一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
PRMU2015-111(2015-12)

［特別講演］　音声認識技術の展開

河原　達也†

†京都大学 情報学研究科
〒606-8501 京都市左京区吉田本町
E-mail:　†kawahara@i.kyoto-u.ac.jp

あらまし　音声認識技術は近年急速な進歩を遂げており、音声書き起こしや音声検索・音声対話・音声翻訳など多くの実用化がされている。本稿では，音声認識の歴史と基本的な方法論を概観した上で、最近の技術革新の基盤となったニューラルネットワークによるモデルとビッグデータパラダイムによる学習について述べる。
キーワード　音声認識，ニューラルネット

[Special Talk]　State of Speech Recognition Technology

Tatsuya KAWAHARA†

†School of Informatics, Kyoto University
Sakyo-ku, Kyoto, 606-8501 Japan
E-mail:　†kawahara@i.kyoto-u.ac.jp

Abstract　Speech recognition technology has made a significant progress over the past decade and has been used in speech transcription, voice search, dialogue and translation systems.　This article gives a brief overview of the history and methodology of speech recognition and then recent technical innovations based on neural networks and the big data paradigm.
Keywords　speech recognition, neural network

1. 音声認識の歴史

音声認識は長い間ＳＦの範疇であった反面、なかなか実用レベルに到達しない技術であった。しかし２１世紀に入って、機械学習の方法論と計算機・情報通信技術(ICT)の進歩に伴って、飛躍的な性能改善を遂げ、様々な実用化が行われた。今では、スマートフォンに搭載されている音声検索やアシスタントアプリは多くの人に認知され[(1)]、音声翻訳アプリも複数リリースされている。また、テレビ放送の字幕付与や国会の会議録作成に音声認識技術が導入されるに至っている。さらにこの数年の間で、ニューラルネットワークに基づくモデルにより一層の性能の向上が実現されている。本稿ではまず、歴史的経緯を簡単に振り返り、最近の技術革新の主な要因について述べる。

音声認識の研究が開始されたのは今から 50 年以上も前に遡る。京都大学では 1960 年頃に単音節単位の認識を行う「音声タイプ」が構築されている[(2)]。その後、音声認識に有効な音響特徴量と、DP マッチングに代表される動的パターンのマッチング手法に関する基礎的な研究が世界中で行われた。これは、パターン認識の観点からはテンプレートベースの方法といえる。特定話者の音声認識は何とか動作しても、多数話者のバリエーションをモデル化するには不十分であった。

これに対して、確率的なモデルを導入することにより解決が図られた。DP マッチングを拡張した形で隠れマルコフモデル(HMM)が導入され、その改良が 20 年以上にわたって行われた。まず、HMM の各状態の音響特徴量のパターンを連続分布でモデル化する混合ガウス分布(GMM)が導入された。そして、これを最尤推定する代わりに、識別誤りが最小化されるように学習（識別学習）するための様々な方法が提案された。2000 年代に実用化された音声認識システムは、基本的に GMM-HMM の識別学習に基づくものである。一方、言語モデルについては、単語の連接規則（文法）をオートマトンで記述したものから、その遷移を確率的なものにし、その確率をコーパスから最尤推定する N-gram モデルに移行していった。以上の変遷をまとめたのが表 1 である。世代の定義は古井[(3)]に従ったものであるが、第 4 世代は著者が追加したものである。この第 4 世代が、ニューラルネットワークに基づくモデルである。音響モデルについては、GMM による確率計算をディープニューラルネットワーク(DNN)に置き換えた DNN-HMM が、言語モデルについては、リカレントニューラルネットワーク(RNN)を N-gram と併用するモデルが一般的になっている。この展開について次節で述べる。

表１　音声認識の方法論の変遷

第１世代	1950　～ 1960 年代	ヒューリスティック
第２世代	1960　～ 1980 年代	テンプレート (DP マッチング，オートマトン)
第３世代	1980　～ 1990 年代	統計モデル (GMM-HMM, N-gram)
３.５世代	1990　～ 2000 年代	統計モデルの識別学習
第４世代	2010 年代	ニューラルネット (DNN-HMM, RNN)

