ISSN 1883-1966(Print)
ISSN 1884-0817(Online)
国立情報学研究所ニュース

NII Today

National Institute of Informatics News

65
Sep. 2014

NII Interview
音声合成と音声認識の組み合わせが豊かな社会を作り上げる

NII Special 1
応用領域が急拡大する「統計的音声合成」技術

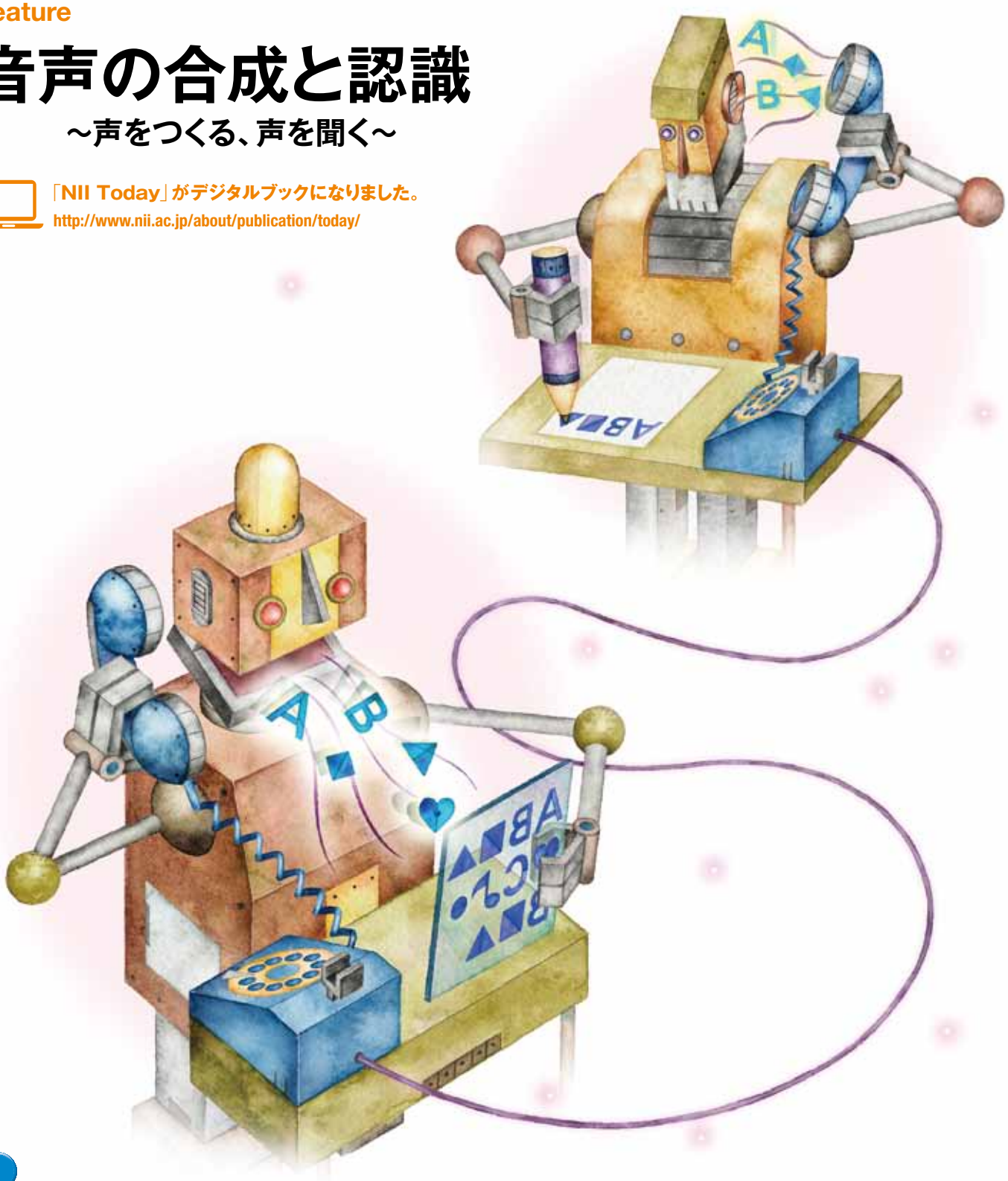
NII Special 2
実用化へ走り出した音声認識

Feature

音声の合成と認識

～声をつくる、声を聞く～

「NII Today」がデジタルブックになりました。
http://www.nii.ac.jp/about/publication/today/



国立情報学研究所
National Institute of Informatics

NII Interview

音声合成と音声認識の組み合わせが豊かな社会を作り上げる

スマートフォンやタブレットを音声で操作する人が増えてきた。また、スマホなどから発せられる音声を聞き、対話をしながら操作する人も増加してきた。前者は音声認識技術であり、後者は音声合成技術によるものだ。この2つの技術は近いようで遠い関係にあったが、ここ数年で急接近している。音声認識に役立つ音声分離などの研究に取り組むNIIの小野順貴准教授と、音声合成を研究するNIIの山岸順一准教授に、音声認識と音声合成の最前線を聞いた。

大河原　ここにきて、音声認識や音声合成が急速な勢いで関心を集めています。その理由はどこにあるとお考えですか。

小野　1つには、スマートフォンという音声入力に最適なデバイスが登場したことが見逃せません。口元にマイクがあることから、音声を正しく認識するデバイスとして非常に有効です。カーナビや家電のように、マイクから離れたところから音声認識をするのに比べて断然有利です。もう1つは、さまざまなデバイスを通じて数多くの音声データを収集できるようになったこと。ここ数年で音声認識の技術が一気に進展していることを強く感じますね。私はスマホでメールを書くのにも、音声入力を使う方が多いですよ。

山岸　現在の音声認識は、HMM（Hidden Markov Model：隠れマルコフモデル）の技術が主流となっています。これは統計的アプローチを用いた手法で、大量のデータ蓄積を行い、それをもとに音声認識のモデルを作り上げるというものです。近年のビッグデータの活用や、ディープラーニングといった新たな潮流がこの技術の進化を支えています。

