9 音声認識の多言語化技術

河村 聡典
（株）東芝 研究開発センター 知識メディアラボラトリー

音声認識技術を応用した製品のグローバルな展開のために必要な，音声認識の多言語化技術について説明する．声調言語への対応，多言語対応開発コスト削減を目指し，①声調認識方式，②音響モデルの言語間適応化方式について新方式の開発を行った．評価実験の結果，いずれも従来方式よりも高い効果を確認した．現在，実際の多言語音声認識エンジン開発への適用を進めているところである．

音声認識の多言語化の必要性

近年，カーナビゲーションシステムの音声コントロール，コールセンターの電話応答システムなど，音声認識技術を応用した製品が世の中で使われるようになってきている．このような製品をグローバル市場で広く普及させることを考えた場合，多言語音声認識に対するニーズはきわめて高く，音声認識の多言語化技術の確立が必須である．

図 -1 は，音声認識エンジン全体の一般的な構成図を示したものである．

- 入力した音声波形データから特徴量を抽出する特徴抽出部
- 「a」という発音はこういう特徴，「i」という発音はこういう特徴を持っている，というような音声の音響的な特徴を記した音響モデル
- 認識すべき単語群とその発音，およびその単語がどのように繋がって発声されやすいかという言語的情報を記した言語モデル
- 抽出された特徴と音響モデル，言語モデルを照合して認識結果を出力する照合部

から構成される．音響モデル，言語モデルは言語に依存する部分であり，多言語対応のためには，各言語ごとのコーパス（データ）から作成して使用する．

音響モデルは，実際に発声された多数の音声コーパスから統計的手法により作成するが，一般に，音声データ量が多いほど高い認識性能が期待できる．しかし，大量の音声コーパスを収録するためには膨大な開発リソースが必要であり，音声認識多言語展開を実現する上でのボトルネックとなっている．

また，多言語展開のためには特徴抽出部も考慮する必要がある．これまでの特徴抽出部は，日本語・欧米言語を中心に検討されてきており，異なる言語特徴，たとえば声調を有する中国語，タイ語などのアジア言語では十分な性能が得られておらず，特徴抽出部を声調言語認識に適したものにする必要がある．

そこで，我々は，上記の課題を解決するために，①声調認識方式の開発と，②音響モデルの言語間適応化方式の開発を行った．以下，それぞれについて詳しく説明する．

声調認識方式の開発

●声調

中国語（北京語・広東語）やタイ語は声調言語と呼ばれ，音節が特定の音高変化のパターンを伴って発音される点で，日本語や欧米言語とは異なる特徴を有している．この音高変化のパターンを声調と呼ぶ．

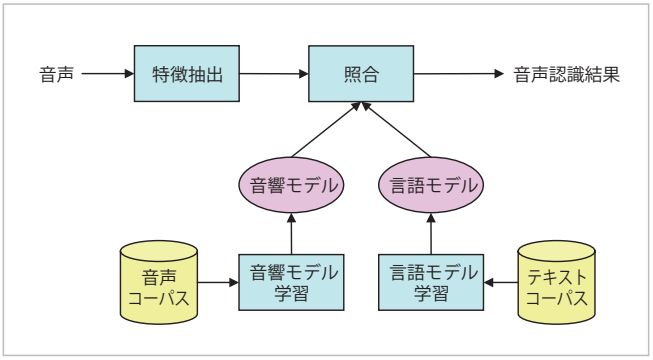

図 -1　音声認識エンジンの構成

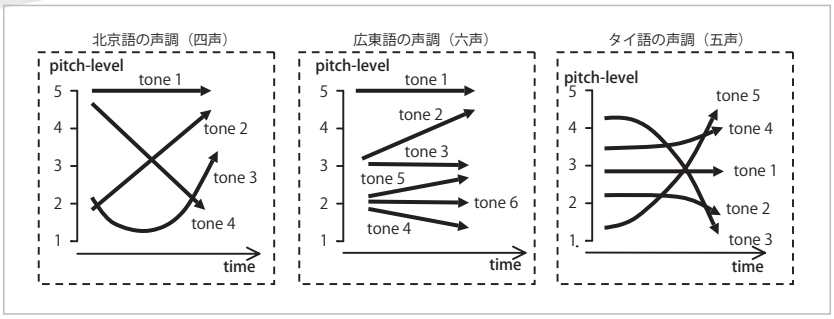

図 -2　北京語・広東語・タイ語の声調

声調言語では，同じ音韻を持つ音節であっても，異なる声調を伴う音節は異なる音節であり，それぞれ異なる意味を持つ．

図 -2 に，北京語，広東語，タイ語の声調の種類を示す．横軸が時間，縦軸がピッチ（音の高さ）を意味する．北京語では音節単位で声調が定義されており，4 種類の声調パターン（四声）を持つ．第一声は高く平ら，第二声は上がり調子，第三声は低く抑える，第四声は急激に下がるパターンをそれぞれ表している．北京語ではこれらの声調の違いにより，たとえば“ma”という音節に対して媽（第一声：お母さん），麻（第二声：麻），馬（第三声：馬），罵（第四声：罵る）と意味が区別される．このように，1 音節内でのピッチの変動によって区別される声調は，曲線声調と呼ばれる．一方，各音節が持つ相対的な音の高低の違いを区別する声調は，段位声調と呼ばれる．たとえば，タイ語では図 -2 に示すように 5 種類の声調が存在し，そのうち第一声・第二声・第四声の 3 種が段位声調である．

