進化を続ける音声認識エンジン「VoiceRex®」

コンタクトセンタ向けのAIを支える基盤技術の1つに音声認識があります．今や多くの場面で利用されるようになり，ビジネスをさまざまなかたちで支えています．その研究開発の歴史は長く，さまざまな技術を積み重ね，現在の水準に至っています．NTTメディアインテリジェンス研究所で開発する音声認識エンジン「VoiceRex®」の歴史を軸にこれらを紹介し，最新のエンジンに導入されているコンタクトセンタでの活用が期待される技術について紹介します．

大庭 隆伸（おおば たかのぶ）/田中 智大（たなか ともひろ）
増村 亮（ますむら りょう）
NTTメディアインテリジェンス研究所

音声認識エンジン「VoiceRex®」の歴史

コンタクトセンタ向けのAIを支える基盤技術の1つに音声認識技術があります．音声認識技術とは入力信号の中に含まれる音声をテキスト化する技術です．NTT研究所における音声認識の研究開発の歴史は長く，半世紀程度に及びます．NTTメディアインテリジェンス研究所では，こうした長年にわたる研究成果を基に，幅広いサービス分野に適用可能なものとして音声認識エンジン「VoiceRex®」を開発し，グループ各社に提供しています．

音声認識をコンタクトセンタ通話の分析に用いるというアイデアは研究開発の当初から存在しました．当時のレベルでは夢のまた夢でありましたが，何十年後かに到達すべき目標でもありました．VoiceRex®（の前身となる音声認識ライブラリ）は，1990年代に入って初めてリリースされましたが，当時のものはキーワードを認識することしかできないものでした．それから現在のように会話等の長い発話が認識可能になったのは2000年のことでした．とはいえ，その性能は人間どうしの会話を認識しようとしても到底正しく認識できないレベルでした．新聞のような書き言葉を，はきはきとした口調で読み上げたような音声でなければ正しく認識できず，認識可能な単語数も極めて限定的でした．

そこから「VoiceRex®」は何段階かの技術革新を経て，性能が飛躍的に向上してきたという経緯があります．まず2008年，日本で初めてのWFST（Weighted Finite State Transducer）と呼ばれる技術を採用しました．従来よりも約100倍の言葉を覚えられるようになり，およそ1000万語の中から最適な単語を選ぶことが可能になりました．この進化は2009年の衆議院での議会録作成システムへの採用につながりました．議会は1人がしゃべる一問一答形式であり，こういう場では90%程度の高い認識性能を達成することができました．本会議場や各委員会等で，人手で行う速記に代わるものとして存在を知らしめることになったのです．

その後も，音声データベースの拡張と整備の進展や，計算機能力の向上を背景とした大規模なデータベースを効果的に活用する技術の創出などがあり，音声認識の性能は徐々に向上を続けていきました．そして，コンタクトセンタの通話の音声認識はいよいよ現実の問題設定となり，2014年，NTTソフトウェアから「ForeSight Voice Mining」*というコンタクトセンタ向けの製品が発表されることになりました．

さて，その数年前，音声認識の研究者コミュニティの中では，ある技術が注目を集めていました．深層学習（ディープニューラルネットワーク）の登場です．深層学習は音声認識にとって大きなパラダイムシフトとなりました．音声信号，すなわち空気振動を「あいうえお」という音の並びに直す音響モデルの性能が極めて高まり，これにより通話の認識率も飛躍的に向上しました．深層学習を採用したVoiceRex®は2014年に商用リリースされています．さらに2015年にニューラルネットワークの一種CNN-NIN（Convolutional Neural Network – Network In Network）を採用し，騒がしい公共エリアでモバイル端末を使った音声認識を行う，「CHiME3」という技術評価国際イベントで参加研究機関中1位を獲得しました．スマートフォンの普及により，外で電話をする機会が増えましたが，そうした人混みなどの周囲雑音がある音声信号に対しても精度良く認識ができるようになりました．

＊ ForeSight Voice Miningは現在NTTテクノクロス株式会社より販売されています．

こうした技術革新を経て音声認識の用途は急速に拡大しました．今やVoiceRex®は多くの商品・サービスで利用されています．特にコンタクトセンタ向けAI関連の商品導入事例は急速に数を増やしており，音声認識を用いた商品・サービスの中でも中核的な存在になってきています．

