Vol.2015-SLP-108 No.6
2015/10/30

IBM Watson の SaaS 型音声認識・音声合成サービス

立花隆輝[†1]　福田隆[†1]　長野徹[†1]

概要：近年，ディープラーニング技術によって音声認識精度は一般に大きく向上した．IBM でも国内外の研究所でニューラルネットワークの構造や学習方法を改善する研究や，特定の用途や環境に適応するための研究などを複数の言語について行っている．基礎的な研究も行っている一方で，企業での応用現場での要求を直接聞き，単語誤り精度の削減のみに留まらない実用性を意識した研究も行っていることに特徴がある．このような研究成果を生かし，IBM は 2015 年 7 月から複数の言語について音声認識と音声合成の Software-as-a-Service (SaaS)型サービスの提供を始めた．本報告では，このサービスにも一部が実用されている IBM の国内外研究所の近年の主な研究成果を概観する．

キーワード：クラウド、SaaS、音声認識アプリケーション、音声合成アプリケーション

IBM Watson Speech To Text Service and Text To Speech Service

RYUKI TACHIBANA[†1]　TAKASHI FUKUDA[†1]
TOHRU NAGANO[†1]

Abstract: Recent advancements in the deep learning technology area have greatly improved the accuracy of Automatic Speech Recognition (ASR) in general. The research laboratories of global IBM are also conducting research in this area toward better topologies and better training methods of neural-network-based acoustic and language models, as well as research of adaptation methods to new environments and new usage scenarios. While rather fundamental research themes are sometimes studied, it is also one of the characteristics of our approaches that we sometimes target on practical usefulness at customers' fields beyond simple reduction of word error rates. Using such research achievements, IBM started new Software-as-a-Service (SaaS) of Automatic Speech Recognition (ASR) and Text To Speech (TTS) in July 2015 for multiple languages. In this paper, we introduce some of recent main research achievements of our research laboratories.

Keywords: Automatic speech recognition applications, Text-to-speech applications

1. はじめに

ディープラーニングを用いた近年の精度向上を背景に，音声認識が本格的に日常で用いられる時代がいよいよ訪れている．人間の音声を「人間が聞き取るように」聞き取るという目標は一見単純に思えるが，技術レベルが今のレベルに達するまで，当初想像したよりもはるかに長い時間がかかった．IBM はこの期間にわたり様々な手法を研究開発し，この研究分野に貢献してきた．そして，その流れの中で，様々な音声関係製品の商用化を行ってきた．その中にはパーソナルコンピューター向けのデスクトップ・ディクテーション製品や，電話自動音声応答システム（IVR）向けのサーバー製品，車載機器や銀行 ATM などの組み込み機器向けの製品などが含まれる．2009 年には IEEE からコーポレート発明賞を受賞したが，これは過去 35 年間に渡って生み出した上述のようなアプリケーションと技術革新に対して与えられたものである．IBM は音声技術を多数の言語に適用しており，日本語については大語彙連続音声認識ソフトウェアを 1997 年に発売した．そして今年，IBM は音声認識と音声合成の実用的なアプリケーションの開発を促進するために，クラウド・ベースの新しいサービスを開始した．本稿では，このサービスにも一部が実用されている IBM の国内外研究所の研究成果を，昨今のものを中心に概観する．

2. IBM の音声認識技術

現在, IBM では DNN や CNN に代表されるディープラーニングの研究に精力的に取り組んでいる．クラウド音声認識サービスにも多くのディープラーニング関連技術を取り入れて，音響モデルや言語モデルの構築に活用している．ディープラーニングが注目され始めた当初は，大語彙連続音声認識における最適なネットワーク構造を検討し，例えば CNN については，畳み込み層 2 層程度，プーリング層において時間方向のプーリングは行わずに，周波数方向のプーリングのみに限定する，畳み込み層の後に DNN 層を結合する，出力層の直前にボトルネック層を挿入するなどの基本的な枠組みを設計した[1]. その後，DNN の識別学習として位置づけられるシーケンストレーニング[2]や，話者適応技術[3], 構成の異なるネットワークの結合学習[4]などを研究し，性能の底上げを図ってきた．さらには，発話区

†1 日本アイ・ビー・エム株式会社東京基礎研究所
東京都中央区箱崎町日本橋 19-21
IBM Watson Multimodal, IBM Japan Ltd.
19-21, Nihonbashi Hakozaki-cho, Chuo-ku, Tokyo 103-8510, JAPAN
E-mail: {ryuki, fukuda1, tohru3}@jp.ibm.com

