

英語音声教育における TTS 合成音声の活用とその問題点

東 淳一 (神戸学院大学)
azuma@gc.kobegakuin.ac.jp

1. はじめに

今日私たちの生活の周りにはTTS合成音があふれている。例えば速報性が要求される高速道路のハイウェイラジオも最近TTS合成音になったし、合成音により読み上げられた音をそのまま放送している日本のFM放送局も存在する。新幹線のアナウンスにもTTS合成音が一部利用されている。今日の社会において、能率的に音声情報を発信するためにかなり多くの部分でTTS合成音が利用されてきており、私たち自身これがTTS合成音であると気づいていないサービスも多い。今後もTTS合成音の利用はますます拡大すると考えられる。

実は音声言語が必要とされる場面は、教育分野においても非常に多い。もちろん我々の日々の講義も音声を使って行っており、また英語教育においてはターゲット言語の音声そのものが大変重要な教材となり使用されている。もちろん教員本人も授業中に英語を発するが、ある程度自動的かつ大量に文型練習を口頭で行わせる、音声の聞き取り練習をさせる、スピーチやプレゼンのモデル音声として音声を提示するなど、生身の教員だけでは対応できないケースも多い。また、ヒアリングの試験では英語のネイティブスピーカーによる音声が必要となり、その場合、アメリカ英語のみならず世界各地の英語の音声が必要とされることもある。現実には、TOEICのリスニングセクションでは、アメリカ英語、イギリス英語、カナダ英語、そしてオーストラリア英語が利用され、受験生は異なるタイプの英語音声に対応できなければならない。大学院生や研究者については、国際学会において英語で発表するケースも頻繁にあり、その場合にはモデルとなる英語音声を聞きつつ何度も発話練習するという必要に迫られることが多い。ビジネスマンの場合にも、英語でのビジネスプレゼンの練習、あるいは英語による交渉のための模擬練習も欠かすことができない。やはりプレゼンのモデルとなる英語音声や、交渉のさまざまな場面を想定した英語での想定問答がネイティブスピーカー発話による音声として提供されれば、非常に有益である。発話練習、スピーチの練習が必要であっても、まずはモデルとなる適切な英語音声のインプットが存在することが重要である。さらに学習者の能力と聴衆の属性、話される内容のジャンルなどに対応した適切な韻律の英語音声が学習者にモデル音声として提供されることが必要である。ネイティブスピーカーによる発話の録音にかかるさまざまなコストを考えれば、このような場面においてもTTS合成音声の活躍が大いに期待される。

上記のようなニーズを考慮に入れ、本研究では、特にAmazon Pollyに代表される、クラウドベースのTTSサービスの実態とその利用方法について述べ、さらに学習者のレベルや産出される英語音声のジャンルなどに対応する適切な合成音声を作り出すには、TTS合成音作成過程で、どのように韻律を制御しなければならないのかについても解説する。

2. ネイティブスピーカーにモデル音声の録音を依頼する際のコストと問題点

2.1. 人材確保の難しさと金銭的成本の問題

英語教員がネイティブスピーカーに録音を依頼し教材を作る、あるいは英語のリスニングテストを作るなどの作業を行いたい場合、いくつかよく問題になる点がある。その一つはコストの問題である。ネイティブスピーカーの先生に録音を依頼するということになるとかなりの録音のための時間を要することになり、またそのためにある程度の謝金を用意するということになるが、その額についてもそう安くはできない。また、これは地域にもよるが録音に対応可能なネイティブスピーカーを確保することは容易ではない。

2.2. 技術的問題

ネイティブスピーカーに録音を依頼する場合、外部からの雑音が遮断された録音室が必要となるが、そのような環境が教育施設に整備されている可能性はかなり低い。また高い謝金を払ってネイティブスピーカーに録音をしてもらったものの、録音終了後にほんのわずかな録音漏れがあった、あるいは録音のミスがあったことがわかった場合、いかにそれを修正するのも大きな問題となる。再録が可能になったとしても、録音機材の録音レベルや録音室内での話者が座る位置の違い、あるいはマイクの位置の違いや話者の健康状況などによっても録音品質や雑音のレベルが大きく変わってくる。このため、後に録音した音声をサウンド編集ソフトを使って最初に収録した音声データに貼り付けても、不自然な音声に仕上がってしまう可能性がある。

2.3. ネイティブスピーカーの問題

ネイティブスピーカーについても、依頼者の思うように素材を発話してくれるとは限らない。当然収録前の打ち合わせにかかる時間も多くなり、また修正を依頼すると、ネイティブスピーカー自身が依頼者の考える韻律は不自然であって自分が正しいと主張する場合もあり、なかなか想定しているような理想の音声の収録ができないこともある。