●声調認識の必要性

声調言語には同音韻・異声調の語彙が多数存在する．そのような語彙の典型例としては中国語の人名が挙げられる．参考のため，文献 1）で報告された中国語の人名バリエーションの調査結果を表 -1 に示す．

表 -1 は，中国語の人名サンプルについて，表記文字あるいは発音に着目した場合のユニークなパターン数を姓名・姓・名ごとに数え上げた結果である．表の CDNs（Character-Distinctive-Names）は表記文字に着目した場合のパターン数を，SDNs（Sound-Distinctive-Names）は北京語発音における音韻と声調の組合せに着目した場合のパターン数を，BSNs（Base-Sound-Names）は音韻のみに着目した場合のパターン数を示している．

	サンプル人数	CDNs	SDNs	BSNs
姓名（2 文字）	137,137	32,212	21,374	15,038
姓名（3 文字）	1,497,923	1,203,471	989,307	691,409
姓（1 文字）	1,635,060	2,626	921	290
名（1 文字）	131,137	3,000	963	309
名（2 文字）	1,447,782	202,197	77,201	19,671

表 -1　中国語の人名バリエーション

たとえば，表の 1 行目は中国語の姓名のうち表記文字数が 2 文字のものの調査結果であり，137,137 サンプルのうち，表記文字に着目した場合には 32,212 個のパターンが，音韻と声調に着目した場合には 21,374 個のパターンが，音韻のみに着目した場合には 15,038 個のパターンが存在することを示している．このとき，SDNs と BSNs の差は，相互に同音韻・異声調の関係にある人名ペアの数に相当する．

表 -1 の SDNs と BSNs を比較すると，いずれの項目においても SDNs が BSNs を大幅に上回っており，中国語の人名には非常に多くの同音韻・異声調の語彙が存在することが分かる．また，人名だけではなく，地名などの固有名詞にも同音韻・異声調の語彙は多く存在する．

このように，声調言語には同音韻・異声調の語彙が多数存在し，そのような語彙を相互に識別するためには声調の違いを認識する技術が必要となる．

●従来の声調認識とその問題点

日本語や欧米言語の音声認識システムでは，音韻を識別するための音声特徴量としてメルケプストラム係数（MFCC：Mel-Frequency Cepstrum Coefficient）と呼ばれる特徴が広く用いられている．しかし，MFCC は音高などの韻律情報を含まない特徴量であるため声調認識にはほとんど効果がない．声調認識のためには韻律情報を含む別の特徴量（トーン特徴量）を追加する必要がある．そのようなトーン特徴量としては，音声の音の高さを表す情報である基本周波数（F0）を用いることが一般的である．図 -3 に，従来のトーン特徴抽出方式を示す．まず F0 を推定し，その時間変化量（時間微分）を計算し，併せてトーン特徴量とする．

F0 の推定方式としては自己相関関数を用いる方法がよく用いられる．音声波形は，有声部（いわゆる声帯が震える音声）において周期的によく似た波形が繰り返されている．この波形の周期を基本周期と呼び，その逆数が基本周波数（F0）である．音声波形を時間方向にずらしながら，それ自身との重なり具合（自己相関関数）を計算すると，等間隔で相関値の山谷が繰り返す構造（調波構造）が形成される．重なり具合が一番強い（自己相関関数のピークが一番大きくなる）ずらし時間量を基本周期として抽出し，その逆数として基本周波数 F0 を求める（図 -4）．

しかし，この F0 推定手法には，背景雑音があるうるさい場所では推定精度が悪い，という問題がある．調波構造自体は雑音環境下でも比較的安定に得られるが，最大ピーク位置は雑音の影響を受けやすく F0 の誤推定につながり，雑音環境下における声調認識性能の劣化の一因となっている．音声認識を応用した製品の実用化では，実環境雑音に対して頑健に動作する音声認識システムが要求される．雑音環境下における声調認識の耐雑音性の改善は，声調言語を対象とした音声認識システムにとって重要な課題である．

●耐雑音トーン特徴抽出方式 SELF+ASC[3)]