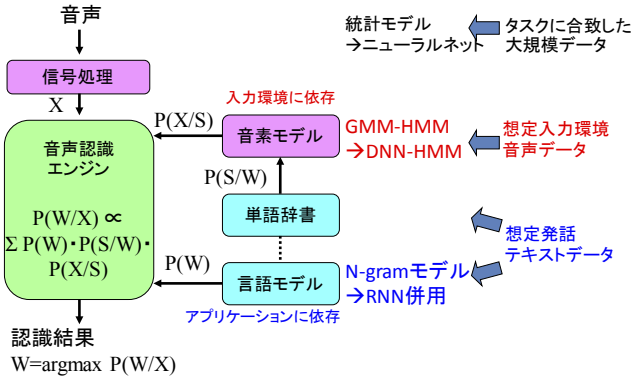

図１　音声認識システムの構成

2. 音声認識の原理とシステム構築法

音声認識は、音声（の特徴量）X が与えられたときにその単語列 W を同定する問題である。これは、以下の式(1)のように、p(W|X)をベイズ則で書き換えて得られる２つの項の積が最大となる W を同定する問題として定式化される。

$$\arg\max p(W \mid X) = \arg\max p(W)p(X \mid W) \qquad (1)$$

実際には、単語は音素などのサブワード単位 S でモデル化され、単語と音素の関係は辞書で決定的に与えられる(p(S|W)={1,0})ので、右辺の中身は以下のようになる。

$$p(W)p(X \mid W) = \sum_{s} p(W)p(S \mid W)p(X \mid S) \approx \max p(W)p(X \mid S) \qquad (2)$$

これは、単語列 W の言葉が音声という雑音のある通信路を伝わってきたのを情報理論に基づいて復号するモデルである。p(W)は（その言語あるいは状況において）単語列 W が生成される先験的な確率であり、p(X|W)は単語列 W（音素列 S）から音声（音響特徴量）X が生成される確率である。

これは、音声認識が２つの確率モデルを推定する問題に分割され、各々が生成モデルとして定式化できることを意味する。具体的に、p(W)を計算するモデルは言語モデルと呼ばれ、時系列(left-to-right)に探索するという制約・相性から単語 N-gram モデルが主に採用されてきた。これは、テキストデータを収集して単語連鎖（２つ組・３つ組）の出現頻度を計数すれば最尤推定できる。ただし、実際にはスムージングを要する。一方、p(X|S)を計算するモデルは音響モデルと呼ばれ、音素の状態毎に音声の音響特徴量の分布を GMM でモデル化する HMM が採用され、EM アルゴリズムによる最尤推定がベースラインの手法となった[4]。

以上述べた原理に基づく音声認識システムの構成を図１に示す。この枠組みは 1990 年頃に確立され、以降四半世紀以上にわたって、世界中（あらゆる言語）において普遍的に用いられてきた。しかしながら、（言語を特定しても）あらゆる用途に用いることができる普遍的・万能な音声認識システムが存在するわけではない。図１に記しているように、音響モデルは、音声認識システムが使われるアプリケーションの入力環境、具体的には音響条件・話者層・発話スタイルに合致するように、データを収集して学習する必要がある。言語モデルと単語辞書は、アプリケーションのタスクドメインに合致するように、想定発話のデータを収集して学習する必要がある。なお、音声認識エンジンは普遍的になっているが、技術的に高度・複雑になっているので、世界中でも著者らが開発してきたJulius[1]を含めて少数になっている 。

要するに、音声認識の原理や音声認識エンジンは普遍的でも、万能な音声認識システムが世の中に存在するわけでない。アプリケーション毎に合致したモデルを構築する必要があり、このモデルの善し悪しが認識性能を左右する。モデルの善し悪しは、最先端（といってもかなり標準的）の技術を用いたとすると、学習データベースの規模が最も重要になる。したがって、音声認識システムの開発は、(1)アプリケーション設計, (2)データ収集, (3)モデル学習という流れにより構成される。

このように音声認識システムの構成論は２１世紀初めに確立されたように思われたが、この５年ほどの間にさらなる技術革新が行われた。一つは、ニューラルネットワークの導入であり、もう一つはビッグデータパラダイムである。著者を含めて多くの人にとって認識性能が本当によくなったと実感できるようになったのは、これら２つによるものである。以下に各々について説明する。

