山岸 順一 Yamagishi Junichi

国立情報学研究所 コンテンツ科学研究系 教授
2018年～24年 CREST研究代表者
2024年よりAIP加速課題代表研究者

特集1 OVERVIEW

他者による音声・動画でのなりすましを防止
セキュリティーやプライバシー保護も実現

近年、あたかも本物であるかのように生成された偽の音声・画像・映像などの「ディープフェイク」が、サイバー犯罪に利用されるケースが散見されるようになった。このような他者によるなりすましを防止し、デジタル技術を健全に発展させるため、国立情報学研究所コンテンツ科学研究系の山岸順一教授は音声合成の領域をはじめとして、セキュリティーや個人のプライバシー保護の実現も目指してさまざまな研究開発に取り組んでいる。

大きく進化した音声合成技術 犯罪に悪用されるケースも

コンピューターを使ってテキストから音声を合成する技術は、時代とともに大きく進化している。現在では各種機器への組み込みなど、多分野での活用が広がっている。中でも「その人らしさ」につながる話者性を再現する音声合成モデル技術は、コンピューターの性能向上とAI（人工知能）による機械学習の進化も追い風となって急速に発展。少量の音声データで、本人そっくりの音声を作れるようになった。近年では、自分の声や既存の音声を用いてAIによる音声クローンを生成できる一般向けのサービスも多くの人に利用されている。

一方で、精巧に再現された人間の音声が悪用される懸念も高まっている。本人になりすまして音声認証システムを突破したり、特殊詐欺に利用したりするといったケースはその一例だ。2023年、米連邦取引委員会（FTC）は「親族を装った電話による詐欺事件の中には、人工音声が用いられた疑いが強いものもある」と警鐘を鳴らしている。事実、この数年の間に、音声だけでなく画像や動画のクローンもAIで大量生成できるようになったが、悪意のあるディープフェイクが社会問題になりつつある（図1）。

また、音声合成技術の進化に派生し、インターネット上に公開された音声から個人を特定することもできてしまう。そうした状況に対して、個人のプライバシーを守るために、音声公開前に話者情報を事前に加工するような「話者匿名化」の技術が求められている。これらの「声」に関するさまざまな問題を解決するとともに、音声合成技術の可能性をさらに広げるための研究開発に取り組んでいるのが、国立情報学研究所コンテンツ科学研究系の山岸順一教授だ。

「敵対的競争型研究」を採用 相反する技術で開発を加速

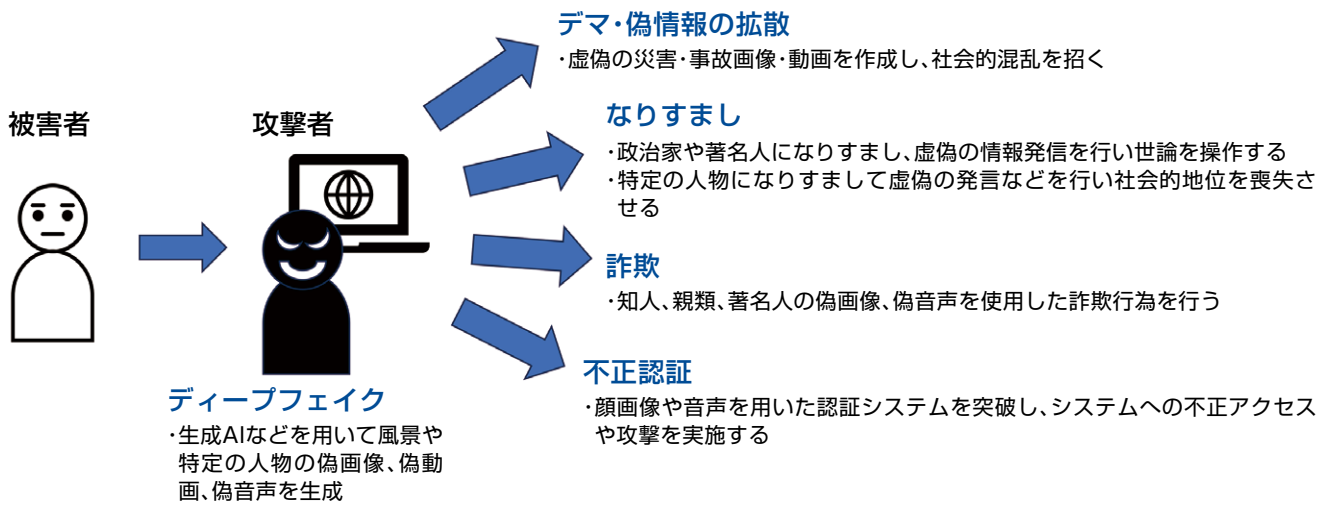
大学時代、学業とバンド活動にいそしんでいた山岸さんは、演奏を通じて「音」に対する興味関心を深めていったという。音と情報処理をテーマとした研究を志すようになり、東京工業大学（現・東京科学大学）博士課程修了後の2006年に渡英し、音声技術に関する研究で長い歴史と実績を持つエジンバラ大学へと研究の場を移す。「そこで音声合成に関する研究開発に取り組むとともに、多くの海外の研究者たちとの個人的なネットワークも構築することができました」と振り返る。