上記問題を解決するために，我々は，耐雑音性の高いトーン特徴抽出方式として，SELF＋ASC（Shift Estimation of Log-Frequency domain ＋ Accumulation of Shifted Coefficients）（図 -5）を新たに開発した．SELF とは，Δ logF0 を logF0 の微分として求めるのではなく，直接，雑音環境下でも精度よく推定する手法である．ASC とは SELF による推定された Δ logF0 情報を用いて，logF0 の推定精度をさらに向上させる手法である．以下，手法を概説する．

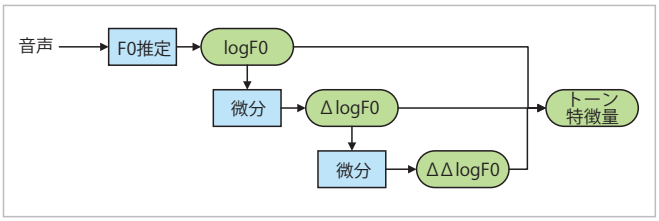

図 -3　トーン特徴抽出ベースライン方式

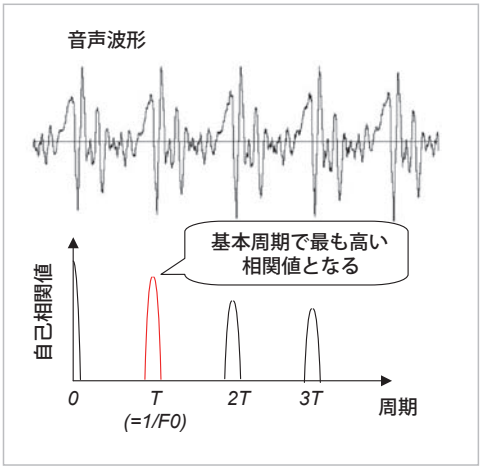

図 -4　自己相関関数による F0 推定

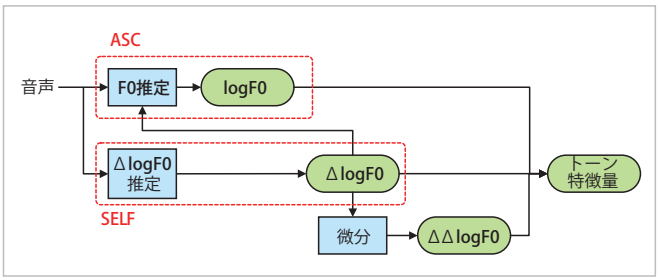

図 -5　SELF+ASC

《SELF》

雑音環境下であっても，自己相関関数の調波構造自体は比較的安定して得られることに着目する．隣り合う時刻フレームの自己相関関数を周期軸を対数化して比較すると，調波構造が平行移動した関係と

