音声と言語の一体型学習に基づく音声認識技術

機械と人のより自然なインタラクションの実現を目的とした，話し言葉の音声認識技術が注目を集めています．話し言葉の音声認識は，世界のさまざまな研究機関が挑戦してきましたが，依然として困難な技術であるとされてきました．そこで私たちは，深層学習と呼ばれる最先端の学習理論を応用し，音声と言語の一体型学習に基づく音声認識技術を提案しました．本稿ではその技術と適用例について紹介します．

久保 陽太郎 /小川 厚徳
堀 貴明 /中村 篤

NTTコミュニケーション科学基礎研究所

音声認識の進展と課題

音声認識とは，マイクで収音した音声の信号パターンから，その内容を単語列として抽出する課題です．ここ数年，音声認識技術は普及の一途をたどり，現在ではさまざまなサービスで音声認識を使うことが可能になってきています．

サービスの展開と同時に，音声認識技術の基礎研究も大きく進展してきました．20年前にはあらかじめ登録した話者の音声を認識するのが精一杯であった音声認識も，10年前には正しく発声されていれば誰の声でも認識できるようになってきました．こうした流れの中で，これまでよりさらに多様な応用形態，例えば講義・講演への自動字幕付与や会議録の自動作成などを可能にする技術として，現在「話し言葉」の認識が大きく注目を集めています．

話し言葉の認識とは，これまで暗黙のうちに仮定してきた，「正しく発声されていれば」という条件を緩和し，正しく発声していない音声であっても，人が人に話すときに通じるレベルの音声であれば高精度に認識するという課題です．正しくない発声は，さまざまな要因が複合的に作用することによって起こっています．会話ではよく，「あー」や「えー」などといったフィラーと呼ばれる発声や，助詞の省略のような文法的逸脱が起こります．音声認識は文法的な知識も利用して認識を行うため，このように文法的に不正確な発話が起こると，手掛かりが不正確になることから精度の悪化が起こります．加えて，発音の間違いも問題となります．人間には正確に聞こえる音声であっても，計算機で分析し観察してみると，母音の省略や不安定な発声運動などをみることができます．このような不正確さも話し言葉の認識を難しくしている原因であると考えられています．

従来，音声の文法的側面は「言語モデル」，また音響的側面は「音響モデル」というモジュールによって表現されてきました．上述したような話し言葉音声認識の課題は，文法的側面，音響的側面の両方に横断して表れるため，非常に困難な問題として認知されてきました．

近年，深層学習（Deep Learning）と呼ばれる機械学習分野の新技術が，音響モデルの高精度化に貢献しました．2011年，深層学習を用いた音響モデルによって，話し言葉の認識精度を大幅に向上させることができるという発表が国際会議であり，聴衆を大いに驚かせました[1]．しかし，この深層学習もまだ音声と言語の両面に横断的に表れる話し言葉特有の問題に片面，すなわち音響的側面からしか対処していません．

本稿では，私たちの提案している，深層学習に基づく音声言語一体型学習技術について紹介するとともに，それがどのように話し言葉音声認識を進展させたかについて，英語の講義音声を例に取りながら紹介します．

重み付き有限状態トランスデューサ

私たちの音声認識システムを動作させるときに大きな役割を担っているのが，重み付き有限状態トランスデューサ（WFST: Weighted Finite-State Transducer）です．WFSTは系列データを変換する規則を表現する，状態と状態遷移からなる抽象的な機械であり，現在の音声認識に必要なさまざまな種類の統計処理はすべてWFSTで表すことが可能であることから，音声認識を記述するための統一的表現法として利用されてきました[2]．本稿で紹介する一体型モデルではWFSTを

用いた統一的な表現を拡張することで，話し言葉における音声と言語の双方に表れる現象の表現を追求します．

音素列を単語列に変換するWFSTの例を**図1**(a)に示します．図中の丸は機械の状態を表し，矢印はその状態からその状態に遷移し得るということを表しています．矢印に付記されている数値は，その状態遷移の起こりやすさを記述する確率で，付記されている2つのシンボル（例えば「“a”/“赤い”」）はその矢印をたどっている最中に“a”を入力として受け付け，“赤い”を出力するということを表しています．またシンボル“ε”は何も入出力しないことを表す特別なシンボルです．この機械による変換処理は初期状態（状態1）からスタートし，矢印をたどり続けて，その過程で入力系列からシンボルを読み込み，出力系列へ変換されたシンボルを出力することで最終状態（状態6）に至るまで続きます．