小野　ただ、音声認識の技術進化には大量のデータが必要となるため、ビッグデータを持っている企業や組織が強くなり、そこにまたデータが集まるという循環が繰り返されます。他が参入しにくくなるという環境になってきていますね。

大河原　音声認識はもともと日本の技術が先行していたはずですが、いまは日本の企業や研究所がそういう立場にはありませんね。それはデータ量が重視されてきたことが原因なのでしょうか。

山岸　そういう側面は否めないといえます。しかし、その一方で、音声合成については、「お家芸」というほどに日本が先行しています。いま私が行っている研究は、声の大規模データベースを使い、ある動詞の中の音素はどんな周波数になるのか、読み上げの声と怒っている声はどんな周波数か、その中間の声はどうなるかといった関数をもとに、平均的な声を作りあげ、そこに個人ごとの声の差を示すデータを組み合わせ、わずか10分程度で、特定の人にそっくりの声を作れるというものです。筋萎縮性側索硬化症（ALS）やがんの手術などで声を失った人も、わずかな音声データがあるだけで、本人の音声を作り上げることができます。

大河原　一方で、ここにきて、音声認識と音声合成の2つの技術が強く結びつくようになってきましたね。

山岸　音声合成と音声認識の技術はまったく別の技術でした。しかし、それぞれの技術が進化し、いずれも統計的アプローチ（隠れマルコフモデル）となったことで、研究者がお互いのコミュニティを行き交うようになり、化学反応が起き始めています。私が研究している音声合成による平均声の開発プロセスも、音声認識の技術からきているものです。

小野　ただし、音声合成と音声認識で求められる技術的要素が異なるという議論もありますね。

山岸　かなりの技術が互いに使えるようになりましたが、細かい部分をみると違う要素が求められます。音声認識は意味がわかればいいので、細かいニュアンスは認識しなくていい。しかし音声合成は、細かいニュアンスまで再現しなくてはならない。同じ統計モデルでも、学習のさせ方や学習する粒度が違うのです。

大河原　音声認識と音声合成の技術を組み合わせると、どんなことができるようになりますか。

山岸　成果の1つに、音声翻訳システムがあげられます。音声認識したものを、機械翻訳し、音声合成をして、あらゆる言語に自動変換してしゃべらせることができます。しかも、自分と同じ声で発す

［予告］
山岸先生が
ボクの声をつくってくれてるよ！
楽しみに待っててね！

情報犬
ピットくん

ることができる。第2外国語を学習するときに、自分の声だとこう発音するべきといったこともわかるようになります。それを発展させると、映画に出演している俳優の声のまま、他の言語でしゃべらせることができます。

大河原　音声認識は人間の耳に、音声合成は人間の口に近づけることが目標となりますか。

小野　音声認識や音声合成は、人間の耳や口を再現しようとしているわけではありません。人間にできることができないことも、また人間にできないことができることもあります。例えば音声認識では、人間はかなり雑音があったり、話者がかなり離れていても音声を認識できますが、こうした状況は音声認識システムにはまだまだ難しい面があります。一方私は、音声認識の前処理として、複数のマイクで特定の音だけを抽出するといった研究を行っていますが、このように混ざった音から、きれいな音を抽出して相手に聞かせることは人間にはできません。

大河原　これからの課題について、お考えをお聞かせください。

小野　音声認識では、遠隔発話において、どこまで人間に近づけるかということですね。この研究が進化すれば、会議の内容を要約して、議事録を自動作成してくれるといったことが可能になります。ロボットが、複数の人がしゃべっていることを自然に理解すれば、SFのような世界がやってくるでしょうね。

山岸　音声合成においては、いかに表現力を発展させるかが課題です。いまの統計的アプローチでは、平均的な表現にしかならないため、ナレーションなどには適していますが、映画のワンシーンの俳優の声といった「声の芸術性」といえる部分には弱さがある。これを解決できないと、聞き手を30分間、1時間と飽きさせない表現ができません。一方で、声が自由に合成できるようなった時に、いかに声の詐称を防ぐかといったセキュアな土壌も作らなくてはなりません。そうしないと、必要な時に、必要な用途に自由に使える技術には発展しえないと考えています。そこが課題だといえます。

インタビュアーの一言

音声合成と音声認識は、数年前までは、「近くて遠い関係」だった。しかし、2つの技術が近づいたことで化学反応が起こり、技術進化を加速させている。その背景には、IT分野における重要な技術トレンドといわれるモバイル、クラウド、ビッグデータ、ソーシャル、アナリティクスが緊密に絡まっている点も見逃せない。2つの技術が我々の暮らしを豊かにすることを期待している。

大河原克行
Katsuyuki Ohkawara
ジャーナリスト

1965年、東京都出身。IT業界の専門紙の編集長を経て、2001年からフリーランスジャーナリストとして独立。25年以上にわたってIT産業を中心に幅広く取材、執筆活動を続ける。現在、ビジネス誌、パソコン誌、ウェブ媒体などで活躍。

小野順貴
Nobutaka Ono
国立情報学研究所
情報学プリンシプル研究系 准教授
総合研究大学院大学 複合科学研究科
情報学専攻 准教授

山岸順一
Junichi Yamagishi
国立情報学研究所
コンテンツ科学研究系 准教授
総合研究大学院大学
複合科学研究科 情報学専攻 准教授

NII Special 1 音声合成

応用領域が急拡大する「統計的音声合成」技術

統計的手法「HMM」による自然な音声合成とは？

いま、音声合成技術は、かつての「宇宙人の声」のような不自然なものから、普通の人間の発話と見分けがつかない高品質なものへと進化している。その背後には、統計的な手法を使った音声合成技術の進歩がある。従来よりも学習データ量、計算データ量ともに劇的に軽減したこともあり、デジタルサイネージやロボット、障碍者支援、携帯デバイスナビゲーションなど、応用領域を急速に広げつつあるのだ。発語機能を失った人の元の声の再現、オリジナル話者の声を使った翻訳の読み上げなど、音声合成研究の最前線について、世界のトップを走る３人の研究者に聞いた。