間検出においてもディープラーニング技術の導入を検討している[5]．その結果，最近では，米国 Switchboard タスク（電話会話音声）において単語誤り率 8%の精度を達成し，研究者らの注目を浴びた[6]. Switchboard に関して人間の聴取能力を単語誤り率に換算すると約 4%であるとの報告があり[7], IBM の音声認識技術が人間の聴取能力に非常に近づいていることを示す一例となった．

ディープラーニングが注目を浴びる以前は，特徴空間やモデル空間の識別学習[8]，VTLN や fMLLR に基づく話者適応化学習が大きな役割を果たし，2004 年には GMM/HMM システムで当時の最高値である単語誤り率 15.2%を達成した[12]. 特徴空間上の識別学習に利用可能な学習データが少量である場合にも環境適応を良好に行う方法として，正則化処理が有効であることを示した[9]．十分な分量の学習データが目標ドメインについて存在しない場合には，それぞれ異なる環境で録音された音声データを変換して利用する手法に大きな効果があった[10]. 逆に, それぞれ異なる環境で録音されたテストデータの精度を向上する手法として，相補的な複数の音響モデルを利用する手法を提案した[11,22]．音響的な大きな課題のひとつは会話に参加する複数話者の発話衝突である．例えば，対面に位置する二話者の会話音声の分離においては，空間的エイリアシングによる対向話者の方向性雑音の漏れが問題となるが, これを確率的なポストフィルタリングという新しい枠組みで改善を図っている[13]．電話のモノラル録音音声の認識においては，二話者による発話衝突がたびたび発生し，発話衝突による連鎖的な認識誤りが問題となっていた．これに対して，発話衝突専用の言語モデルを導入するという新しい技術を導入し，検討を重ねている[14].

ディープラーニングについても関連技術を簡単に述べておく. ディープラーニングに基づく IBM の標準的な音響モデル構築手順は，まず識別的事前学習によりネットワークを生成的に構築し，その後，クロスエントロピーに基づく誤差逆伝播法によりファインチューニングを行う．事前学習およびファインチューニングの際には，学習データをフレーム単位でランダムに入れ替え，数百フレーム毎にパラメータの更新を行うミニバッチ学習を採用している．その後，状態依存最小ベイズリスクを目的関数としたヘシアンフリー最適化アルゴリズムによるシーケンストレーニングを行い, 識別的な基準でモデルの更新をさらに繰り返す. シーケンストレーニングは，平均して相対的に 10%程度の誤りを削減することが示されている[2].

これらの基本的要素に加わる最近の知見として Dropout 処理を組み込んだ Maxout ネットワークの活用や，出力層のユニット数の拡充，また，畳み込み／非畳み込みネットワークの結合学習がさらなる認識性能の向上をもたらすことなどを見出した．音声認識において Maxout ネットワークの利用は，学習データが限定的な場合のみ大きな効果があるとされてきたが[15]，学習の過程で Dropout の割合を調整することにより，大規模データが利用可能な場合においても大きな改善を得られることを確認した[16]． 出力層のサイズに関しては, GMM/HMM が中心的に用いられていたころ，リーフサイズが 10K 以上の音素決定木は認識性能の改善にあまり効果がなかったが[4]，ディープラーニングにおいては，10K 以上の状態を持つ音素決定木が着実な改善を与えることを見出した．前述の Switchboard システムでは前後 3 音素からなる 32K もの状態数を採用している[6]．DNN の学習時に入力層で特に中心付近のフレームに重みを置くことによる改善も発見した[17]. 一方, 話者認識の分野で検討が盛んな i-vector を，DNN への入力の一部として併用する検討も行っている[3]．これは，DNN の入力特徴量に話者固有の情報を追加することによって，ディープラーニングにおける話者適応化学習を実現しようとするものである．

他方，音響モデルだけでなく，言語モデルにもディープラーニングの技術を取り入れている．標準的には，修正 kneser-ney スムージング法で構築した単語単位の 3-gram もしくは 4-gram を利用しているが, セカンドパスのリスコアリングに，指数モデルに基づくクラス言語モデルである model M[18]や，ニューラルネットワーク言語モデル（NNLM）[19]を使っている．これらの要素技術は，Switchboard タスクにおいても大きな効果をあげた．