日本に帰国後も音声合成に関する研究開発を続ける中、2018年に米IT企業が肉声との差がほとんど生じないニューラルネットワークによるテキスト音声合成方式を発表した。山岸さんらも、この方式に対抗すべくニューラルネットワークと信号処理を融合した音声合成方式「ニューラル・ソースフィルター・モデル（NSF法）」を発表した。「この時点で音声合成と肉声の差は十分に小さく、実用に耐えうる方式が確立できました」と山岸さんは語る。

この時、音声合成技術が1つの節目を迎えたものの、次の研究の方向性を決める岐路に立たされたという。「その模索を続ける中で、音声合成の進化とともに今後は『悪用されるリスク』も増大するのではないかと思い至ったのです」。そこで、音声のセキュリティー対策とプライバシーの保護を強化する新技術創出を目的にJSTのCRESTで研究を開始した。「これまで、声のアイデンティティーに関する分野間の壁を取り除くと同時に、話者性のモデル化技術を高精度化することで、音声による生体認証の安全性と強度を高めるさまざまな研究開発を進めてきました」。

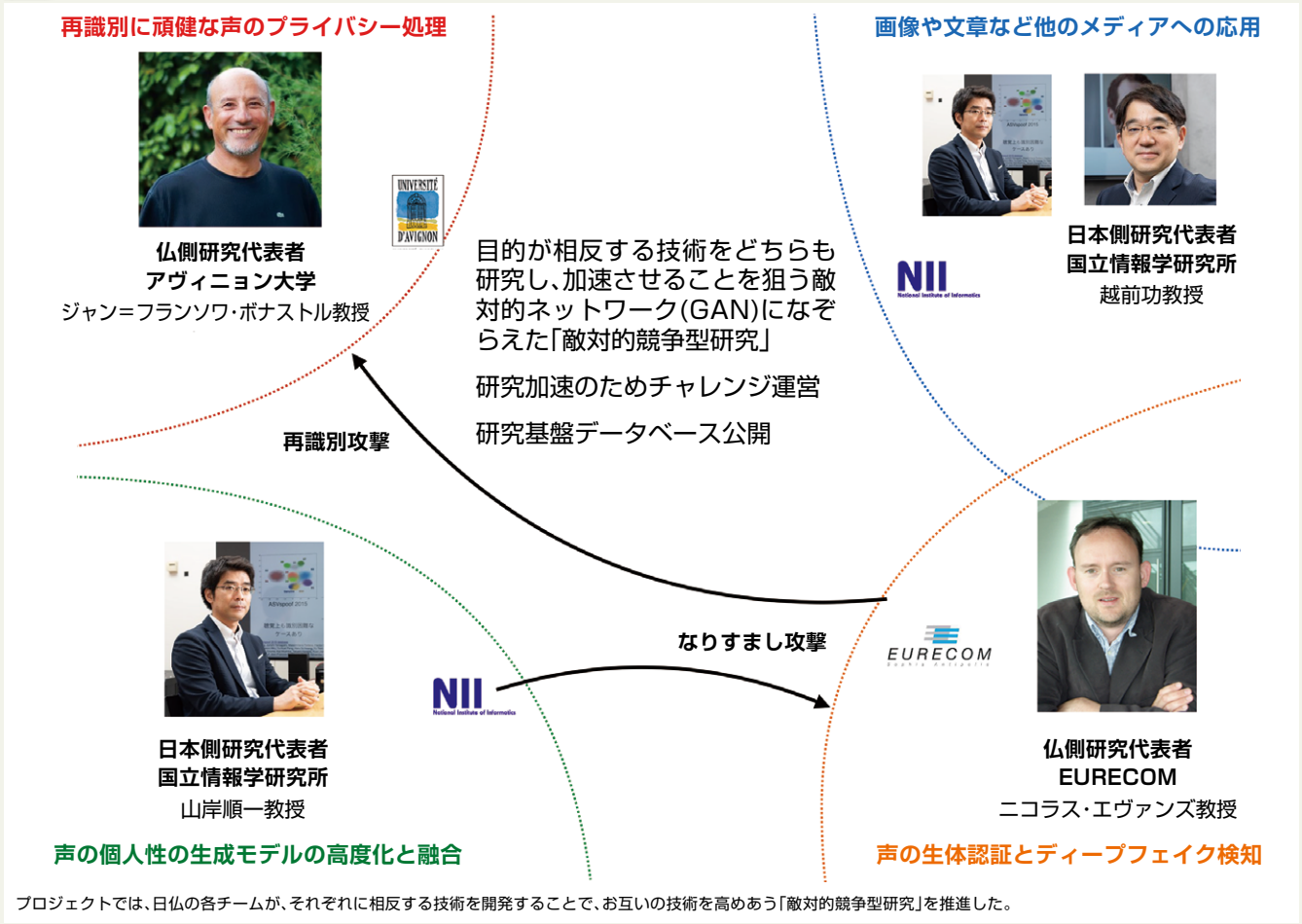
このプロジェクトのユニークな点は、国立情報学研究所、仏アヴィ

図1 社会問題化するディープフェイク



近年のAIの急速な進化に伴い、デマや詐欺、不正認証などにつながるディープフェイクが社会問題化している。

図2 プロジェクト体制



プロジェクトでは、日仏の各チームが、それぞれに相反する技術を開発することで、お互いの技術を高めあう「敵対的競争型研究」を推進した。

ニョン大学、仏研究センターであるEURECOMの日仏共同研究チームが相反する技術をお互いにぶつけ合うことで、各々の研究開発を加速させる「敵対的競争型研究」を採用していることだ(図2)。具体的には、山岸さんらが新しい音声合成モデルを開発したら、生体認証を突破できるかも併せて検証し、EURECOMはそれを防御するより高度な生体認証技術を開発する。さらにアヴィニョン大学はその高度な生体認証技術にも特定されない音声匿名化技術を開発する、といった複数の「矛と盾」による枠組みだ。

「各チームが開発した音声合成に関する技術を用い、お互いに『闘いあう』ことで、研究の成果を高めるとともに、さらに応用範囲も広げていこうと考えました」と山岸さんはこの構想の狙いを説明する。プロジェクトは4つのテーマに分かれて研究開発を進め、これまでに数々の成果を上げている。