山岸順一
Junichi Yamagishi
国立情報学研究所
コンテンツ科学研究系 准教授
総合研究大学院大学
複合科学研究科 情報学専攻 准教授

音声合成技術を劇的に変えた「統計的方法」とは？

徳田　音声合成は1950年代から取り組まれてきましたが、現在につながる音声合成は1970年代から発展しました。この頃は、ある音のあとに次の音がつながるときにどのようなルールがあるのかを調べ、そのルールを音声合成に利用する方法でした。

山岸　人間は肺からの空気で声帯をふるわせて出る音を、口腔や鼻腔で共振させて発声します。舌や口の形で共振周波数を調整し、音色を加えて声にしています。昔の音声合成は、声帯から出てくる音がどう変化するか、共振周波数の変化ルールを導きだし、テキストに応じてルールをベースに音源に変調をかけるやり方で音声を合成していました。その音声は、古いSF映画に出てくる「コンピュータの声」や「宇宙人の声」のような、独特の響きを持っていました。

徳田　中には非常によい結果を出した研究もありましたが、ルール作りには研究者の個性が出てしまいます。また1人の声のルールは作れても、例えば男・女・若年・壮年などといったさまざまな属性の人のルールを作ろうとすると何十年もかかってしまうので、柔軟性に乏しいという欠点がありました。

そんな状況を変えたのが、この頃から急速に発展したコンピュータ技術です。従来よりも大量のデータを高速に処理できるようになったことから、大量の音声を録音し、「コーパス」と呼ばれるデータベースにして、そこから音声を拾って、いわば切り貼りするようにして自然な音声を合成できるようになりました（波形接続型音声合成）。こうすると、単音節以外にも単語や文節など、ある程度連続した音声のつながりを利用できるので、ずっと自然な合成音声になります（単位選択合成）。

このコーパスベースの音声合成は、1980年代にはコールセンターなどの電話自動応答システムやパソコンのテキス

徳田恵一
Keiichi Tokuda
国立情報学研究所 客員教授
名古屋工業大学
大学院工学研究科 教授

徳田恵一先生はロンドンからスカイプで参加

ト読み上げソフトに応用され始め、90年代には利用が広がって一大ブームとなりました。この頃の技術は現在も広く利用されています。いまネットで人気のボーカロイドの音声合成技術も80年代の音声合成技術をベースにしたものです。

山岸　しかしこの方式は、例えば数十～数百時間におよぶ音声データの収録を必要とし、大規模なデータベースを使わなければなりません。徳田先生と戸田先生、そして私は同じ研究機関で協働して100時間の音声収録を行いましたが、1年もの期間がかかりました。人間の声は体調によりさまざまに変化するので、利用できる音声を収録するのに10倍くらいの時間が必要になるからです。

徳田　より自然な音声にしようとすると、膨大な音声データが必要になり、インデックスづけなどの後処理も大変です。例えば「ドラえもん」や、人工知能と人間が恋をする映画「her／世界でひとつの彼女」のように、自然に会話ができるような機械を想定すると、無限の多様性を持つ音声合成が必要になり、データ収録も無限に必要になってしまいます。また、データベースサイズが大きくなると、携帯端末のような記憶容量も性能も低い端末での利用は難しくなります。

そこで、もっと柔軟で効率のよい音声合成の方法はないかと研究した結果、私たちが見出したのが「HMM（hidden markov model：隠れマルコフモデル）」という統計学の確率モデルでした。

HMMによる音声合成の仕組みと3つの利点

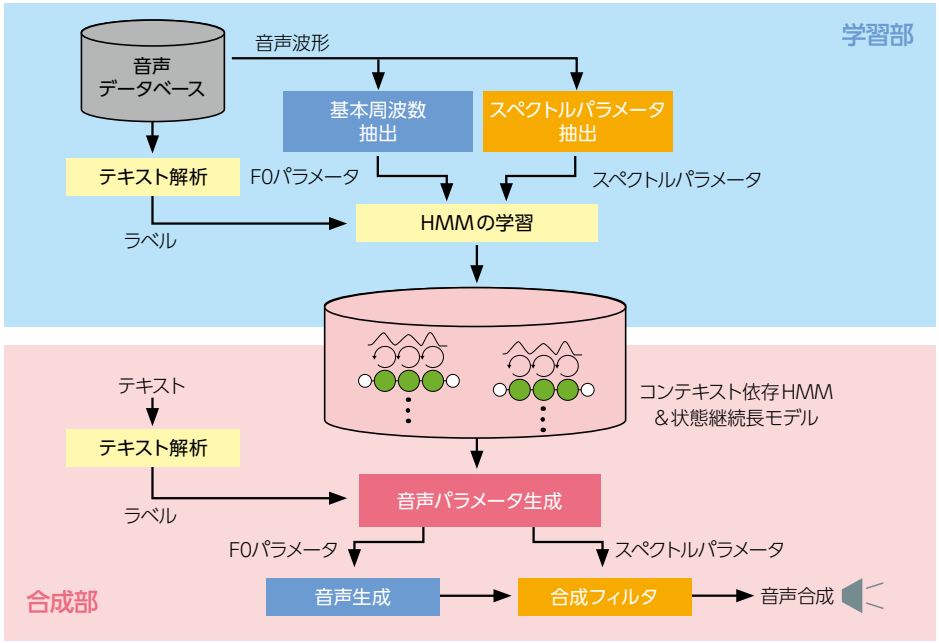
戸田　HMMによる音声合成は、コーパスに蓄積されたテキストと音声波形の対応関係を、関数としてコンピュータに学習させるところから始まります。これは音声認識で成果を挙げた手法で、抽象化された関数を利用することで、音声波形に揺らぎがあっても、その背後に潜んでいる特性を見出すことができます。例えば同じ「おはよう」という言葉でも、発声するたびに違った波形になるものですが、統計的な手法によれば共通のパターンが計算によって割り出され、テキストとの対応はどれも同じ「おはよう」になるのです。徳田先生が、世界で初めてHMMから直接音声を合成する技術を開発されました。