一方，数多くのお客さまにご利用いただくことで，新たな課題も生まれてきています．その１つは話題の多様性です．現在の音声認識技術では，入力される通話の話題が既知であると精度面では有利に働きます．例えば，電話窓口を提供している企業ごとに取り扱っているサービス名は違っています．また，同じ会社の電話窓口でも入会・退会を受け付ける窓口，苦情を受け付ける窓口，技術的な質問に答える窓口などに分かれています．そのため，現状では会社ごと，もしくはコンタクトセンタの単位で，話題に関するモデル（言語モデル）を個別にチューニングする作業を行っています．

もう１つの課題は，話の流暢さへの対応です．人間どうしの会話では明確に発音しないという現象が頻発します．例えば，文末の「〜します」をアルファベットで書くと「shimasu」ですが，「shim」くらいまでは発音するが，残りの「asu」は音はあるもののはっきりとは発音されず，リズムだけ表現するといったことが頻発します．こうした音の不明瞭さ（場合によっては発音をしていない現象）に対しては，言葉の文脈的な観点から予測するような枠組みが必要になります．以降では，最新版VoiceRex®に搭載されている，これらの課題を克服するための新しい技術について紹介します．

会話コンテキスト言語モデル

言語モデルとは単語のつながりを予測するモデルです．直感的には文が日本語として正しいかどうかを判定するといった役割を果たします．音声認識エンジンには，ある単語は直前のN-１個の単語に依存すると考えてモデル化を行うN-gram言語モデルと呼ばれる確率モデルが利用されています．Nが大きいと単語の組み合わせが爆発的に増加することからNはたかだか３〜４程度です．そのため，局所的なコンテキスト（文脈情報）のみを考慮したモデル化となってしまっています．短い発話であればN-gram言語モデルでも十分ですが，発話が長くなればなるほど，より長いコンテキストを考慮することが必要になるのです．

そこで近年，ニューラルネットワークを用いた言語モデルが注目を集めており，特にリカレントニューラルネットワーク（RNN）言語モデルと呼ばれる長距離のコンテキストを扱うことが可能なモデルが注目を集めています．音声認識で用いる際は，一度音声認識をして得られた複数の音声認識結果候補の文に対して，RNN言語モデルで算出したスコアを加味することで，最終的な認識結果文を決定するという使い方をします．これをリスコアリング法と呼びます．

さて，RNN言語モデルの問題点の１つは発話単位のコンテキストの利用にとどまっている点です．コンタクトセンタの通話では，発話をまたぐコンテキストの利用は極めて重要です．例えば，オペレータがお客さまの問合せに対して回答することを考えると，その回答内容はお客さまの問合せ内容に関連する内容のはずです．そこで私たちはそのような発話をまたぐ長期的なコンテキストを考慮した，会話コンテキスト言語モデル技術[1]を開発しました．実際に音声認識する際には，逐次的に前の発話の音声認識結果をコンテキストとして利用します．VoiceRex®では，会話コンテキスト言語モデルを発話ごとにリスコアリング法により適用することで，より良い文（正解に近い文）を選び直しながら，それをコンテキストとして与え，次の発話の音声認識を行うことができる実装になっています．そのため，各発話の入力ごとに，会話コンテキスト言語モデルの恩恵が得られるのです．

ニューラル誤り訂正言語モデル

話の流暢さにより明瞭に発音されない音があることは述べましたが，そこには一定の傾向があります．前述のような「〜します」といった文末表現，助詞，「ありがとうございます」といった頻出表現などで多くみられ，それぞれで音声認識は毎回同じような誤りを起こしてしまいがちです．発音のあいまいさに起因する事象以外にも，頻出の誤りパターンがいくつか存在します．こうした誤りの偏りを見つけて，まるごと修正するアプローチを誤り訂正と呼びます．

私たちはRNN言語モデルに対して音声認識誤りを考慮できる枠組みを取り入れたニューラル誤り訂正言語モデル技術[2]を開発することで，認識精度の向上に成功しました．具体的にはEncoder-Decoderモデルと呼ばれるニューラルネットワークを導入し，音声認識誤りを含む音声認識結果から正解文を推定するといった機構を与えました．ニューラル誤り訂正言語モデルの学習には，一度音声認識して得られた音声認識誤りを含む文と，それに対応する正解文を用意して，前者と後者の関係を学習させます．これにより誤りの傾向とそれを修正する方法を同時に獲得できるのです．