1 http://julius.osdn.jp/

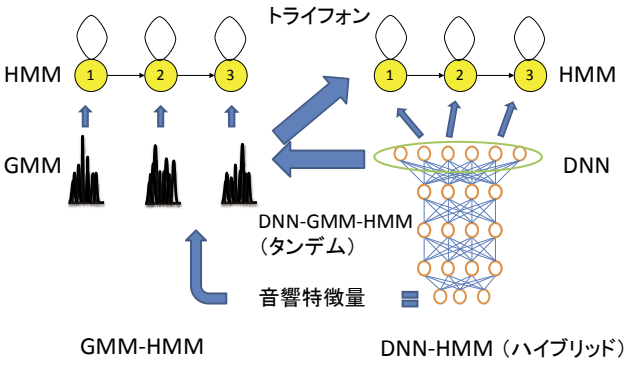

図２　GMM-HMM と DNN-HMM

3. ニューラルネットワークの“逆襲”

音声認識においてニューラルネットワークを用いることは、1990 年頃にも盛んに研究が行われ、音素識別などではよい性能が報告されたものの、確率的な枠組みに基づく連続音声認識システムでは統計的なモデルである混合正規分布(GMM)に基づく隠れマルコフモデル(HMM)が標準的になった。これに対して、近年再びニューラルネットワークが注目されるようになり、数年のうちに主流になった。1990 年頃と比べて、入力特徴量（セグメント：数百次元）、出力カテゴリ（トライフォン状態：数千クラス）、中間層の層・ノード数ともに、巨大化したのが最大の特徴であり、ディープニューラルネットワーク(DNN)と呼ばれる。多層のネットワークを逐次的に事前学習した上で、全体をバックプロパケーション学習するディープラーニングにより、大規模なネットワークの学習が可能になった[5,6,7]。

最初は音素識別を行う音響モデルの置換が主であったが、その後、単語予測を行う言語モデルや雑音・残響の抑圧においてもニューラルネットワークの導入が進み、現在では国際会議等においてニューラルネットワーク以外の論文は少数派になっている。

3.1. DNN-HMM による音響モデルの構成

音素識別を行う音響モデルに DNN を用いる直接的な方法は、GMM-HMM における各状態の GMM による確率計算を DNN に置き換えるものであり、DNN-HMM ハイブリッドシステムと呼ばれる。別の方法として、DNN の出力もしくは中間層の値を特徴量（当該層のノード数を抑えた場合、ボトルネック特徴量と呼ばれる）として用いて、GMM を学習する DNN-GMM-HMM タンデムシステムもある。これらを図２に示す。

DNN-HMM ハイブリッドシステムでは、図２の右側に示すように、音響特徴量を DNN に入力する。DNN では出力層まで順次各層の計算を行う。各層のノード j では、これに結合する前の層のすべてのノード i の出力値 y_i と結合の重み w_{ij} の積和を求め、バイアス項 b_j を加えたものに、非線形関数を適用することで出力値を得る。

$$y_j = f(\sum w_{ij} * y_i + b_j)$$

中間層では、この関数にしきい値関数を模したシグモイド関数（２クラスロジスティック回帰）かハイパボリックタンジェント(tanh)関数が用いられるが、出力層では、全ノードに対する事後確率を計算するために softmax 関数（多クラスロジスティック回帰）が用いられる。出力層のノードは、HMM の各状態に対応付けられるが、一般的な音声認識ではトライフォンモデルの共有状態となる。これは、先行音素と後続音素の文脈を考慮したもので、クラスタリングを行っても数千個のオーダになる。DNN による確率計算は、GMM と比べて計算量が大きいが、単純な行列計算の組合せであるので、GPU による高速化が容易であり、リアルタイムの認識も十分に可能である（GPU なしでは困難である）。

各種の音声認識タスクにおいて、DNN-HMM が従来の GMM-HMM を凌ぐ認識精度を得られることが示されている。種々のベンチマークの結果を表２に示す。音素認識から大語彙連続音声認識までの様々なタスクにおいて、誤り率を概ね 20～30%削減している。単一の方法により認識精度がこれほど大幅に改善したことは、著者の知る限りほとんどなく、非常に画期的なことであった。

DNN が GMM に比べて優れている理由については様々な説明がされているが、最大の理由は、識別器に特徴抽出を統合して最適化しているためであろう。ただし画像認識では、特徴抽出を明示的に行う CNN が一般に用いられ、各層の解析も行いやすいのに対して、音声認識においては各層の役割や意味は明示的でない。

表２　GMM-HMM と DNN-HMM の比較

	学習データ(時間)	GMM-HMM 単語誤り率	DNN-HMM 単語誤り率
TIMIT 音素認識	10	27.3%	22.4%
Switchboard 電話音声	300	23.6%	17.1%
Google 音声検索	5870	16.0%	12.3%
JNAS 日本語新聞記事	85	6.8%	3.8%
CSJ 日本語講演音声	257	20.0%	16.9%

