り替えて適用することが可能な Tulip の有効性が言える．

TV 番組検索タスクにおいても，ベースラインと比較した場合にはもちろん，並列認識と比較しても提案法である Tulip が良い精度を得られていることが分かる．このことから，提案手法である Tulip はタスクに依存せず効果があると言える．

高次言語知識による効果の考察

前章で説明した評価結果から，カテゴリ推定結果を利用したTulip方式において次のことが言える．まず，カテゴリ非依存の言語モデルを用いた場合と比較して，特定カテゴリに特化したモデルを切り替えて利用することで，キーワード正解率の向上が得られる．次に，実利用場面に近い環境で収録したフィラーや言い直しなどの話し言葉現象を含む自然なデータに対して，時間情報に基づいた非連接な長距離の依存性を考慮することで精度向上の効果がある．

すなわち，認識が困難な話し言葉現象を含むような発話においても，特定のドメインに特化することでキーワードの認識精度を確保し，音声認識の適用範囲を拡大する可能性を示した．

音声認識における言語モデルの今後の課題

音声認識の適用範囲を広げるための課題とその解決の試みについて，主に言語モデルの側面から議論した．音声認識の適用範囲を広げるためには，統計的言語モデルの利用において学習データ不足の問題を解決しつつ，話し言葉現象への対応が必要である．そのための1つの試みとして，言い直しなどの困難な話し言葉現象を含むような発話においても，従来のbigramやtrigramより高次な言語知識を利用することで認識精度を確保する手法として，CRFによるカテゴリ推定結果を利用した言語モデルの時間依存線形補間方式とその効果について解説した．しかし，今回のカテゴリ推定精度にはなお改善の余地があると考えられる．統計的モデルを用いる以上，素性や学習データの不足は常に課題となる．また，方式の性質上，キーワードのみを連続する発話のような，言い回し表現を含まない場合には効果が小さい．このような場合には，複数の発話から文脈を推定するなどさらに広範囲の情報を用いることが必要である．さらには，今回は議論しなかったが，発音の変形や発音の変形など話し言葉現象に見られる音響的な課題も残されている．

音声認識の実用化を推進するためには，その適用範囲を広げるためにさまざまな場面において精度良く効率的に言語制約をかける仕組みが必要である．今回解説した手法はその1つの取り組みではあるが，広く話し言葉現象を考えるとまだまだ部分的な対応にとどまっているのが現状である．今後，話し言葉でより顕著になる発声変形など，さらなる調査・分析と課題抽出，そして課題への対処のサイクルを回していく必要がある．

参考文献

1) 古井貞熙，前川喜久雄，井佐原均：科学技術振興調整費開放的融合研究制度：大規模コーパスに基づく『話し言葉工学』の構築，日本音響学会誌，56(11), pp.752-755 (2000).
2) http://www.kokken.go.jp/katsudo/seika/corpus/
3) Lafferty, J., et al. : Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data, Proc. of ICML, pp.288-298 (2001).
4) http://crfpp.sourceforge.net/

(平成22年8月30日受付)

花沢　健（正会員）k-hanazawa@cq.jp.nec.com
1997年日本電気（株）入社．音声認識，音声翻訳の研究開発に従事．現在，情報・メディアプロセッシング研究所主任研究員．日本音響学会，人工知能学会各会員．