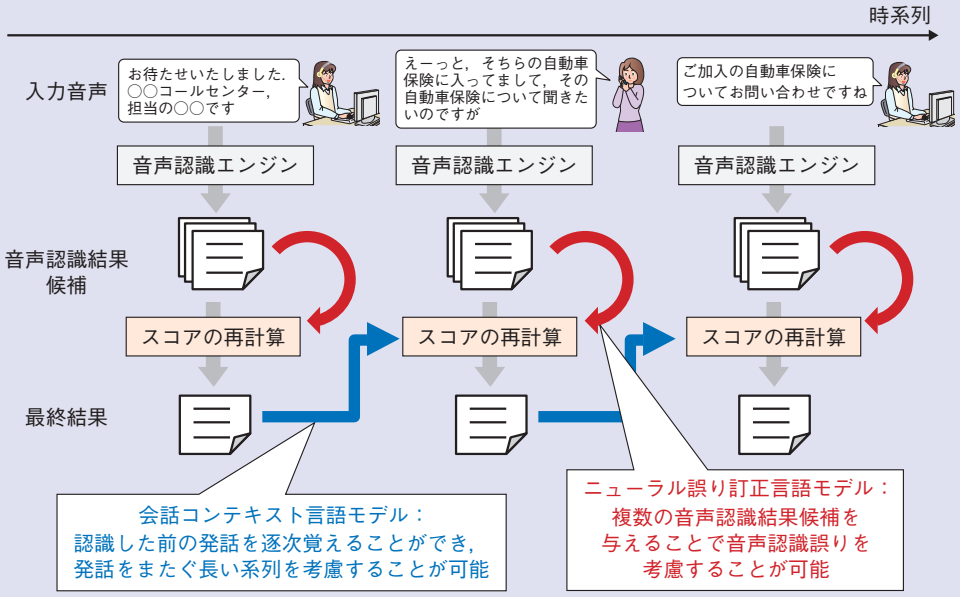


図　会話コンテキスト言語モデルとニューラル誤り訂正言語モデルを併用した音声認識

実際に音声認識を行う際には，会話コンテキスト言語モデルと同様にリスコアリング法を適用します．会話コンテキスト言語モデルと併用することが可能で，両言語モデルの恩恵を受けた認識結果を発話ごとに得ることができます．この様子を**図**に示します．通話の中で入力信号から発話区間が切り出されるたびに，認識結果候補を複数出力し，両言語モデルを適応して最終的な認識結果を求めます．会話コンテキスト言語モデルに関しては，その最終的な認識結果を次の発話を認識する際のコンテキストとすることを繰り返していくかたちをとっています．

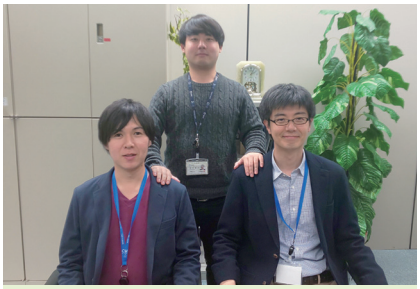
今後の展開

これまで対象としていたコンタクトセンタの多くは，お客さまのほうからコンタクトセンタに通電するという形式がほとんどでした．その場合，オペレータとお客さまは１度きりの会話をするので，お互いに比較的丁寧に話をしようとします．それは音声認識精度の観点でも都合の良い現象といえます．しかしながら，コンタクトセンタのAI関連製品の導入が進むにつれ，会社側からお客さまに連絡をするといったシーンにも用いられるようになってきました．この場合，お客さまごとに担当者がつき，複数回にわたって同じ人どうしで会話を行うため，よりフランクな会話が行われており，音声認識の観点では課題となっています．これまでも，音声認識は適用先を拡大にするのに伴い，従来よりも認識の難しい状況に直面してきました．そのたびに，それを打開するための研究開発を行うことで進化を続けてきたのです．今後も，実際に音声認識をご利用いただいている方々が直面している，新たな課題に挑戦しながら，VoiceRex®は進化を続けていくでしょう．

■参考文献

(1) R. Masumura, T. Tanaka, A. Ando, H. Masataki, and Y. Aono：“Role Play Dialogue Aware Language Models Based on Conditional Hierarchical Recurrent Encoder-Decoder,” Proc. of Interspeech 2018, pp.1259-1263, Hyderabad, India, Sept. 2018.

(2) T. Tanaka, R. Masumura, H. Masataki, and Y. Aono：“Neural Error Corrective Language Models for Automatic Speech Recognition,” Proc. of Interspeech 2018, pp.401-405, Hyderabad, India, Sept. 2018.



（左から）増村　亮/ 田中　智大/ 大庭　隆伸

音声認識は多くのビジネス案件で活用されており今後も拡大が期待されます．そこで，いくつかのグループ各社を通して，音声認識をお試しいただける環境を提供しています．ぜひ「NTT 音声認識」と検索してみてください．

◆問い合わせ先

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
TEL　046-859-2943
FAX　046-855-1054
E-mail　takanobu.oba.ec@hco.ntt.co.jp