上段３つは文献(4)より引用，下段２つは著者による

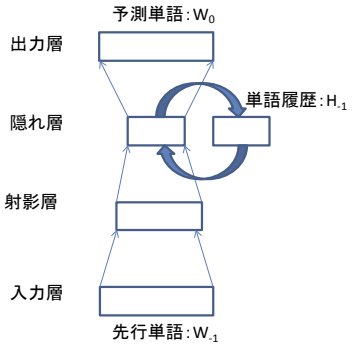

図３　リカレントネットワーク(RNN)言語モデル

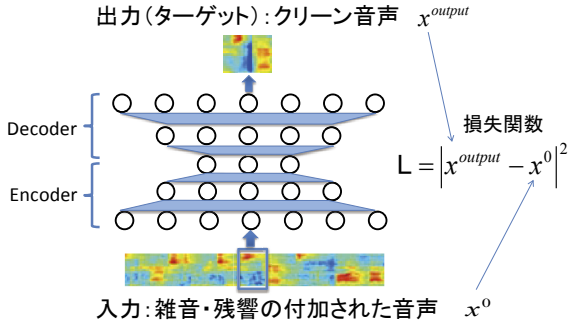

図４　デノイジングオートエンコーダ(DAE)による雑音・残響抑圧

従来は、当該フレームのメル周波数ケプストラム係数(MFCC)やその回帰係数(ΔMFCC)などが主な特徴量として用いられてきたが、DNN を用いる際には、比較的広い範囲（前後 11 フレーム程度）のフィルタバンク出力をそのまま用いるのが最も効果的とされている。“生”の周波数特徴量を与えて、特徴抽出もニューラルネットワークの学習に委ねるブラックボックス化の発想といえる。これに対して、GMM-HMM を学習するには、統計的推定の信頼性の点から特徴量の次元をあまり大きくできず、しかも無相関にすることが望ましいとされていた。そのため、MFCC や ΔMFCC に変換していたのであるが、このような単純な特徴抽出が性能のボトルネックになっていたことを示唆している。

3.2. RNN による言語モデル

音声認識における言語モデルは、単語列の履歴から次の単語を確率的に予測し、尤度を与えるものである。これにより単語候補を絞ったり、言語的に意味のある単語列が認識結果に現れやすくする。従来は、N 単語列の頻度に基づく N-gram モデルが一般的に用いられてきたが、履歴長 N をあまり大きくできない（通常は N=3）という問題があった。これに対して、中間層の出力を次の入力にフィードバックさせるリカレントニューラルネットワーク(RNN)を用いたモデルの導入が進められている[8]。これを図３に示す。入力単語を少ないノードの数値データに射影する層を別途用意する。中間層はこれと履歴を符号化したものと捉えられ、N-gram モデルと比べて非常に長い履歴を考慮することができる。ただし、N-gram モデルの方が低頻度語のスムージングが効果的に行えることもあり、N-gram モデルと併用・線形補間する場合が多い。リアルタイムの認識に組み込むのは容易でないが、従来の N-gram モデルで生成した n-best 候補に対してリスコアリング（別の高精度なモデルで尤度を再評価）する枠組みで概ね 5～10%程度誤り率の改善が得られることが報告されている。

3.3. DAE による雑音・残響抑圧

音声認識を実環境に適用する際に雑音や残響の抑圧が大きな課題となる。従来はスペクトル減算やウイーナフィルタのような統計的な手法が一般的であったが、最近ニューラルネットワークの検討が進められている。主として、図４に示すようなデノイジングオートエンコーダ(DAE)が用いられる。これは、入力を多段のネットワークにより符号化した後に、復号化するものであるが、雑音や残響が付加された音声（前後のフレームも含めた特徴量）を入力とすることで、元のクリーンな音声を推定・復元するように学習する。学習には、ウイーナフィルタと同様に入力とターゲットの二乗誤差最小化基準が用いられるが、非線形なネットワークを用いるのが特徴である。これにより、従来手法と同等以上の性能が実現されている。特に、学習の際に想定されていた雑音・残響条件とミスマッチがあっても頑健に動作する傾向が確認されている[9]。

3.4. 話者・環境適応

音声認識では、新しいタスク・環境や話者に徐々に適応する機能が重要である。DNN を用いる認識手法は、GMM や N-gram などの確率統計モデルに比べて高い性能を実現するが、少量のデータで適応を行う方法が確立されていない。

