親しみやすいコンピュータ音声合成

Talking with Computers - Future Styles of Communication

ニック・キャンベル
Nick Campbell
ATR音声翻訳通信研究所
ATR Interpreting Telecommunications Research Laboratories

音声による会話は我々人間にとって最も自然で好まれるコミュニケーション方法である。そして人間と会話できるコンピュータがあれば、情報へのアクセスはより簡単に、より一般的なものになるだろう。コンピュータによる音声合成が発明されて約半世紀が経つにもかかわらず、まだその技術は身近に利用できるものとは言い難い。しかし、音声合成の技術はコンピュータや通信によるコミュニケーションをいっそう身近で一般的なメディアとしうる可能性を秘めていることを紹介する。

近年、Ｗｏｒｌｄ　Ｗｉｄｅ　Ｗｅｂの成功から一気に加速したオンライン情報網や移動体通信はかつて見られないほどの勢いで我々の生活に浸透し始めている。より性能の高いコンピュータとテレコミュニケーションを基盤とした情報網を使うことにより、情報化社会に住む我々の生活は一層便利で快適なものになった。しかし、情報化社会の進化は同時に新しいタイプの社会格差を引き起こした。コンピュータのキーボードを自由に操り、情報化社会にアクセスできる人とコンピュータを使えないがために情報化社会から取り残される人々の差である。一方、携帯電話が近年、急速に普及したのは、我々人間にとって最も自然で簡単な「声」によるコミュニケーション方法に拠るところが大きい。１９６０年代、音声合成の研究は盲人のための朗読機械として開発が進められていた。しかし、現在の音声合成技術は「声」を媒介にオンライン情報網と会話するかけ橋として、コンピュータの複雑な操作を行わずに、誰もがどこからでも簡単にアクセスできるよう、研究が進められている。

通常、音声合成器にテキストを朗読させる場合、区切りの位置やアクセントはもちろん、文章の構成・要約などを考慮した形、つまり我々が理解しやすい形にして機械に朗読させる必要がある。しかし、新聞やＷｅｂをはじめとするテキストはその様な朗読用の形態になっておらず、また、オンライン情報網で頻繁に使われている電子メールのテキストなどは口語体に近く、そのまま読み上げても意味がわからないこともある。合成音声の「声」によって情報にアクセスする場合、これら様々な形態のテキストと対話できるよう、わかりやすい形態に変換したうえで、必要な情報だけを出力できるような処理技術が必要になる。その他にテキストの内容に応じて声のトーン、話し方、ピッチやスピードを変えるだけでなく、情報の意図をより的確に伝えるため笑いなどの非言語的要素も必要となるかもしれない。

これら多様な形態のテキスト内容を合成音声によって情報をに変換出力する際、声のトーンや韻律がテキスト内容と異なる印象を与えないよう注意が必要となる。文字から音声合成への変換の際に生じる誤りを最小限にするには、限られた韻律特徴のみを用いて合成音声を生成するしかないが、その方法だと単調で非常に聞きづらいものになってしまう。あるいは伝達誤りの危険性を冒してでも、よりわかりやすく、聞きやすい韻律を伴った合成音声にするかどうか、今後その伝達の誤りに対しての責任問題を含め考慮が必要である。

Ｗｅｂを始めとするオンライン情報網の発展に伴い、今後、音声合成の研究で急がれるのは、多言語音声合成である。現在、多言語音声合成の技術は翻訳通信分野での活用が期待され、音声認識及び言語翻訳の技術と組み合わせて、電話回線やコンピュータネットワークを通じて提供するサービスが考えられている。

現在、我々が日常的に接してる情報をはじめ、Ｗｅｂや電子メールなどのテキストは日本語や英語など様々な言語を含んでいることが多い。従来の技術では日本語で記述された部分は日本語話者の声を、英語の部分は英語話者の声を使い発話されていたため、ちぐはぐな印象が否めなかった。しかし、ＡＴＲ音声翻訳通信研究所では同じ話者で、異なる言語にも対応可能な音声合成の研究を進めている。その手法は、異なる言語で記述されている部分は、一度異なる言語でのケプストラム・データを生成し、ネイティブスピーカーの韻律を再現する。そのデータをもとの言語コーパスから音声波形を選択する際のターゲットとして使う手法をとっている。例えば、日本人のＡさんの声で英語の合成音声を生成する場合、英語話者の声で合成した発話をＡさんのコーパスデータから音声波形を単位選択する際のターゲットとして用いると、ネイティブの発話に近い韻律を持った合成音声ができる。この方式だと同一話者の声で何カ国語にも対応が可能で、しかも自然な韻律を伴った合成音声となる。しかし言語ごとに使用される音素や発音が異なるため、ソースとなるコーパスに類似する音素が含まれない場合、必ずしも「ネイティブに近い音」が選択できないこともあるが、現段階では最も発話ターゲットに近い合成音声を生成する方法と言えよう。

音声合成の研究で、現在直面している問題は、音質に対して確立された評価規準がない点が挙げられる。かつての音声技術は了解度のみの評価で十分であった。ハスキンスによる変則的な文章やハーバード大学が考案したテスト用文章は、単語の理解力、アクセントの位置、デュレーションなどを評価するため使用されてきたが、合成音声のレベルが向上すると共に、その評価目的は自然性へと移行した。例えば人間が外国語を話す場合、既に評価体系が確立されているため評価が可能だが、合成音声の場合、生成された音声のレベルそのものと、音声コーパス提供者である話者の特徴や話し方をとらえているかどうか総合的な評価が必要となる。合成音声が必ずしも人間と同じ様な自然な声を発話する必要はないと言う考えもあるが、やはり聞きやすさの点からもより人間に近いほど、コミュニケーションの伝達はスムーズに行えるであろう。

興味深いことに、合成音声が人間の声に近づくにつれ、その合成音声に対する評価は厳しくなっている。以前なら合成音声は言葉を聞き取れる程度でよかったものが今では本人の声をどれだけ忠実に再現しているかが問われるようになった。生成された合成音声をコンピュータによる発話としてとらえるか、または人間の発話としてとらえるかによってその評価は分れる。

また、合成音声が人間の声に近づくにつれ、声に対して肖像権や著作権を適用し、保護する法的問題が生じてきている。今まで合成音声に対してこれらの権利が問われなかったのも、音声合成で作られた「声」が本人の声とは見分けがつかないほど似ていなかったためである。従来の音声合成の生成方法では、本人の声をサンプリングしても、韻律の調整や信号処理が施されていたため、生成された合成音声は良くても「声が似ている」程度でしかなかった。しかし、ＡＴＲが開発した多言語音声システムＣＨＡＴＲで用いられている波形接続法による合成音声は、かなり忠実に本人の声が再現できるようになったため、注意して聞かないと区別がつかないほどのものもある。現在、我々の研究に声のサンプル（コーパスデータ）を提供してくれている人々の権利を守り、音声合成技術が悪用されないためにも法的な面での整備が急がれる。

この様にコンピュータによる音声合成をより身近な技術として活用するには、幾つかの課題が残されている。しかし、音声合成技術はキーボードを必要としない「声」によってコンピュータを操作し、「合成音声」によって情報を誰もが自由にどこからでも聞きだすことができる平等な近未来のコミュニケーション方法の１つとしてその役割が期待される。