WFSTを使う利点の1つとして，その合成アルゴリズムの研究が進展していることが挙げられます．単語の連なりを考えて，不自然な文の確率を小さくするWFSTの例を図1(b)に示します．このWFSTは単語列の起こりやすさを考えているだけなので，実際にシンボルの変換は行わないのですが，図1(b)のように入力と出力に同じものを持つWFSTとして書くことでWFSTの一種として考えることができます．合成アルゴリズムは図1(a)と図1(b)の2種類のWFSTを受け取り，(a)のWFSTの変換結果をそのまま(b)のWFSTに入力した際に考え得るすべての入出力系列のパターンを列挙するようなWFSTを図1(c)のように生成します．この合成アルゴリズムを利用することで，音声の短時間スペクトルから得られる音響パターンから音素状態と呼ぶ中間表現に変換するWFST，音素状態の連なりを音素に変換するWFST，そして音素から単語への変換を表すWFST（図1(a)），単語列の確率を示すWFST（図1(b)）のすべて

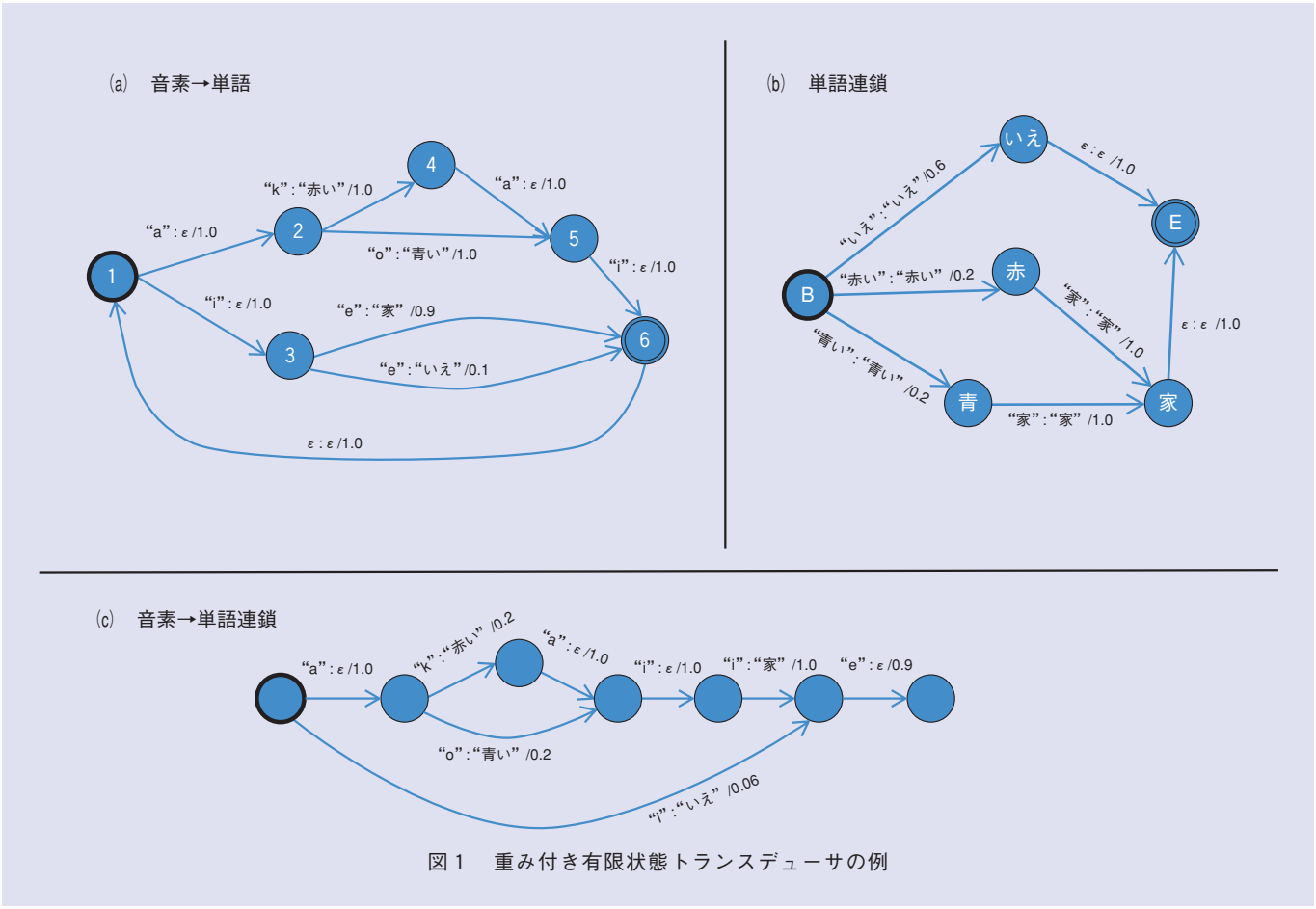


図1　重み付き有限状態トランスデューサの例

を合成することで，音声認識を表現する巨大なWFSTを構築することができます．音声認識はそのようにして合成したWFST上で音響信号から単語への変換としてもっとも確率の高い経路を最短経路問題として探索することによって実行されます．