最もナイーブな方法は、適応データで追加的にバックプロパゲーション学習を行うものであるが、大規模なデータで学習されたネットワーク全体のパラメータは膨大であるので、少量のデータで再学習を行うのは不安定になる恐れがある。そこで、正則化などを導入したり、特定の層やノードのみを話者・環境依存のものとして設定し、これらのみを学習する方法が考えられている。あるいは、ネットワークの一部のパラメータ、例えば各ノードの出力のゲイン値のみを学習することも検討されている。

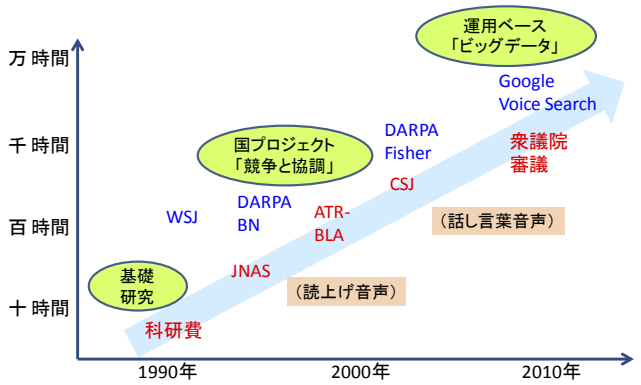

図５　代表的なデータベースの構築時期とデータ量

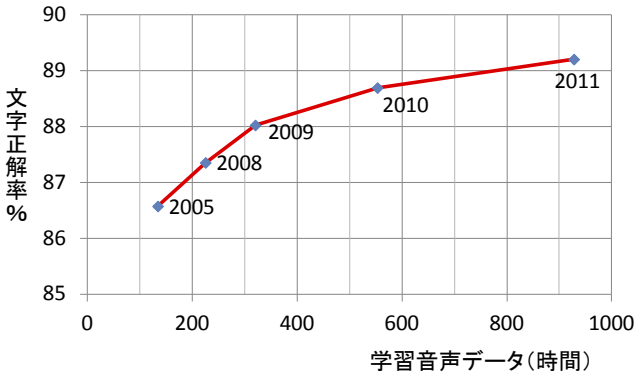

図６　国会審議音声認識におけるデータ量と認識率

応用例は限られるが、クラスタ毎（例えば男女別）にモデルを構成・学習しておいて、その線形補間値のみを適応データで推定することも考えられる。

別のアプローチとして、あらかじめ入力に話者や環境に関する補助情報を追加して、ネットワークを構成・学習する方式がある。例えば、話者性をコンパクトに表現した i-vector や環境雑音の成分を追加することが代表例である。この場合ネットワーク自体を再学習する必要はない。

4. データベース構築の限界 --ビッグデータパラダイム

音声認識システムの構成において、タスクに合致した学習データ量が鍵であることは先に述べた。そのため、研究コミュニティを挙げて、大規模な音声・テキストデータベースの構築が進められた。

図５に、代表的な音声データベースの構築時期とデータ量（時間数）をプロットしたものを示す。時代とともに、対象が読上げ音声から話し言葉音声に移行し、それに伴ってデータサイズが大規模化していることがわかる。さらに、図６に著者らが開発している国会審議の音声認識システムの音響モデルの学習音声データ量と認識精度の関係を示す[10]。線形ではないが、単調に改善していることがわかる。言語モデルの学習テキストデータ量についても、また他のシステムでも同様の報告がされている[11]。

それではどのようにして、これだけ大規模なデータを集めるのであろうか。音声に限らず、文字や画像などのパターン認識の研究においては、単独の研究機関でデータベースを構築するのが困難なため、研究コミュニティで協力してデータを収集することがよく行われてきた。実際にこの「協調と競争」パラダイムは、1990 年代に世界的に成功を収めた。

しかし最近では、この「データを頑張って集める」という発想自体が限界になってきている。実際に、そうやって頑張って集められるのはせいぜい数十～数百時間が限界である。また、被験者を集めて収集したデータが、実際のユーザが発話するものと適合するかも不明である。したがって、リアルなデータを自然に集積できる枠組みを構築することが考えられた。このようなビッグデータパラダイムが、音声認識の最近の成功の鍵となっている。