実用の現場に目を向けると，認識結果に含まれるすべての単語（文字）が等しく重要なわけではないケースが多い．音声認識の認識精度の評価基準としては単語誤認識率や文字誤認識率を用いるのが一般的であるが，これらの尺度が認識結果の有用性を必ずしも反映していないという課題がある．誤りの挿入・脱落・置換誤り数に加え，誤りの連続数や位置，キーワードの誤り数などを基にランダムフォレストを用いることで人間の感覚スコアを精度高く予測する手法を開発した[20]. また, 音声認識結果から重要語を高精度に検出する研究も行っている[21].

3. IBM の音声合成技術

テキスト音声合成システムは，フロントエンドとも呼ばれる言語処理部とバックエンドとも呼ばれる波形生成部から構成されるのが一般的である．

(1) 言語処理部

IBM の日本語音声合成システムはフロントエンドに，＜表層，読み，アクセント，品詞＞の 4 つ組の N-gram 列を同時に推定するという統計的フロントエンド[23]を利用する．これは，文を音読する際のアクセントも含めてアノテーションをテキストに付随させたコーパスに基づいて 4 つ組の連接確率を学習するアルゴリズムであって，アクセント結合によるアクセント変化の予測も学習データにおける統計情報に基づいて行う．しかし，テキストコーパスの収集は

比較的容易であるのに対して，アクセントを付随させたコーパスの作成作業にかかるコストは大きいという問題がある．そのコストを削減する方法が幾つかある．まず，各語について可能なアクセント型の集合があらかじめわかっているという前提ではあるが，アクセントの付随していない音声データのアクセントを自動的に推定することで，フロントエンドの学習に利用可能なアクセント付きコーパスを作成する方法がある[24]. また，アクセント移動ルールに関する知見を，上述の枠組みにアクセントクラスによって導入する手法も精度の改善に効果があった[25].

(2) 波形生成部

一方でバックエンドには波形重畳編集を利用する．システム構築時には，まず話者依存 HMM によって音声のアライメントをとる．ここで音素当たり 2 または 3 状態を割り当て，この 1 状態相当の波形が最小の接続単位となる．状態ごとに音素コンテキストを説明変数とした音響決定木が構築され，素片は各 leaf に分類される．一方，先のアライメントを利用して F0 値，継続時間長，音圧を推定する決定木と，スペクトル連続性コストに関する決定木がそれぞれ構築される．日本語や中国語などの高低アクセント言語では F0 の絶対値よりも F0 の傾きを直接予測するモデルに効果がある[26]. 合成時には，言語処理部の出力を特徴量として，まず韻律が各決定木によって推定される．次に，推定した韻律を目標として，目標からの誤差および素片の接続性に関するコスト関数を最小化する素片が one-pass DP によって各 leaf から選択される．ただし連続する素片の選択には報酬を与える．目標韻律まで PSOLA などで素片を変形した上で重畳し，合成音声が作成される．ただし F0 に関しては目標 F0 をスムージングしたものが使用される[27,28].

4. おわりに

クラウド・ベース音声認識・合成サービスでは従来のアプリケーションとは異なる制約や要求もある．そのような制約を満たしつつさらに精度を向上する技術の開発が今後の課題である．