音声認識全体を表現するWFSTを構築するためには，合成前の要素WFST（例えば，図1(a)および(b)）をつくらなくてはなりません．これまでは専門家の知見やルールを用いることで状態遷移を示す矢印の設計を行い，その状態遷移に関連付いた確率値を機械学習技術によって推定するといったことが行われてきました．

しかし，これまでの要素WFSTを個別に推定する方法では，話し言葉のように，音響パターンと言語パターンに横断的に表れる現象をうまく表現するように確率値が調整できません．

深層学習による音響モデル

近年，トロント大学を中心とした研究グループから，深層学習を用いた音響モデルによって，事前に複雑な音響パターンの補正を行うことなく，高精度な認識ができることが示されました[3]．深層学習は，30年以上前より続くニューラルネットワーク（NN）と呼ばれる手法が進展することによって，これまで学習が困難であった種類のNNの利用が可能になり確立された分野です．

本稿で対象とするタイプのNNは，入力を表すベクトル$\boldsymbol{x}=(x_1, x_2, \cdots, x_D)^T$が入力されたときの出力$\boldsymbol{y}=(y_1, y_2, \cdots, y_{D'})^T$を以下のような数式で計算するアルゴリズムで，パラメータ，すなわち$w_{i,j}^{(l)}$および$b_j^{(l)}$を適切に調整することで$\boldsymbol{x}$と$\boldsymbol{y}$の関係を表現します．

$$
\begin{aligned}
y_j(\boldsymbol{x}) &= h_j^{(L)}(\boldsymbol{x}), \\
h_j^{(l)}(\boldsymbol{x}) &= f\left(\sum_{i=1}^{D} w_{i,j}^{(l)} h_j^{(l-1)}(\boldsymbol{x}) + b_j^{(l)}\right) \\
h_j^{(0)}(\boldsymbol{x}) &= x_i \\
f(z) &= \frac{1}{1+e^{-z}}
\end{aligned}
\qquad (1)
$$

ここでeはネイピア数（自然対数の底）です．

学習データとして用意した$\boldsymbol{x}$と$\boldsymbol{y}$のペアを良く表現するよう，パラメータ（$w_{i,j}^{(l)}$および$b_j^{(l)}$）を調整することで，運用時に新しい$\boldsymbol{x}$に対応する$\boldsymbol{y}$の推定値を得ることができます[4]．音声認識の場合，$\boldsymbol{x}$は音響パターンを表現したベクトル，$\boldsymbol{y}$は音響モデルで利用される確率値として応用されます．式(1)はL回再帰する再帰関数によって計算可能であり，このLを層数と呼びます．

深層学習は，これまで困難であるとされてきた層数の大きなNNを学習する方法です．深層学習によるNNの学習では，事前学習と呼ばれるステップを導入します．事前学習では，$\boldsymbol{y}$は用いずに，$\boldsymbol{x}$の効率的な表現を実現するためのパラメータを追求します．このステップを通常の学習，すなわち$\boldsymbol{x}$と$\boldsymbol{y}$の関係の表現を追求する学習の事前に行うことで，層数の大きなNN（例えば$L>3$）のパラメータを適切に調整できることが示されました．

多層のNNは入力パターンを少しずつ変化させながら，最終的に確率値を計算する手法なので，超多層のNNが可能になることで，入力パターンの高度な補正と確率値の計算を同時に一枚岩のモデルで表現することが可能となりました．

深層学習による音声言語一体型学習モデル

WFSTは合成アルゴリズムによって，音響パターンの確率値と単語の並びの確率値をすべて包含した巨大な一体型WFSTを構成することができます．また，深層学習による音響モデルは音響パターンの確率計算において，入力パターンの補正と確率値の計算を一体型のモデルで行うことができます．以上の2つを統合することによって，入力パターンの補正法，音響パターンの確率値の求め方，単語の並びの確率値の求め方をすべて一体として最適化するのが，私たちの提案している「深層学習による音声言語一体型学習モデル」です[5]．

音声言語一体型学習モデルでは，合成した後のWFSTの各状態遷移の確率値として，従来のように要素WFST内の確率値の積を利用することをやめ，直接それを深層学習によって計算します．具体的には式(1)における$\boldsymbol{y}$を音響モデルで利用される確率値ではなく，合成後のWFSTの状態遷移確率そのものを表現するとして学習を行います．

これによって，従来暗黙のうちに導入されていた，音響的な多様性と言語的な多様性が独立，すなわち相互に関連せずに起こるという仮定から脱却することに成功しました．話し言葉特有の現象は音響言語横断的に起こることが多く，相互に依存しているという観察から，この拡張は自然であるといえます．また，提案法もWFSTの構造を利用しているため，従来の音声認識器