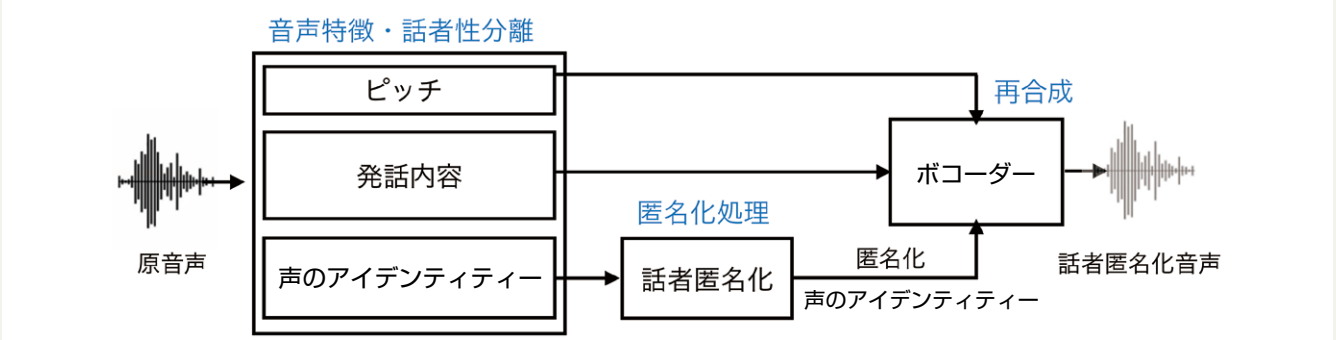
音声を強調・匿名化する技術 駅とテレビ番組に社会実装

1つ目のテーマは、声のアイデンティティーを再現するための生成モデリング技術の融合と高精度化。音声合成・声質変換という2つの異なるタスクを同時に実施可能な生成モデルのほか、音声の明瞭性指標を活用し、雑音の中でも音声が最も明瞭に聞こえるように音声を自動変換する「音声強調」も提案している。これをさらに応用し、雑音の中でも合成した音声を聞き取りやすくするシステムも開発。同技術は、音声合成システムを提供する国内IT企業に導入され、現在、東海道新幹線の各駅ホームのアナウンスにも採用されるなど、すでに社会実装済みだ。

2つ目のテーマは、音声によるディープフェイクの検知高度化。セキュリティー上の危険性を軽減する技術を開発しており、防御技術の研究開発でこれまでにも数々の成果を上げている。その一例が、ディープフェイク音声を検知するための防御モデルの学習に必要な大規模音声データベースの構築だ。「2019年に公開してから現在に至るまで約80万回ダウンロードされるなど、標準データベースとして広く利用されています」と山岸さんは語る。

3つ目のテーマは、話者匿名化による音声のプライバシー保護。公開されている音声から個人を特定したり、ディープフェイク音声を生成したりすることが容易になっている。そこで、山岸さんらは話者情報を加工してプライバシーを保護する「話者匿名化技術」を開発した(図3)。すでに日本放送協会(NHK)の番組にお

図3 話者匿名化技術の仕組み



音声をピッチ、発話内容、声のアイデンティティーの3要素に分解。このうち、肉声のアイデンティティーを匿名化するために「K匿名化」と呼ばれる手法を用いる。K匿名化は、個人が特定される確率を「K分の1」に変換するもので、その値は任意に設定することが可能。ボコーダーは、音声を分析し合成するための技術や装置を示す。

いて、インタビューを受けた人の音声を匿名化するために利用されている。従来の匿名化では、まるで事件の犯人かのような低くてくぐもった音声に変化されていたが、明瞭性を維持したままの匿名化に成功した。

4つ目は、画像や文章など、他メディアへのディープフェイク対策への応用だ。ディープフェイクの対象は画像や動画などへも広がりを見せている。「そこで私たちも、さまざまなメディアを対象として、ディープフェイクを見破るためのツールを開発しています」。その成果として、ディープフェイク顔映像検知を実現するための技術を開発。さらにディープフェイク検知に必要な処理を手軽に利用可能なプログラム「SYNTHETIQ VISION（シンセティック ビジョン）」を開発し、複数社にライセンスを提供し、著名人のディープフェイク映像検知のために利用されている（図4）。