徳田　HMMには95年から取り組んでいます。仕組みを簡単に言えば、まず学習データから音声信号の基本周波数（声の強さや抑揚にあたるもの）とスペクトル※1（声道での共振にあたるもの）を抽出し、テキストとの対応をHMMによりモデル化（関数化）してモデルとします。読み上げるテキストを解析した結果をこのデータベースに照合して、統計的に最も正解に近い基本周波数パラメータで音源を生成し、同様に生成したスペクトルパラメータによって音色を合成するという方法です（**図1参照**）。

これには次の3つの利点があります。

※1 **スペクトル**　音に含まれる周波数成分を波長の順に並べた強度分布のこと。

図1：HMM音声合成の枠組み



戸田智基
Tomoki Toda
国立情報学研究所 客員准教授
奈良先端科学技術大学院大学
情報科学研究科 准教授

障碍者支援に貢献する音声合成

●省メモリでモバイルデバイスにも音声合成システムが搭載可能に

自然な音声合成に必要なデータ量は、わずか1～2MB程度と劇的に軽減できます。デジタルサイネージやモバイルデバイス、マスコットロボットなどでも、その端末内で音声合成を簡単に行うことができます（**写真1**）。

●言語依存性がなく、多言語適用が容易。歌声合成にも利用可能

言語依存性がほとんどないので、ある言語で開発した音声合成ソフトはほとんどそのまま他の言語でも利用できます。現在40カ国語以上の言語に適用されています。

この柔軟性は、「音声」を「歌声」に、「テキスト」を「歌詞付き楽譜」に置き換えれば、同じ仕組みで歌声合成も可能にします。Aさんがいくつかの曲を歌って機械に学習させると、それらとは別の曲の歌詞付き楽譜を機械に入力するだけで、機械がAさんの歌声で歌ってくれます。また、自分の声によるボーカルアシストも可能です。この技術は「CeVIO Creative Studio」というフリー/有償ソフトとして一般向けに提供されています（**写真2**）。

●パラメータ調整で音声を「真似る」「混ぜる」「作り出す」ことが可能

パラメータの調整だけでいろいろな声が作り出せることも特長です。感情表現をつけたり、別の人の声を真似たり、複数の人の声を混ぜ合わせたり、実際にはない音声を作り出したりすることができます。

戸田　いろいろな声をパラメータ変更だけで作り出せる技術には、エンターティンメント以外に医療・福祉分野で大きな期待が寄せられています。ここまでのお話はテキストを機械で読みあげることが前提ですが、類似した技術によれば、例えばAさんの話をマイクで拾い、リアルタイムでBさんの声に変換してスピー

教職員・学生が
自由に登録！

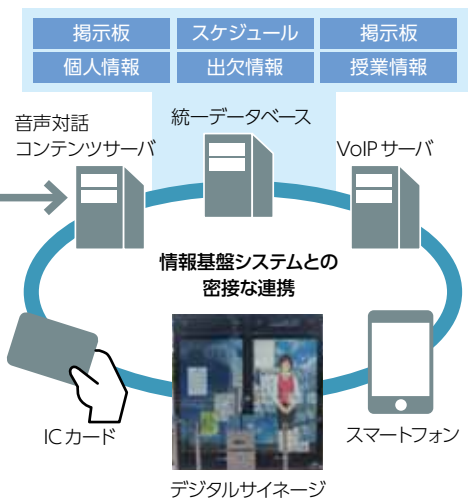


写真1：名古屋工業大学のキャンパスナビゲーション用に作られたデジタルサイネージ。音声認識 / 音声合成 / CG システムを搭載し、CG キャラクターが、訪問者と音声でインタラクティブに会話する。このシステムに使われた音声合成システム「HTS」と音声認識エンジン「Julius」は、「MMDAgent」という名前でオープンソース・ソフトウェアとして公開されている。

カーから流すことができます。これを応用すれば、外国の方が発声した外国語を日本語の話者の音声に変換して、その話者自身の声色で出力することもできるわけです。また、同様にして、そのままでは聴き取りにくい音声を自然で明瞭な音声に変換することができるようにもなります。

例えば、喉頭がんなどで声帯を切除した人は、食道の入口をうまく振動させて発声する「食道発声」や電気式人工喉頭とよばれる補助器具を用いた発声を練習することが多いのですが、これは通常の音声と比べるとはかなり不自然で聴き取りにくいものとなってしまいます。その際、健康なときのその人の声のサンプルがあれば、音声変換装置を通して元のその人自身の声に近い自然な話し方に変換できるというわけです。

山岸　サンプル音声が少なくても、統計的な音声合成技術を使えば、かなり自然にその人の声に似せられるところがポイントですね。10分程度の録音データがあれば、その人の声による音声合成が可能です。また複数の人の声から「平均声」を作成することも簡単です。

これは構音障碍者の支援に大きな役割を果たします。筋萎縮性側索硬化症（ALS）など構音障碍が急速に進行する病気の人は、自分の声で意思伝達できなくなっていくことに苦しみ、周囲の人も話が聴き取れないことに悩むことが多いのです。そこで元気な頃の音声録音データを用いて音声合成を行う会話支援器を利用すると、聞き取りにくくなっていく一方の音声を明瞭な音声に補正して出力することができます。録音データは数分程度で大丈夫です。

あるスコットランドのALS患者の場合は、近隣の20名の方々が協力して音声を録音し、その平均パラメータにより本人の発音に近い違和感のない音声を合成することに成功しました。方言が強い地方だけに、一般的な平均声では満足できなかったのですが、この方法だとこれまで自分が話してきた方言混じりの話し方が再現でき、アイデンティティが取り戻せたと喜んでいました。