3. TTS 合成音声を使用する利点と利用シーン

3.1. TTS 合成音声使用の利点

テキスト入力を行い、ポンとクリックすれば音声合成されるという利便性だけではなく、TTS 合成音声を利用することにはさらに複数の利点がある。もちろんもっとも重要な点は音声教材の作成が極めて簡単になるということである。2 番目として、音声はデジタルデータとして電子的に蓄えられるため、その音声素材の共有は極めて容易になることがある。つまり他の教員、学習者との音声ファイルの共有がきわめて容易になり、またその配布も簡単になる。3 番目にあげるべき点は音声データ編集の容易さである。かつてのアナログ音声データを編集あるいは修正しようとする大変な労力と時間がかかった。しかしデジタルデータであれば適切なソフトにデータを瞬時に読み込み、編集・修正が終われば瞬時にデータを書き込むことができる。また4 目目にあげるべき点として、かつてのオーディオテープのようにメディア素材そのものがさほど短時間に劣化することがないとい

うことがある。このため保存したメディアの経年変化についてはさほど神経質になる必要がない。

3.2. 英語教育での TTS 合成音声の利用シーン

TTS 合成音を英語教育の現場で活用する場合、想定されるシーンとしては以下のようなものが考えられる。

- リスニング練習用のモデル音声，指示のための音声として。
- リスニングテスト用のモデル音声，指示のための音声として。
- シャドーイングやオーバーラッピング練習のためのモデル音声として。
- パブリックスピーチやプレゼンテーションの練習のための音声モデルとして。
- 対話練習用の音声素材として。学生の発話部分を無音にしておき相手方の音声のみを作成しておく等。
- 動画制作において。ナレーションやアフレコ，あるいは TTS の複数の音声による擬似的な対話音声の作成と挿入等。
- 教員の研修用の音声素材として。特に小学校での英語教育のための教員研修等。

リスニング練習，あるいはリスニングテストについては Moodle などの LMS を使い，オンラインで実現することが可能である。次の図 1 は，TOEIC のリスニング問題で絵を見てその描写として正しい英語の発話を選ぶ問題を，サンプルとして Moodle で作成してみたもののスクリーンショットである。



図1:Moodle 上でのリスニング問題のスクリーンショット

図中のプレーヤを再生すると1～4の4種類の音声の流れ、写真の描写として正しいものを選ぶ形のものである。なお、新しいTTS合成音活用シーンとしては、動画でのナレーションやアフレコの用途が考えられる。音声の収録以上にビデオの収録は時間がかかり、さらに何か不備が後で見つかったとしても撮り直しが大変困難である。そういうケースに備えて、話者の姿を写さなくても良いシーンや、静止画を次々と表示しバックでナレーションを入れるだけで済むような場面も作っておけば、音声についてはTTS合成音で十分に間に合う。また、登場人物が日本語で話している場面で英語での音声の吹き替えを行う必要がある場合にも、TTS合成音の使用が便利である。

4. Amazon Polly の利用について

ほんの2, 3年ほど前まではTTSといえばスタンドアロンタイプのソフトが普通であり、ソフトにはTTS音声は含まれておらず、同時に好みのTTS音声を購入しなければならないという場合も多かった。東(2018)で報告したように、現在ではAmazon, Google, Microsoftがそれぞれクラウド上でWebサービスを展開しており、それぞれにおいてTTSのサービスも利用できる。この3社のうち、Amazon PollyがWeb上のGUIベースのインターフェース上で音声合成を実現できるため、特に気軽に使える。紙面の関係で手続きは省略するが、使用に当たってはあらかじめアマゾンウェブサービス(以下AWS)に登録しておく必要がある。Amazon Pollyをキーワードに検索し、登録されたい。なお、ログイン後にはAWSマネジメントコンソールの画面に移動するので、そこで「サービスを検索する」の部分からAmazon Pollyもしくは単にPollyで検索をかける。すると、以下の図2のようなコンソール画面に到達する。すでにテキストを入力してあるが、中央の四角い部分にテキストを打ち込み、「音声を聞く」スイッチをクリックすれば音声を再生し、「ダウンロードMP3」をクリックするとMP3形式で音声ファイルをダウンロードできる。2019年7月末からは、24000Hzサンプリングの高品質音声を提供されており、その場合にはエンジンをスタンダードではなくニューラルに切り替える。Amazon Pollyの課金情報は、<https://aws.amazon.com/jp/polly/pricing/>に記載されているが、原則1件あたり1,000文字のリクエストを1,000件行った場合(つまりテキストの長さにして合計100万文字)スタンダード音声は4米ドル、ニューラル音声は16米ドルであり、大変安価である。合成音について使用目的の制限はない。