以下にその２つの典型的な事例について述べる。

4.1. 携帯端末用クラウドサーバ型システム

携帯端末、特にスマートフォンのアプリケーションでは、クラウドサーバ型の音声認識が用いられている。これは、携帯端末に入力された音声をパケット化し、ネットワーク経由でサーバに送信して、認識処理を行うものである。これにより、端末の処理能力・記憶容量を気にせずに、大規模なモデルを用いた高精度な音声認識が可能になった。さらに重要な点は、ユーザの発した音声データをサーバ側に蓄積できることである。サービスは無償のものが多く、利用者は数百万人にも達する[11]ので、リアルなデータが巨大な規模で蓄積されている。Google の英語の音声検索では、発話データが１万時間規模になっている[7]。

4.2. 会議音声と会議録の活用

会議や講演などの話し言葉の音声認識システムを構築するには、そのような音声とその忠実な書き起こしテキストを用意する必要がある。会議や講演は毎日のように行われるので、その音声を収録すること自体は容易である。しかし、これらには通常書き起こしがない。議会の場合は逐語的な会議録が作成されるが、忠実な書き起こしではなく、そのままでは音声認識のモデル学習には使えない。そこで著者らは、会議録のテキストから実際の発言内容を確率的に予測する枠組み

を考案した。例えば、「あのー」などのフィラーがどこに入りやすいかも予測することができる。この枠組みによって、会議録から話し言葉の統計的言語モデルを推定するとともに、会議録と音声から発言内容を復元し、千時間規模の会議音声からほぼ自動的に音響モデルの学習が可能になった[10]。この効果が図 6 に示されている。

5. 今後の展望

音声認識はかなり高度になり、実用レベルになったとはいえ、基本的には（能力の高い）外国語話者の域を出ない。一般人の話し言葉にはほとんど対応できないし、騒音下ではとたんに性能が低下する。母語話者のようなリスニング能力が実現されるのは想像できないくらい先のことのように思われ、それにはまだまだ素朴なブレークスルーが必要と思われる。

例えば、現在の音声認識システムでは、周波数特徴量に関する音響モデルと局所的な単語連鎖に基づく言語モデルの尤度のみしか用いていないが、韻律に関するモデルや、意味や話題を考慮した高次・大局的な言語モデルを組み合わせることが期待される。そのためには、伝統的な式(1)の定式化から脱却し、一般的な情報統合の枠組みを構成する必要がある[12]。

文　献

[1] 河原達也. 音声対話システムの進化と淘汰 ―歴史と最近の技術動向―. 人工知能学会誌, Vol.28, No.1, pp.45--51, 2013.

[2] T.Sakai and S.Doshita. The Phonetic Typewriter. Proc. IFIP Congress 62, pp.445-450, 1962.

[3] S.Furui. Selected Topics from 40 Years of Research on Speech and Speaker Recognition. Proc. InterSpeech, pp.1-8, 2009.

[4] 鹿野清宏, 伊藤克亘, 河原達也, 武田一哉, 山本幹雄. 音声認識システム. オーム社, 2001.

[5] 久保陽太郎. 音声認識のための深層学習. 人工知能学会誌, Vol.29 No.1, pp.62-71, 2014.

[6] D.Yu and L.Deng. Automatic Speech Recognition – A Deep Learning Approach. Springer, 2015.

[7] G.Hinton, L.Deng, Y.Dong, G.E.Dahl, A.Mohamed, N.Jaitly, A.Senior, V.Vanhoucke, P.Nguyen, T.N.Sainath and B.Kingsbury. Deep Neural Networks for Acoustic Modeling in Speech Recognition. IEEE Signal Processing Magazine, Vol.29, No.6, pp. 82-97, 2012.

[8] T.Mikolov, M.Karafiat, L.Burget, J.Cernocky, and S.Khudanpur Recurrent Neural Network Based Language Model. Proc. INTERSPEECH, pp.1045-1048, 2010.

[9] M.Mimura, S.Sakai, and T.Kawahara. Reverberant speech recognition combining deep neural networks and deep autoencoders augumented with phone-class feature. EURASIP J. Advances in Signal Processing, Vol.2015, No.62, pp.1--13, 2015.

[10] 河原達也. 議会の会議録作成のための音声認識－衆議院のシステムの概要－. 情報処理学会研究報告, SLP-93-5, 2012.

[11] 辻野孝輔, 栄藤稔, 磯田佳徳, 飯塚真也. 実サービスにおける音声認識と自然言語インタフェース技術. 人工知能学会誌, Vol.28, No.1, pp.75-81, 2013.

[12] 河原達也. 音声認識の方法論に関する考察―世代交代に向けて―. 情報処理学会研究報告, SLP-100-3, 2014.