誰もが、できるだけ自分の声で話したいのです。音声合成技術を用いた会話支援器は低コストで入手できるようになりますが、それを障碍者1人ひとりが生かして使うには、できるだけ広い範囲で、たくさんの人の音声データの集積をしておくことが肝心です。世界で音声データを収集して利用可能にする「ボイスバンク」プロジェクトが行われており、日本ではNIIで「日本語ボイスバンクプロジェクト」を推進しています※2。

写真2：HMMによる音声合成技術を利用した「歌声合成」ソフト「CeVIO Creative Studio」。キャラクターボイスは「さとうささら」などのパッケージが提供されている。テキスト音声の合成、歌声の合成に、「元気」「怒り」「哀しみ」などの感情パラメータを加えたり、タイミングやピッチ、音量を変化させて声質、声色を変化させることができる。（左：トークトラックの編集画面例、右：ソングトラックの編集画面例）

画像提供：CeVIOプロジェクト（販売元）株式会社フロンティアワークス　（イラスト）斎藤将嗣 http://cevio.jp/others/CCS/

明瞭性は「人間以上」音声合成が拓く未来

徳田　戸田先生は音声入力ができる人向けの支援技術を、山岸先生は発音が難しい人のための支援技術を研究・開発しておられるわけですね。そんな研究が音声合成の応用領域をどんどん広げてくれています。人工知能と人間が自然に対話する時代はすぐそこまで来ています。耳障りな声でなく、生身の人間同様に気軽に機械とおしゃべりできるような技術開発を加速していくつもりです。

山岸　HMMを使った音声合成は、世界の音声合成研究者が合成音声の自然さを評価するリスニングコンテストで「人間と同等の明瞭性」を持つと世界で初めて認められました。また、「騒音下では人間の声よりも明瞭に聴き取れる」という評価も得られました。ある意味では人間の声よりも高品質の声を手に入れたことになります。

（取材・文＝土肥正弘）

※2　ボイスバンクプロジェクト　音声の障碍患者の生活の質を向上させることを目的に、本人以外の参加者の声を収集する取り組み。声のデータを混ぜ合わせてテンプレートとして利用することで、本人の声による音声合成システムを容易に、素早く構築できる。http://www.nii.ac.jp/research/voicebank/

下記をクリックして動画・音声ファイルをご覧ください。

動画　自分の声でコミュニケーションする音声情報処理 ― 山岸順一 准教授
http://www.yourepeat.com/watch/?v=CSPP_z0GfzQ

動画　【MMDAgent】初音ミクとおしゃべりできるソフトをつくってみた ― 徳田恵一 教授
https://www.youtube.com/watch?v=hGiDMVakggE

音声　統計的手法に基づく音声変換を用いた音声生成機能の拡張 ― 戸田智基 准教授
http://isw3.naist.jp/~tomoki/NII/DemoVC_Toda@NAIST.pptx
※PPTファイルがダウンロードされます。

NII Special 2 音声認識

実用化へ走り出した音声認識

母国語のようなスムーズな会話の実現へ

現在、コンピュータ技術の進展や膨大な音声データの集積などに伴い、音声認識の実用化に向けた研究が加速している。一方で、本格的な実利用が始まり、期待が高まる中で、いくつかの課題も見えてきた。音声認識技術の進化の歴史と実用化に向けた取り組み、そして現状の課題について、音声認識研究の専門家である京都大学の河原達也教授と独立行政法人 情報通信研究機構（以下NICT）ユニバーサルコミュニケーション研究所 音声コミュニケーション研究室の堀智織室長に、NIIで音の信号処理等の研究を手掛ける小野順貴准教授が話を伺った。

クラウド化により進化する音声認識技術

小野　まずは音声認識技術の歴史からお聞かせください。

河原　研究が開始されたのは、今から50年以上前に遡ります。海外ではベル研究所などいくつかの機関で研究が行われていましたが、日本でも先駆的に取り組まれており、1962年に京都大学が「音声タイプライター」を開発しました。これは、「あ・お・い」といった単音節を認識するものでした。その後、現代の音声認識システムの元となる技術ができたのが1990年頃です。それは、スペクトル包絡※を表現する特徴量と統計的分布の状態遷移モデル（HMM：Hidden Markov Model、隠れマルコフモデル）に基づくもので、以降約20年が経過しましたが、音声認識の基本的な枠組みはほとんど変わっていません。

※スペクトル包絡：音声の特徴量の中でもっとも重要な、スペクトルのなだらかな変動。

小野　音声認識の具体的なしくみについてご説明いただけますでしょうか。

河原　音声認識に必要な主な要素は、音響モデルと単語辞書・言語モデルです（図1）。音響モデルは日本語の各音素の周波数パターンを記憶したもの、言語モデルは日本語の単語の典型的な並びを記憶したものです。

堀　このように音声認識は複数の技術を組み合わせ、音声を文字データとして抽出します。例えば、音の並びを単語に置き換える際に単語の辞書だけでは十分な推定ができないため、音の並びから単語の候補を出し、さらに言語モデルに基づき単語列のもっともらしさも考慮して、確率的にもっとも正しいと推定される候補を選択するといった具合です。

小野　そうした中で、近年、技術的なブレイクスルーがあったわけですね。

河原　ええ、音響モデルなどの統計モデルの洗練や学習データの大規模化、そして計算機の処理能力の大幅な向上です。コンピュータの小型化、高性能化に続いて、現在ではスマートフォンなど携帯端末も高性能化しています。その一方で、ネットワークが高速化したことで、クラウドサーバ型のシステムが実現されました。こうした超大規模なサーバ、およびデータの活用により、現在の音声認識は端末側ではなく、バックグラウンドのサーバ側で行われ、これまでになかったような高精度な処理が実現されつつあります。