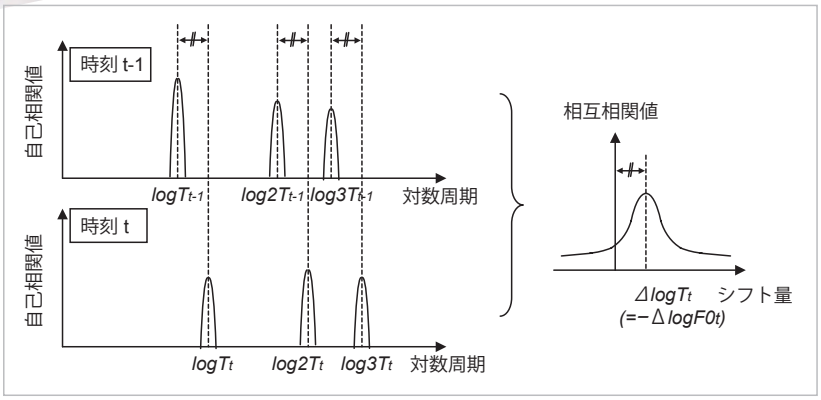

図 -6　SELF によるΔ logF0 の推定

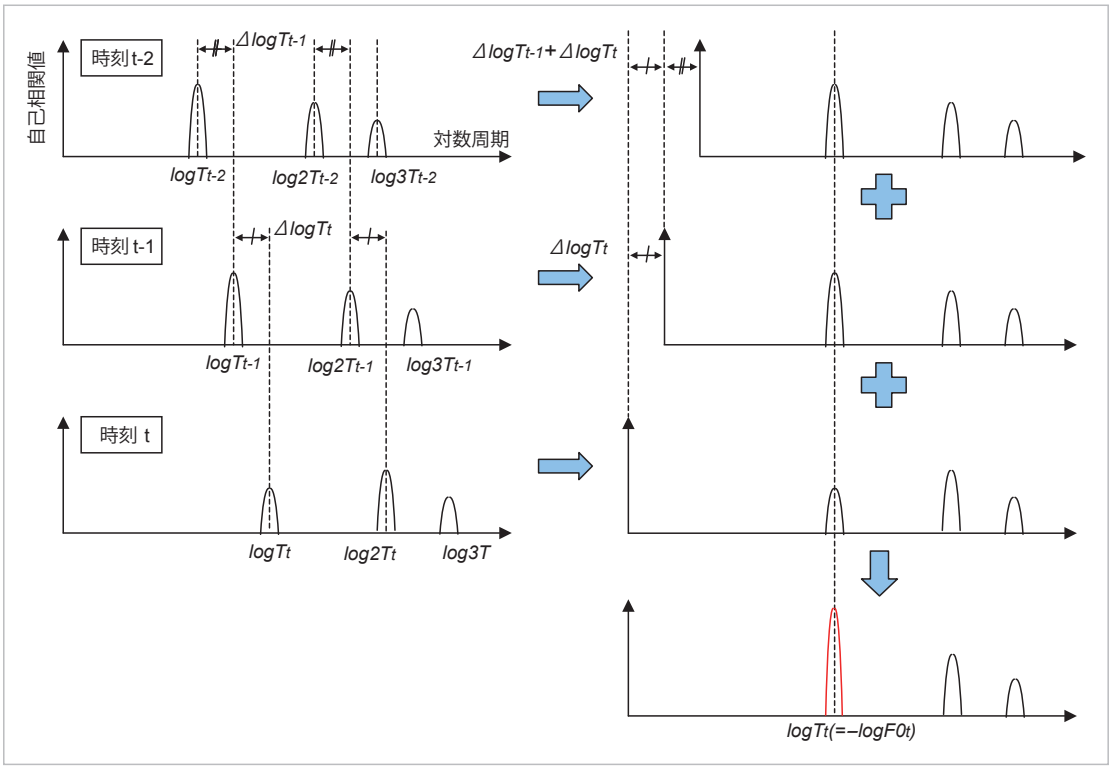

図 -7　ASC による logF0 の推定

なるため，お互いの相互相関関数を計算すれば単峰性のピークが安定して得られる(図 -6)．このピーク位置のシフト量は隣り合うフレームの対数基本周期の変化量Δ logT を表しており，これから対数基本周波数の変化量Δ logF0 を求めることができる．このように，SELF は調波構造に着目することにより，雑音環境下でも安定してΔ logF0 を抽出する手法である．

《ASC》

ASC は，SELF で得られたΔ logT（＝−Δ logF0)を用いて，過去フレームの自己相関関数をシフトして足し込むことにより，雑音環境下でも，自己相関関数の最大ピーク位置を安定して抽出する手法である(図 -7)．当該時刻フレームの自己相関関数だけを用いる従来手法と比べて，logF0 の推定精度が向上する．

●耐雑音トーン特徴を用いた声調認識の評価

SELF ＋ ASC で抽出されたトーン特徴量と MFCC を用いた声調認識実験を，北京語を対象に行った．比較のためベースライン方式の F0 抽出方式としては，予備実験の結果，既存方式の中で最も性能の良かった手法[2)]を用いた．

声調認識性能そのものに着目するため，発声内容の音韻は既知，声調は未知という条件における音節単位の声調認識を評価タスクとした．評価には，背

	0dB	5dB	10dB	15dB	20dB	25dB
Baseline	55.9%	31.7%	20.2%	18.0%	18.6%	19.5%
SELF+ASC	28.2%	20.0%	17.5%	17.5%	17.3%	17.0%
改善率	49.6%	36.9%	13.4%	2.8%	7.0%	12.8%

表 -2　北京語の声調認識誤り率の比較

	5dB	10dB	15dB	20dB	clean	平均	改善率
(1) Baseline	33.92%	23.42%	20.67%	20.19%	20.58%	23.76%	-
(1)+ 従来法	35.06%	19.50%	15.06%	13.97%	14.25%	19.57%	17.63%
(1)+ 提案手法	22.64%	14.08%	11.53%	11.53%	10.56%	14.47%	39.10%

表 -3　北京語の連続数字発声認識誤り率の比較

景雑音のない静かな環境で収録した音声に，高速道路走行雑音を人工的に重畳したコーパスを用いた．重畳時の SNR は，実車内で収録した音声コーパスから求めた SNR 分布の範囲に合わせて（0，5，10，15，20，25）dB の 6 種に設定した．

実験結果を表 -2 に示す．Baseline はベースライン方式による声調認識誤り率，SELF ＋ ASC は今回開発手法による声調認識誤り率，改善率はベースライン方式に対する誤り改善率を示している．

ベースライン特徴量と提案特徴量（SELF＋ASC）を比較すると，すべての SNR 条件において，提案特徴量（SELF＋ASC）がベースライン特徴量に対して認識誤りの改善を示していることが分かる．特に，低 SNR 条件で改善率が大きく，提案特徴量により声調認識の耐雑音性が改善されることが確認された．