より高度な検知手法を開発 未知のディープフェイクに挑む

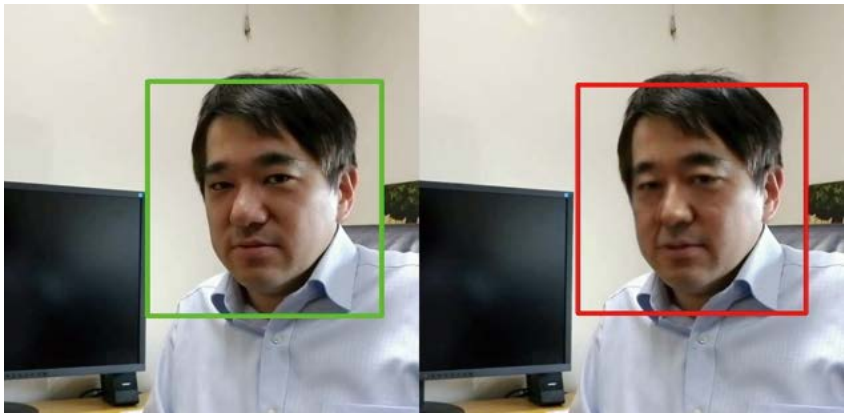
そして現在、JSTのAIP加速課題において、フェイクメディア検知技術の社会実装加速と普及を目標に掲げ、さらなる研究開発にまい進している（図5）。山岸さんは「先のCRESTの強化発展として、この課題ではディープフェイク検知技術を音声と映像も融合したマルチメディアに対応させるほか、新たに出現することが予測される未知のディープフェイクに追従可能な機械学習法の開発にも取り組んでいます」と説明する。

その成果の1つが、新たなディープフェイクの出現にも対応可能な機械学習アルゴリズムの開発だ。従来、ディープフェイクを検知するためのAIを実現するには、膨大なデータをAIに与え、学習させる必要がある。しかし、学習が完了した後に出現した新たな未知のディープフェイクを高精度に検出することは容易ではなく、一定の周期で検知モデルの再学習を手動で行うことが必要だったのだ。

これに対し、山岸さんらが開発したアルゴリズムは、新たに出現したディープフェイクの検知に必要なデータをAI自身が判断し、自動的にデータベースを拡張させることができる。「これにより、未知のディープフェイクにも追従可能な仕組みの実現を大きく前進させることができました」と新アルゴリズムの意義を語る。また、フェイクメディアを検知するための基盤モデルに関する研究成果も出ている。5万6000時間に達する人間のリアルな音声と約1億4200万枚の画像データを用いた基盤モデルを利用したディープフェイク検知モデルを構築したのだ。