堀　ビッグデータは1つのキーワードとなっています。実世界で利用できる音声認識の実現にあたって、大規模な語彙を格納したデータベースが不可欠ですが、現在ではWeb上にテキストや音声の膨大なデータが存在しています。そうした新たな大規模な情報を利用しつつ、世界中の動画や音声に対して字幕化や検索のためのインデックス化、さらに翻訳する技術が各研究機関や組織で研究開発されています。

さまざまな場面での応用が進む

小野　進化を続ける音声認識技術ですが、現在、どこまで応用が進んでいるのでしょうか。

河原　主要なアプリケーションとして、機械に発話することで何らかの作業を行わせる「音声インタフェース」と、人間同士の自然な会話を認識し、自動的に記録や字幕を作るといった「音声をコンテンツとして扱うもの」の2つに大別されます。前者では、10年ほど前から音声タイプ／音声入力ワープロなど、パソコンのディクテーションソフトや、カーナビなど音声によるコマンド入力で実用化されてきました。また、電話や携帯電話による予約、問い合わせといった音声による情報アクセスでも活用されています。近年では、クラウドサーバによる音声認識の性能向上、およびスマートフォンの高性能化と普及に伴い、携帯端末における音声入力のニーズがより高まりつつあります（図2）。一方、後者の「音声コンテンツ」では、テレビ放送の字幕付与や議会の会議録作成などにおいて有効活用されています。

図 1：音声認識の仕組み。Wは単語列、 Pは音素列、X は音響特徴量を表している。

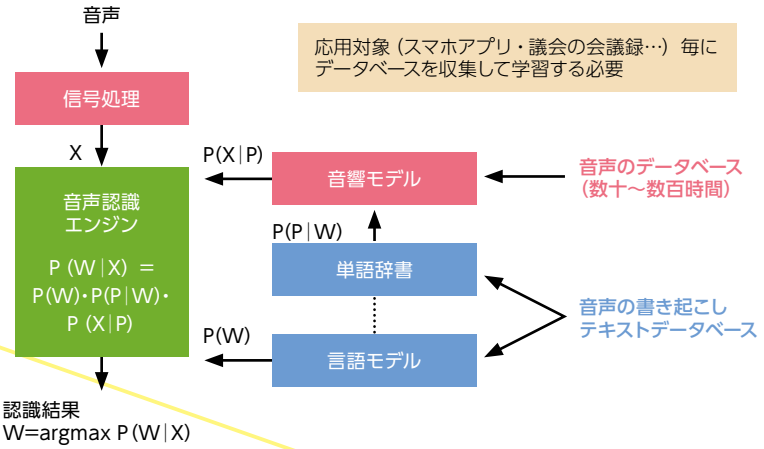
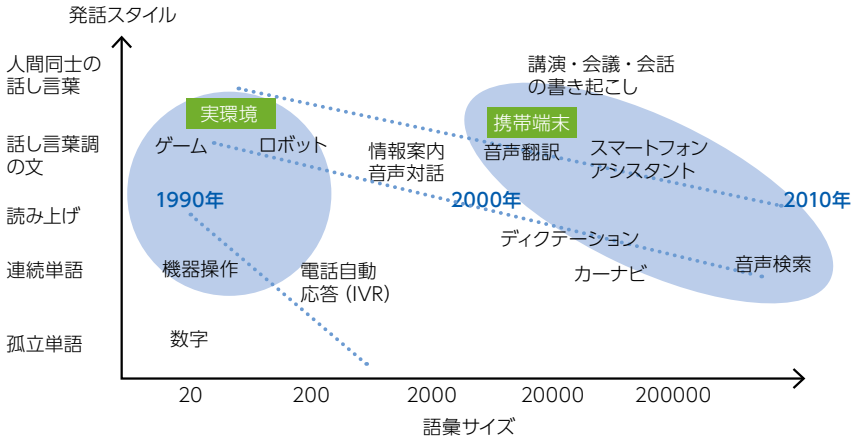


図 2：音声認識技術を活用したアプリケーションの例



堀　私が所属するNICT 音声コミュニケーション研究室では、主に2つの研究を進めています。1つが、人間と人間、人間と機械の簡便なコミュニケーションを実現するための音声インタフェースの研究、もう1つがWEB上の音声データに対して字幕付与、検索のための自動インデキシング技術の研究があります。前者の研究では、音声翻訳技術を用いて母語の異なるユーザー間のコミュニケーションを実現する音声翻訳システムの開発や、音声認識や合成技術を用いて聴覚障碍者の方と健常者の方のコミュニケーションを支援するシステムの開発を進めています。後者の研究では、多言語のニュース音声を認識して日本語に翻訳する、自動インデキシングを行うことで検索語に基づき音声を検索する、音声以外の音響イベントを抽出する、などの研究開発を進めております。

NICTで行っている多言語音声言語処理の研究を日本だけで実現するのは難しいことから、世界23カ国28研究機関が参加する研究共同体「U-STAR（Universal Speech Translation Advanced Research：ユニバーサル音声翻訳先端研究）コンソーシアム」により、世界規模の音声翻訳研究ネットワークの実現、

聞き手
小野順貴
Nobutaka Ono
国立情報学研究所
情報学プリンシプル研究系 准教授
総合研究大学院大学 複合科学研究科
情報学専攻 准教授

河原達也
Tatsuya Kawahara
京都大学
学術情報メディアセンター／情報学研究科
教授

そして音声認識を活用した多言語でのコミュニケーションの実用化を目指しています。

話し言葉を認識することの難しさ

小野　さらなる実用化に向けた課題には、どのようなものがあるのでしょうか。

河原　現在の多くの音声認識システムは、機械を意識して事前に内容を考え、簡単な文章を丁寧・明瞭に発声することを想定したものがほとんどです。そうしたケースであれば認識率も90%には達していますが、人間同士の会話に目を向けた場合はどうでしょうか。講演や議会のように公共の場で話す状況、それもスタジオやヘッドセットマイクなどで収録したものについては、かなり高い精度で音声を認識できるようになっているものの、家の中や街中での雑音が多い環境での会話や、日常会話のように発話のバリエーションが多様なものは、まだまだ精度を高めることが困難です。