●提案方式の更なる可能性

トーン特徴を用いることで，声調以外にも音韻そのものの識別性能も向上できる可能性がある．そこで，提案した特徴量を用いて，北京語の連続数字発声（4 桁～ 8 桁）の認識実験を行った [4]．結果を表 -3 に示す．表中の数字は誤認識率を示している．Baseline はトーン特徴を用いず MFCC のみを用いた場合，（1）＋ 従来手法は MFCC に従来のトーン特徴を追加した場合，（1）＋ 提案手法は MFCC に今回開発したトーン特徴を追加した場合を意味する．

表 -3 より，トーン特徴を用いることで数字認識性能が改善されること，さらに，今回開発した手法を用いることで認識性能が大幅に向上することが確認された．これは，提案したトーン特徴量の導入により，声調の違いではないものの，よく似た間違えやすい音の識別能力が向上したためと考えられる．このように，提案した特徴量は，声調認識だけではなく，数字認識のような声調の違いが直接的に認識性能に影響をしないようなタスクの性能改善にも効果がある．

音響モデルの言語間適応化方式の開発

●音声コーパス収集コストの問題

音声認識エンジンでは，ある言語の音響モデルとして，統計的手法に基づき当該言語の大量の音声コーパスで学習した音響モデルが用いられる．音響モデルは，ある言語における音声のさまざまな変動，たとえば，前後の音韻環境や話者性の違いによる音声の多様性を十分にカバーするために，大量の音声コーパスで学習されることが望ましい．統計的手法に基づく音響モデル学習では，学習用の音声コーパスの規模が大きくなるほど，学習の信頼度・安定性が向上し，精度の高い音響モデルを得ることができる．つまり，音声認識の性能は，音響モデルを学習するための音声コーパス量に依存する．

しかし，新規言語の大規模な音声コーパスを収集するには，膨大なコストが必要となる．音響モデル学習用の音声コーパスは，通常，数百人から数千人規模の話者がそれぞれ数百～数千程度の語彙を発話した音声から構成されており，その収集にかかるコストは非常に大きい．このコストが，音声認識技術

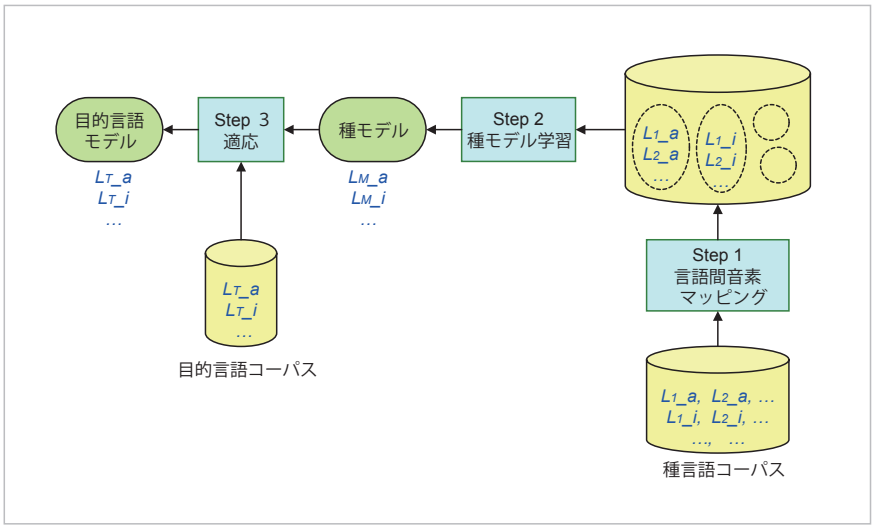

図 -8　音響モデルの言語間適応（従来方式）

を応用した製品をグローバル展開する際のボトルネックとなっている．

この問題を解決する手段として，比較的小規模な音声コーパスしか利用できない場合でも，目的言語の音響モデルを高い精度で学習する言語間適応（CLA：Cross-Language Adaptation）[5] と呼ばれる手法がある．

●従来の言語間適応化方式

言語間適応は，ある言語の音響モデルを学習する際に，他の言語群の既存の音声コーパスを用いて学習した音響モデルを「種」モデルとして，少量の目的言語音声コーパスを用いた適応を実施することで，目的言語の音響モデルを得る手法である．目的言語とは異なる他の言語群の大規模音声コーパスを利用することで，小規模音声コーパスしか利用できない言語に対しても，当該言語の音響モデルを高精度に学習できることが期待される．音響モデルの言語間適応の一般的プロセスは，図 -8 に示すように 3 つのステップからなる．

《言語間音素マッピングの作成》

言語間適応では，まず，目的言語と種言語との間で，相互に類似する音素の対応関係（マッピング）を作成する．この対応関係を「言語間音素マッピング」と呼ぶ．代表的なマッピング作成手法として，目的言語・種言語の音素を IPA（International Phonetic Alphabet；国際音声記号）に代表される汎言語の音素シンボルで表記し，音素シンボルが一致または類似する音素群を対応付けるという手法がある．たとえば，IPA 表記で「a」と表される母音は，言語 L1 の音素 L1_a，言語 L2 の音素 L2_a，…などすべての言語において共通の同じ母音 a として扱う，という考え方である．