のために開発されてきた，高速な音声認識技法が提案した理論の上にも展開できます．

この手法を用いることで，話し言葉音声の1つである講義音声の認識において，従来の性能を大きく超える結果を得ました．講義音声認識における単語誤り率を**図2**に示します．図2で「従来法」と示してあるのが深層学習導入以前の音声認識率です[6]．「深層学習」で大幅に誤り率が低下していることから，いかに深層学習が音声認識の研究者に衝撃を与えたか見ることができます．「一体深層学習」と示してあるものが，提案した手法によるものです．従来法を大きく超えるだけではなく，深層学習の結果をも大きく上回っていることが分かります．

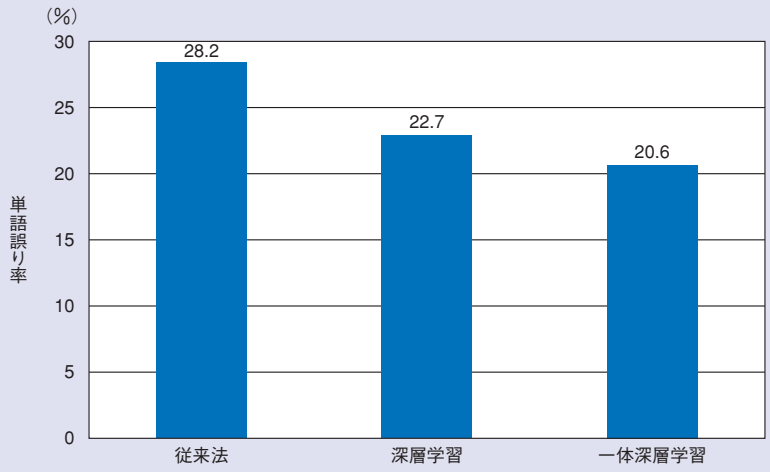


図2　英語講義音声認識実験における単語誤り率

今後の展開

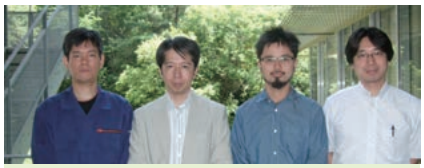
今後の展開には2つの方向性が考えられます．1つは，より深層学習と音声認識に関する深い理解を追求するという方向です．深層学習は，理論的解析が非常に困難なことから，その優位性は現在まで，実験によってしか示されていません．しかし，今後安心して深層学習技術を応用していくにあたり，深層学習のどこがどのように優れているのかを理解する必要があると考えています．

もう1つの方向性は，より現実的な計算時間で動かすための検討になります．現在の深層学習も十分現実的な時間で動作させることが可能なのですが，現在はGPU（Graphic Processing Unit）を用いた高速化を積極的に活用しており，普通のPCで高速に動作させるには，高度な検討を必要とします．またパラメータの調整プロセスはGPUを用いても非常に長い時間を必要とし，これも実サービスを展開するうえでは問題になり得ると考えています．

以上2つの方向性を通して，音声認識技術がより多様な応用形態で，より高い精度を発揮することができるよう，検討を続けていきます．

■参考文献

(1) F. Seide, G. Li, and D. Yu：“Conversational Speech Transcription Using Context-Dependent Deep Neural Networks,” In Proc. of INTERSPEECH 2011, pp.437-440, Florence, Italy, Aug. 2011.
(2) 堀・塚田：“重み付き有限状態トランスデューサによる音声認識,” 情報処理堀・塚田：“重み付き有限状態トランスデューサによる音声認識,” 情報処理，Vol.45，No.10，pp.1020-1026，2004.
(3) 久保：“ディープラーニングによるパターン認識,” 情報処理，Vol.54，No.5，pp.500-508，2013.
(4) 中野：“ニューラル情報処理の基礎数理,” 数理工学社，2005.
(5) Y. Kubo, T. Hori, and A. Nakamura：“Integrating Deep Neural Networks into Structured Classification Approach based on Weighted Finite-State Transducers,” In Proc. of INTERSPEECH 2012, Portland, U.S.A., Sept. 2012.
(6) E. McDermott, S. Watanabe, and A. Nakamura：“Discriminative training based on an integrated view of MPE and MMI in margin and error space,” In Proc. of ICASSP 2010, pp.4894-4897, Dallas, U.S.A., March 2010.



（左から）小川　厚徳/ 中村　篤/
久保　陽太郎/ 堀　貴明

音声認識技術の実用化研究開発を推進するNTTメディアインテリジェンス研究所音声言語メディアプロジェクトとも連携を取りつつ，音声認識技術の適用可能範囲を拡大していきます．

◆問い合わせ先
NTTコミュニケーション科学基礎研究所
メディア情報研究部　信号処理研究グループ
TEL　0774-93-5310
FAX　0774-93-1945
E-mail　kubo.yotaro@lab.ntt.co.jp