堀　母国語の人間同士が行っているコミュニケーションのように、考えながら話したり、発声が明瞭でなくとも認識できるようにすることが課題となっています。

河原　実際に今の音声認識技術は、人間の言語処理とは異なり、意味を理解しているわけではありません。また、音声認識は先述したように、モデルを構築していくためのデータの収集が鍵となります。しかし、会話のデータはバリエーションが数多く存在しており、単にそれらを数多く収集して蓄積すれば、解決するかというとそれほど単純ではありません。多様なデータを的確に扱える技術を実現してくことが当面の課題の1つでしょう。

堀　一方で、いかに少ない学習データで大量の学習データと同等の性能を実現していくかが問題となるので、少資源で高精度化できる技術を創出してくことも重要です。また、これまではコンピュータの性能の制約から表層的な部分だけを扱わざるを得ませんでしたが、現在では大規模コンピュータによる膨大な計算が可能になってきているので、膨大な情報を取り込んでより高精度なモデルを学習することも不可欠です。

河原　音声認識を母国語での会話レベルにまで高めていくためには、今よりもう1段も2段も進化させていかなければなりません。そのためにも、統計的な学習理論をさらに追究していくことが必要です。

堀　音声認識は使用してもらえなければ進化できないので、さらにユーザーの実生活とリンクさせていく必要があります。今後は、音声認識技術を世界に広げていくことで、一般のユーザーからのさまざまなフィードバックを収集し、新しい課題の発見や、その解消に向けた研究を行っていく必要があります。そのためにもすべての関連機関が協力しあって技術を育成し、発展させていくことが重要でしょう。

（取材・文＝伊藤秀樹）

堀 智織
Chiori Hori
独立行政法人 情報通信研究機構（NICT）
ユニバーサルコミュニケーション研究所
音声コミュニケーション研究室 室長

河原達也先生の研究

国会会議録で活躍する音声認識技術

議会では長らく手書きの速記によって会議録が作成されてきた。しかし、速記者の新規養成の廃止に伴い、2011年に衆議院では、河原達也教授らの音声認識技術を用いたシステムを導入。同システムにより、すべての本会議・委員会の審議で、発言者のマイクから収録される音声に対して音声認識が行われ、会議録の草稿が生成されている。国会の審議音声を直接認識するシステムは、世界でも初めての事例だ。

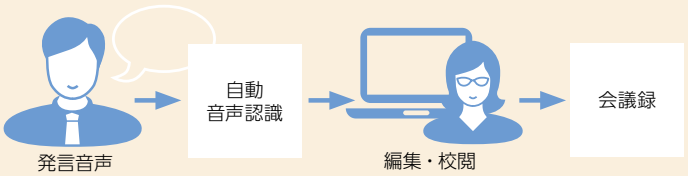
そのしくみについて、河原教授は以下のように語る。

「まず衆議院の審議音声と忠実な書き起こし（実際の発言内容）からなるデータベース（＝コーパス）を構築しました。会議録の文章との違いを統計的に分析、モデル化した結果、「えー」や「ですね」といった冗長語の削除を中心として、約13%の単語で違いが見られました。この統計モデルに基づいて、過去10年以上分、約2億単語にもおよぶ大量の会議録テキストから、実際の発言内容を予測する言語モデルを構築しました」

さらに、この言語モデルと音声を照合することで、約500時間分の審議音声から音響モデルを構築。これらのモデルは半自動的に追加学習や更新を可能とするもので、今後の総選挙や内閣改造に際しても話者集合の変化を反映し、持続的に性能を改善できるという。本格導入に先駆け2010年の試験導入で性能評価を行ったところ、会議録と照合した音声認識結果の文字正解率は89%に到達。この音声認識結果を速記者が専用エディタで修正・編集することで会議録原稿を作成するシステムの有用性が検証され、本格的なシステム運用が開始された。

「2011年に行われた118会議で評価したところ、平均文字正解率は89.8%となり85%を下回る会議はほとんどありませんでした。本会議に限れば、ほぼ95%にも達しています。しかし満足はしていません。もう一段性能を上げることで、その他の応用にもつなげてきたいと考えています」（河原教授）

衆議院の会議録作成システム



音声認識を活用した会議録作成システムのイメージ。2011年度以降、衆議院のすべての本会議・委員会の審議を処理している。国会審議を直接音声認識するシステムの導入・運用は世界初。

堀 智織先生の研究

言葉の壁を打ち破る音声翻訳アプリの開発

小野　現在、NICTでは、世界23か国、28の研究機関と連携した研究共同体「ユニバーサル音声翻訳先端研究コンソーシアム」（U-STAR）において、国際連携による自動音声翻訳システムの研究開発を進められています。そして2012年に、U-STARでは、世界人口の約95%をカバーする「多言語音声翻訳システム」が開発されましたね。堀室長はその開発を手掛けていらっしゃいますが、取り組みについてお聞かせください。

堀　U-STARでは、各国の研究機関との連携により、現在、音声入力17言語、テキスト入力27言語、音声出力14言語に対応した多言語音声翻訳システムを開発しており、本システムは2010年にNICTが国際標準化（ITU-T勧告書F.745およびH.625に準拠）した技術を採用しています。NICTが運用するコントロールサーバと、加盟する各研究機関がそれぞれ運用する音声認識、機械翻訳、音声合成のサーバをネットワーク型音声翻訳通信プロトコルで相互接続するとともに、利用者にはクライアントアプリケーションを介して、多言語の音声翻訳サービスを提供するというものです。