《種言語コーパスを用いた種モデルの学習》

次に，目的言語のある音素のモデルに対して，当該音素に対応付けられた種言語の音素の学習用音声コーパス（種言語コーパス）を用いて種モデルを学習する．ここで種モデルとは，後述する適応の「種」となるパラメータを与える音響モデルであり，目的言語の音響モデルの近似となることが期待される．

《目的言語モデルへの適応》

最後に，種モデルのパラメータを初期値として，目的言語音声コーパスを用いた言語間の適応を行い，適応後のモデルを目的言語モデルとして取得する．目的言語のある音素のモデルに対して，種モデルのパラメタを適応の初期値として利用することで，少量の目的言語コーパスを用いた適応であっても，精度の高い目的言語モデルを取得できる．

種モデルから目的言語モデルへの適応手法としては，ブートストラップ法や，話者適応に用いられる MAP 推定や MLLR などの適応手法の利用，および，それらを組み合わせた手法の利用が提案されている．

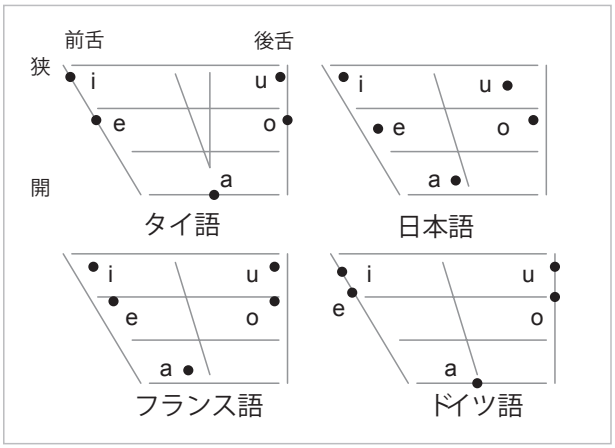

図 -9　母音四辺形における母音の位置

●従来の言語間適応化方式の問題点

IPA を介した言語間音素マッピングを利用する場合，目的言語と種言語のある音素が同一の IPA シンボルに対応する場合でも，それらの音素が同一の音響的特徴を有するとは限らない．

図 -9 は，タイ語・日本語・フランス語・ドイツ語の 5 つの母音に関して，母音四辺形中の位置を図示したものである．母音四辺形とは，音声学において，母音発声時の口の開き具合，舌先の位置によって母音を表現する方法であり，発声器官による音の違いが表現されている．ここで例示するように，同一の IPA シンボルに対応する母音であっても，言語ごとに母音四辺形中の位置は異なる．すなわち，同一の IPA シンボルに対応付けられる音素であっても，言語間で音響的特徴の差異が存在する．

したがって，単純な IPA によるマッピングから作成した種モデルは言語による差異を含んだものとなり，その近似精度は高くなく，結果として言語間適応により作成された目的言語の音響モデルを用いた認識精度に悪影響を及ぼすという問題がある．

● CLA-AT

そこで，我々は，種モデルの学習に適応学習（Adaptive Training）の概念を導入することで，目的言語に対する種言語の音響的特徴の差異を補正し，より近似精度の高い種モデルを学習する手法 CLA-AT：Cross-Language Adaptation with Adaptive Training を開発した．

CLA-AT による種モデル学習プロセスを図 -10 に示す．CLA-AT では，少量の目的言語コーパスから作成した初期モデルに対して，種言語コーパス特徴量の線形変換パラメータを最尤基準で推定し，線形変換後の種言語コーパス特徴量で種モデルを学習するという適応学習の仕組みを導入した．

CLA-AT の概念を図 -11 に示す．図 -11 は，タイ語を目的言語に，日本語・フランス語・ドイツ語を種言語とした場合の，提案手法に期待される効果の模式図である．図中の矢印は，適応学習時の音響