参考文献

1) T. N. Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhadran, “Deep Convolutional Neural Networks for LVCSR", Proc. ICASSP, pp. 8614-8618, 2013.
2) B. Kingsbury, T. Sainath, and H. Soltau, “Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization,” Proc. Interspeech, 2012.
3) G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, “Speaker adaptation of neural network acoustic models usin I-Vectors,” Proc. ASRU, pp. 55-59, 2013.
4) H. Soltau, G. Saon, and T. N. Sainath, “Joint training of convolutional and non-convolutional neural networks,” Proc. ICASSP, pp. 5609- 5613, 2014.
5) S. Thomas, S. Ganapathy, G. Saon, and H. Soltau, “Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions,” Proc ICASSP, pp. 2519-2523, 2014.
6) G. Saon, H. J. Kuo, S. Rennie, and M. Picheny, “The IBM 2015 English Conversational Telephone Speech Recognition System,” Proc. Interspeech, pp. 3140-3144, 2015.
7) R. P. Lippmann, “Speech recognition by machines and humans,” Speech Communication, vol. 22, no. 1, pp. 1-15, 1997.
8) D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, G. Zweig, “fMPE: Discriminatively trained features for speech recognition,” Proc. ICASSP, pp. 961-964, 2005.
9) T. Fukuda, O. Ichikawa, M. Nishimura, S. Rennie, V. Goel, “Regularized feature-space discriminative adaptation for robust ASR,” Proc. Interspeech, pp. 2185-2188, 2014.
10) 市川治，福田隆，立花隆輝，” 大規模音声データを異なる音響環境向けの音響モデル学習データに変換するオーディオマッピング技術,” 音講論（秋），2014.
11) T. Fukuda, R. Tachibana, D. Willett, and Z. Puming, "Ensembles of Dissimilar Acoustic Models Based on Big Data for Large Vocabulary Continuous Speech Recognition, " IEICE Trans. on Information and Systems, Vol.J98-D, No.8, pp.1162-1170, August 2015.
12) H. Soltau, B. Kingsbury, L. Mangu, D. Povey, G. Saon, and G. Zweig, “The IBM 2004 conversational telephony system for rich transcription,” Proc. ICASSP, pp. 205-208, 2005.
13) O. Ichikawa, T. Fukuda, and R. Tachibana, “Effective speech suppression using a two-channel microphone array for privacy protection in face-to-face sales monitoring,” Acoustical science and technology, 2015.
14) 鈴木雅之，倉田岳人，長野徹，立花隆輝，“発話衝突に頑健なモノラル録音された電話会話の音声認識モデル,” 音講論（秋），3-2-2, 2015.
15) X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, “Improving deep neural network acoustic models using generalized maxout networks,” Proc. ICASSP, pp. 215-219, 2014.
16) S. Rennie, V. Goel, and S. Thomas, “Annealed dropout training of deep networks,” Proc. SLT, 2014.
17) G. Kurata, and D. Willett, “Deep neural network training emphasizing central frames,” Proc. Interspeech, pp. 3595-3599, 2015.
18) S. F. Chen, “Shrinking exponential language models,” Proc. NAACL-HLT, pp. 468-476, 2009.
19) A. Emami and L. Mangu, “Empirical study of neural network language models for Arabic speech recognition,” Proc. ASRU, pp. 147-152, 2007.
20) N. Itoh, G. Kurata, R. Tachibana, M. Nishimura, “A Metric for Evaluating Speech Recognizer Output based on Human-Perception Model,” Proc Interspeech, pp. 1285-1288, 2015.
21) 長野徹，倉田岳人，鈴木雅之，立花隆輝，西村雅史，“大語彙連続音声認識と音節 N-best 音声認識を用いたキーワード検索の高精度化,” 情処論，Vol. 56, No. 8, pp. 1646-1656, 2015.
22) R. Tachibana, T. Fukuda, U. Chaudhari, B. Ramabhadran, P. Zhan, “Frame-level AnyBoost for LVCSR with the MMI criterion,” Proc. ASRU, pp. 12-17, 2011.
23) 長野徹，立花隆輝，西村雅史，「コーパスベース日本語音声合成フロントエンド」，電子情報通信学会論文誌 D Vol.J93-D No.10 pp.2096-2106 (2010-10).
24) R. Tachibana, T. Nagano, G. Kurata, M. Nishimura, N. Babaguchi, "Automatic Prosody Labeling using Multiple Models for Japanese," in IEICE Trans. on Information and Systems, 2007.
25) T. Nagano, R. Tachibana, N. Itoh, M. Nishimura, “Improving phoneme and accent estimation by leveraging a dictionary for a stochastic TTS front-end,” Proc. of ICASSP, 2008.
26) S. Shechtman, R. Tachibana, “Efficient Gradient F0 Tree Model for Prosody Modeling and Unit Selection Applied for the Embedded US English Concatenative TTS,” in Proc. of ICASSP, 2009.
27) R. Fernandez, Z. Kons, S. Shechtman, Z.-W. Shuang, R. Hoory, B.

Ramabhadran, and Y. Qin. The IBM “Submission to the 2008 Text-to-Speech Blizzard Challenge.”
28) J. F. Pitrelli, R. Bakis, E. M. Eide, R. Fernandez, W. Hamza, and M. A. Picheny., “The IBM Expressive Text-to-Speech Synthesis System for American English,” IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14(4) (July 2006): pp. 1099-1108.