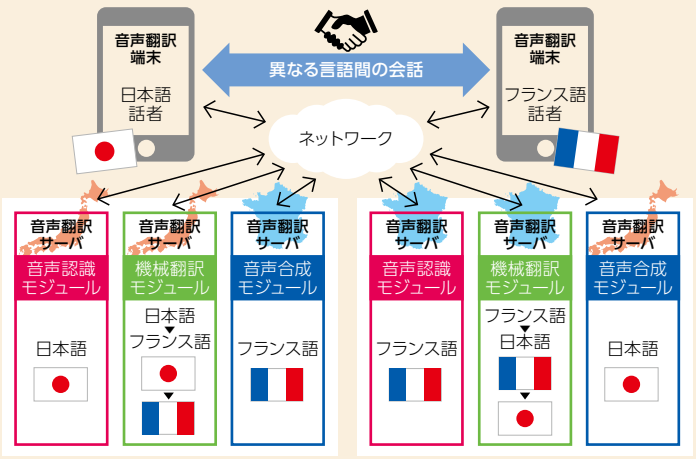
小野　一般の人も利用できるのですか？

堀　ええ、実証実験を行うためにiPhone向けの音声翻訳アプリケーション「VoiceTra4U」が一般公開されています。VoiceTra4Uは、発話された音声を認識し、簡単な操作で会話内容を翻訳するというものです。対応言語は、アジア、ヨーロッパの30種類以上をカバーしています。また、1台で音声翻訳し、翻訳結果を確認できるだけでなく、最大5台の端末を接続し、リアルタイムで会話ができる「チャットモード」も搭載しています。話した内容が、話し相手のそれぞれの言語に翻訳される、というわけです。

小野　これは使えますね。今後の展開をお聞かせください。

堀　VoiceTra4Uは言語だけではなく、視覚障碍を持った方との音声による会話や、聴覚障碍を持った方とのテキストによる会話も可能です。言語の壁だけでなく、コミュニケーションモダリティの壁も越えることも目指しています。

U-STARによるネットワーク型音声翻訳の概要図



NII essay

声が伝わる、声で伝える？

津崎 実
Minoru Tsuzaki
京都市立芸術大学 音楽学部 教授

今回の特集テーマは「音声合成と音声認識」ということであるが、私の研究の核は聴覚なので、私自身は音声よりはもう少し初期的なところが常に気になっている。一般的に音声と言えば、人間の声がまず思い浮かべられ、その声が動物の鳴き声と異なっている特徴は言葉を伝えていることと考えられる。

言葉とは言語内容であり、言語内容は意味や意志を伝えることは言うまでもない。だから選挙活動中の候補者が、「皆さまの一人ひとりの声を、国会に届けるためにも、是非私に1票をお願いします！」と言ったときには、別にその候補者は音声信号を国会で再生することを約束しているのではなく、みんなの意見を代表して国会で活動をすることを約束しているのである。この言語情報と音声信号とのリンクを機械にも可能とするのが、音声認識や音声合成の技術の大目的である。その技術はきっと便利であり、いろいろな局面で有効活用ができそうであり、活用されつつある。

しかし、ここでひとつ忘れてならない音声の側面があると思う。その側面はより根源的であり、誰もが無意識に享受しているがために、言語情報の側面に隠れてしまいがちである。その側面とはパーソナリティーである。自然環境では、声には必ずその主がいる。これは決して声のおまけではない。person、即ちペルソナは、語源としてper (through、by way of ; ～によって) と sona（音）であるという説がある。ペルソナに対して「仮面」という意味が英和辞典を引くと出てくるが、これはギリシャ時代の仮面劇では、役者の特定が声によってしかできなかったという側面と、さらに達者な役者はつけた仮面によって、その声までも自在に変えていたということに由来するものである。つまり、個人の特定に声を用いる戦略は昔から存在する。そもそも動物が声を出し始めたのも、個体識別を暗闇などでも可能とするための戦略だったと考えられる。

音声認識と合成ができる機械が、生活空間に人間と親和性高く同居するためには、このペルソナの側面を上手にハンドリングする必要があるかもしれない。人間らしくない機械が妙に人間くさく話すのはユーザーを混乱させかねず、また世界一優秀な音声合成器がすべて同じ声であちこちで公共の案内を話し出すというのも困りものである。コンシェルジュ・ロボットが、いくらこちらが言っている内容をしっかりと認識できても、「ところでどちら様でしたっけ？」という態度だったら腹が立つであろう。

巷ではボーカロイドという仮想的な存在が一定のファンを獲得しており、何を隠そう筆者もそのファンの一人であるが、彼女（？）たちが「不気味の谷」に落ちることなく、受け容れられている最大の要因は、実は声から入っていったことでペルソナがそこに存在していたからかもしれない。

編集後記

今回の特集は、NIIの前副所長で、長く本誌の編集長を務められた故・東倉洋一先生のご専門でもあった音声について取り上げました。東倉先生のご業績は周知の通りですが、なかでも最先端の音声研究を牽引してきたATR研究所の設立と後進の育成に努められたことは、音声研究の発展を大いに促しました。いまや音声合成・認識技術は、暮らしに欠かせない主要技術の１つになっています。改めて東倉先生が遺された偉大な足跡に思いを馳せるとともに、心よりご冥福をお祈りいたします。

情報から知を紡ぎだす。
NII
国立情報学研究所ニュース [NII Today] 第65号 平成26年9月
発行：大学共同利用機関法人 情報・システム研究機構 国立情報学研究所
〒101-8430 東京都千代田区一ツ橋２丁目１番２号 学術総合センター
発行人：喜連川 優　監修：佐藤一郎
表紙画：城谷俊也　写真撮影：川本聖哉 / 佐藤祐介
デスク：田井中麻都佳　制作：ノーバジェット株式会社
本誌についてのお問い合わせ：総務部企画課 広報チーム
TEL：03-4212-2164　FAX：03-4212-2150　e-mail：kouhou@nii.ac.jp

情報犬ビットくん
(NIIキャラクター)

http://www.nii.ac.jp/about/publication/today/