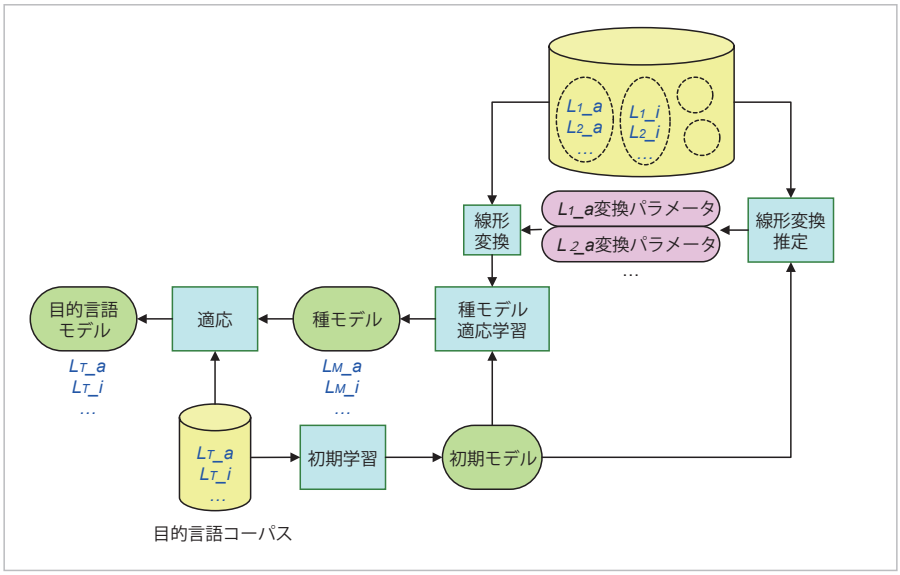

図 -10　CLA-AT による言語間適応

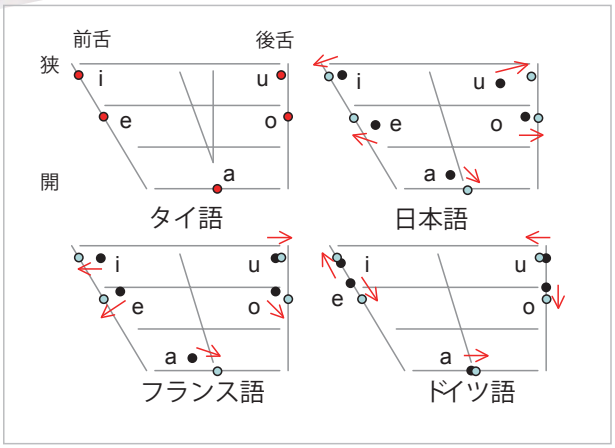

図 -11　言語間の音響的特徴の差異の補正

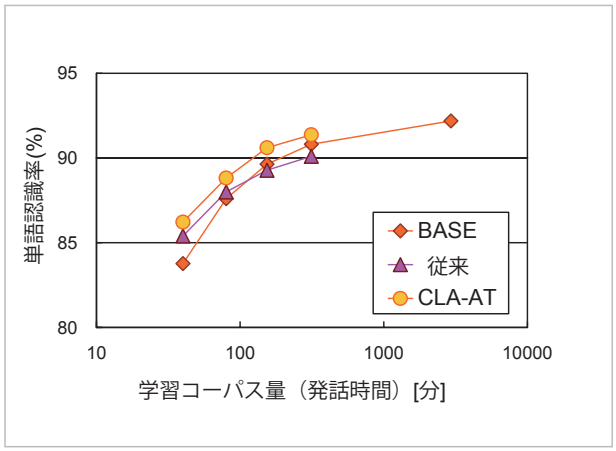

図 -12　従来手法と CLA-AT の比較（広東語）

的特徴の補正を例示している．

提案手法ではこのように，種言語の音響的特徴を目的言語の音響的特徴と一致させるよう補正を行ってから種モデルを学習することにより，目的言語モデルに対する種モデルの近似精度を向上させる．この結果，少量の目的言語コーパスで高精度な目的言語の音響モデル学習が可能になる．

● CLA-AT の評価

音響モデルの学習に用いる目的言語音声コーパス量（発声時間）と認識率の関係を調べることで，提案手法の効果を確認する実験を行った．

目的言語はタイ語，広東語とした．種言語は，アメリカ英語，イギリス英語，ドイツ語，フランス語，スペイン語，イタリア語，オランダ語，北京語と，目的言語のうち学習の対象ではない言語を用いた．

種言語の音声コーパスは，各言語につき約 200 名の話者の発声した 300 ～ 1000 分間程度の発話コーパスからなる．目的言語の音声コーパスのサイズは，一人当たりの発話時間の制限ではなく，話者数を制限することで音声コーパス量を調整した．

評価に用いる音声データおよび音声認識タスクは，車載応用を想定して，走行中の車内で発声された 50 人名発話に対する 100 人名認識とした．

図 -12 に，広東語についての従来手法との比較実験結果を示す．BASE とは，言語間適応なしの場合，すなわち目的言語コーパスのみで学習した音響モデルを用いた場合である．CLA-AT は提案手法である．音響モデルの音素単位はモノフォンと呼ぶシンプルなものを用いた．横軸は学習に用いる目的言語のコーパス量（対数軸），縦軸は認識率を表しており，左上に位置するほど，少ない学習コーパス量で高い認識性能が得られる優れた方式であることを意味する．

図 -12 より，従来手法では言語適応の処置を施しても，学習コーパス量に対する認識性能において明確な改善は見られなかったが，提案手法ではより少ないコーパス量で同等の認識性能が得られることが分かる．