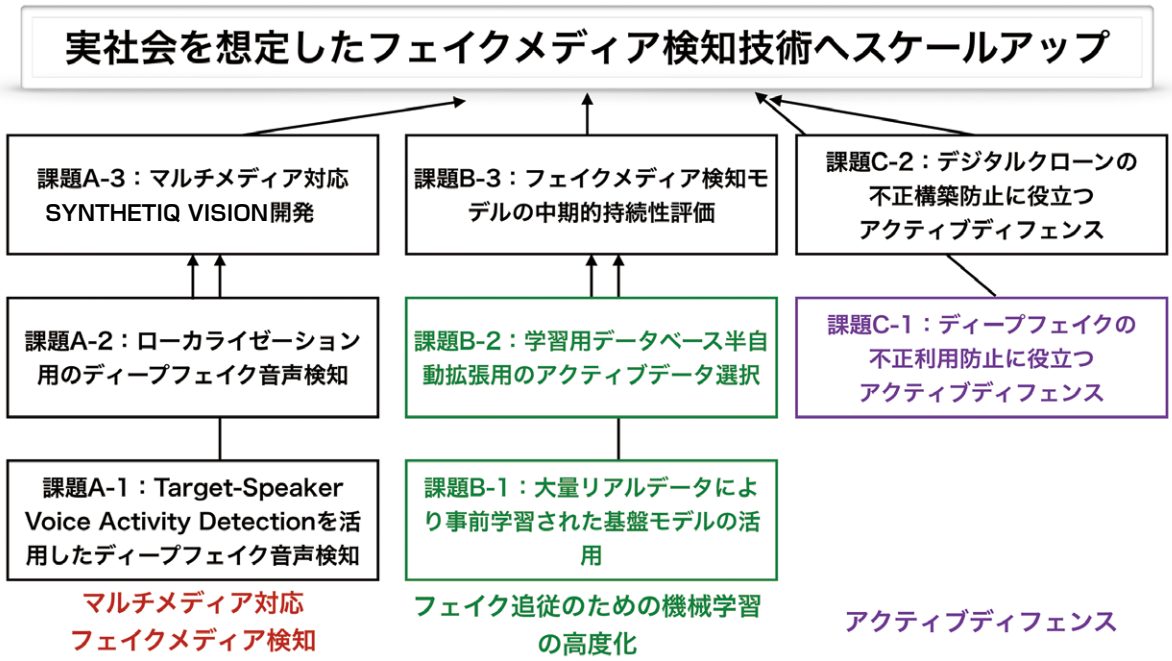
「これにより、リアルな人間だけが持つ特徴に基づき、ディープフェイク音声や画像を見抜く仕組みを実現できました。画像の場合、誤認識率を5パーセントにまで抑えられるなど、高精度な検知を実現しました」。このほかにも、画像や音声がオリジナルであることを証明したり、ディープフェイクに使用された元画像や音声をオリジナルのものに戻したりするといった技術の開発にも取り組んで

図4 SYNTHETIQ VISIONの判定イメージ



緑枠で囲まれた左側の画像が本物であり、赤枠で囲まれた右側の画像がディープフェイクで生成された偽物であると示される。人間の目では区別が付かない。

図5 社会実装を目指すAIP加速課題の研究構想



AIP加速課題において山岸さんが定めた目標。より社会実装を目指した、さまざまな技術開発、仕組みの実現が掲げられている。

いるという。

現在の研究領域に固執せず 社会での利用見据えた視野を

音声認識・合成の分野において、時代の要請に応じた先進的な取り組みに果敢にチャレンジし、数々の成果を上げてきた山岸さん。CRESTに採択された2018年から現在に至るまで、約160本もの査読付き論文を書いた。18年に発表した、ディープフェイク映像検知モデルの「MesoNet（メゾネット）」を発表した論文は、これまでに1600回以上も引用された。ディープフェイク映像の検知と改ざん領域の特定を同時に実行可能な「Capsule-forensics（カプセルフォレンジックス）」を発表した論文は、ここ5年間の中でインパクトの高い生体認証分野の論文として2回も選ばれている。

計画的に研究開発を進めてきたように思えるが、決して常に先を見据えて研究活動をしてきたわけではないという。「大学生の時には、サウンドエンジニアに憧れていて、スタジオで働こうと考えたこともあるんです」と笑って打ち明ける。自身が持つ音と情報処理に関する知識を生かせる研究職の道を歩むことになってからも、2018年には音声合成技術が一定のレベルに達したことで一度は研究が行き詰まった。

しかし、そこでディープフェイクに対する検知・防御という新たな目標を探り当てたことが、現在につながっている。この経験から、若手研究者には、現時点での自分の研究分野やスタイルに固執せずに「これから自分が向かう先に何があるのか」を常に考えつつ、日々の研究活動にいそしんでもらいたいと語る。「基礎研究に専念するだけでなく、実際に社会で利用してもらうために必要な努力を惜しまないでほしいと願っています」。山岸さんは、次代を担う研究者に向けてそう語った。

（TEXT：佐宗秀海、PHOTO：石原秀樹）

今やディープフェイクは、大きな社会問題となっています。これまでも音声合成を基軸にさまざまな防御のための技術を実現してきましたが、引き続き、現在のプロジェクトを加速させ、企業や日本社会のディープフェイク対策に貢献していきたいと考えています。