引き続き，音響モデルの音素単位をトライフォンと呼ぶ緻密なものに変更した場合の実験結果を図 -13（広東語），図 -14（タイ語）に示す．図中の「半減」とは，BASE 手法に対して，性能をそのままでコーパス量を半減した曲線である．この図より，提案手法は言語適応なしの場合と比較して，非常にコーパス量が少ない場合を除き，ほぼコーパス量半減を達成できていることが分かる．

音声認識の多言語化のまとめ

音声認識の多言語化の必要性を述べ，そのための 2 つの課題を挙げ，その課題を解決するために開発した，①声調認識方式と，②音響モデルの言語間適応化方式について紹介した．いずれの方式も，これ

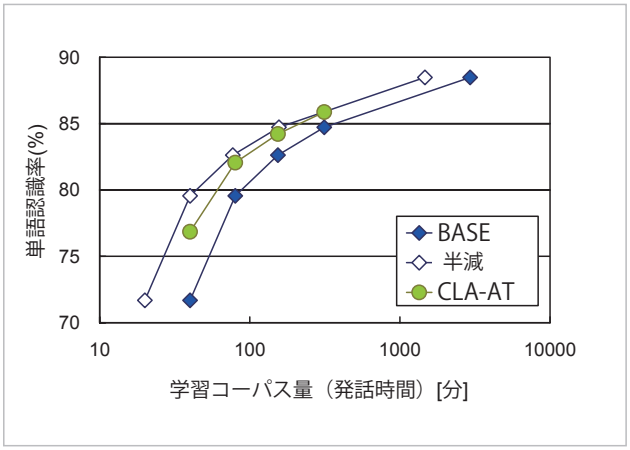

図 -13　CLA-AT の効果（広東語）

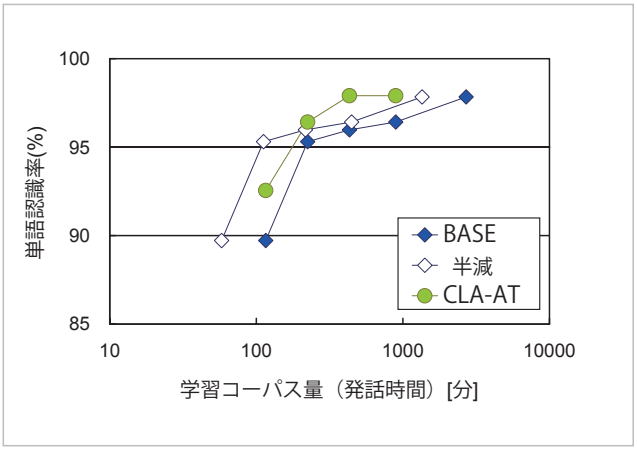

図 -14　CLA-AT の効果（タイ語）

までの実験により基本的な効果は確認できており，現在，実際の多言語音声認識エンジンの開発への適用を進めているところである．

特に，企業の立場から音声認識をビジネス視点で捉えた場合，多言語対応に要する開発コストの削減は重要な課題である．その意味では，今回は，図 -1 の音響モデルの開発コスト焦点をあてた言語間適応化方式の取り組みについて紹介したが，音響モデル以外にも言語モデルの開発もコスト高の要因となっている．今後は，言語モデル開発コスト削減のための技術，たとえば，良質のテキストコーパスを安価にかつ大量に収集・作成するために必要な技術や，言語モデルの適応化技術なども，ますます重要になるであろう．

参考文献

1) Zhang, Y., Medievski, A., Lawrence, J. and Song, J. : A Study on Tone Statistics in Chinese Names, Speech Communication, Vol.36, pp.267-275 (2002).
2) Ghulam, M. et al. : A Noise-robust Feature Extraction Method based on Pitch Synchronous ZCPA for ASR, Proceedings of INTERSPEECH 2004, Jeju Island, Korea (2004).
3) Kida, Y., et al. : Robust F0 Estimation based on Log-time Scale Autocorrelation and its Application to Mandarin Tone Recognition, Proceedings of INTERSPEECH 2009, Brighton, U.K. (2009).
4) Zhao, R., et al. : Using Duration and Pitch for Mandarin Digit String Recognition, Proceedings of ICASSP 2010, Dallas, U.S.A. (2010).
5) Schultz, T. and Waibel, A. : Language-independent and Language-adaptive Acoustic Modeling for Speech Recognition, Speech Communication Vol.35 (2001).

（平成 22 年 9 月 6 日受付）

河村　聡典（正会員）akinori.kawamura@toshiba.co.jp
昭和 62 年京大・工・電気卒．平成元年同大学院修士課程修了．同年（株）東芝入社．主として音声認識・文字認識の研究に従事．現在，同社研究開発センター知識メディアラボラトリー研究主幹．電子情報通信学会，日本音響学会各会員．