通常、合成音はニュース音声をモデルとして合成される。このため発話速度はかなり早く、韻律についてもニュース番組のアナウンサーのようなスタイルになり、イントネーションについてもあまり大きなF0の起伏はない。フレーズの長さも長く、ピリオドでのポーズの長さもさほど長くはない。このままでは日本の英語教育用の英語音声としては利用が困難かもしれない。しかしながら、Amazon Pollyでは合成音作成者が合成音にお好みの韻律を付与できる。この場合には、コンソールの「プレーンテキスト」ではなく「SSML」のタブを選択し、そちらにテキストを入力する。

テキスト読み上げ機能

音声の聴き取り、カスタマイズ、ダウンロード。準備が整ったら統合。

ウィンドウにテキストを入力するか貼り付けて、言語と地域を選択し、音声を選択します。次に、[音声を聴く]を選択し、アプリケーションとサービスに統合します。

3000 文字以内の場合、すぐにリッスン、ダウンロード、または保存できます。100,000 文字までの場合、タスクは S3 バケットに保存する必要があります。

ブレンテキスト SSML ⓘ

If we use today's text-to-speech technology, our teaching will be of course much easier. First, creation of teaching materials including audio will be quite easy. Second, as the audio part is digitally stored as a computer file, sharing audio materials with other instructors will be also very easy. Third, modification of the analog audio material is quite difficult and requires a complicated process. However, the synthesized voices are quite easy to modify. Reuse of the text-to-speech synthesized voices is quite simple and done in a fraction of a second.

560 文字を使用

デフォルトのサンプルレートは、スタンダードエンジンの場合 22,050Hz で、ニューラルエンジンの場合 24,000Hz です。

デフォルトのテキストを表示 テキストを消去

エンジン ⓘ

☒ スタンダード ☐ Amy, 女性
☐ ニューラル ☒ Emma, 女性

言語とリージョン

英語 (英国)

▶ 音声を聴く

📄 ダウンロード MP3

サンプルレート: 24,000Hz

図2: Amazon Polly のコンソールのスクリーンショット

下の図 3 が SSML 対応のコンソールに必要な SSML タグを打ち込んだ場合のスクリーンショットである。

ブレンテキスト SSML ⓘ

<speak><amazon:breath/> If we use today's text-to-speech technology, <break time=".4s"/>our teaching will be of course <emphsis level="moderate">much easier</emphsis>. <break time="1s"/><emphsis level="moderate"><amazon:breath/>First,</emphsis> <break time=".3s"/><amazon:breath/>creation of teaching materials including audio, will be quite easy.<break time="1s"/> <emphsis level="moderate"><amazon:breath/>Second,</emphsis> as the audio part is digitally stored as a computer file,<break time=".3s"/> <amazon:breath/>sharing audio materials with other instructors will be also very easy.<break time="1s"/> <emphsis level="moderate"><amazon:breath/>Third,</emphsis> modification of the analog audio material is quite difficult, <amazon:breath/>and requires a complicated process.<break time="1s"/> <amazon:breath/>However,<break time=".3s"/> the synthesized voices are quite easy to modify.<break time="1s"/> <amazon:breath/>Reuse of the text-to-speech synthesized voices is quite simple, <amazon:breath/>and done in a <emphsis level="moderate">fraction of a second.</emphsis> </speak>

1097 文字を使用

デフォルトのテキストを表示 テキストを消去

図3: Amazon Polly の SSML モードでのコンソールのスクリーンショット

SSML タブ右横のクエスチョンマークをクリックすると、SSML に関する説明が表示される。また、https://docs.aws.amazon.com/ja_jp/polly/latest/dg/supported-ssml.html では Amazon Polly が現段階でサポートする SSML タグがその使用法とともに示されている。実際のところ

ろ、SSML で韻律のファインチューニングを行うことで、それぞれの教育シーン、あるいは学習者のニーズにうまくマッチした韻律をもつ合成音を作成することが可能である。

なお、Azuma(2008)あるいは Azuma(2010)で述べたように、従来のスタンドアローン形式の TTS 合成ソフトの利用については、合成された音声ファイルの利用についての制約が多く、作成したファイルを学習者に渡したり、誰もがアクセスできる Web サイトで音声のストリーミングを行ったりということは、ライセンスの関係で大変困難であった。この点、最近登場したクラウドベースの TTS 合成サービスは僅かな課金はあるものの、作成された音声ファイルの配布についての制約がなく、英語教員にとっては大変使いやすい環境が整ったといえるだろう。

5. おわりに

クラウドベースの TTS 合成音作成サービスの Amazon Polly を中心に、TTS 合成音の英語教育シーンでの活用法について述べた。たしかに人的コスト、時間的・金銭的成本を大幅に削減しつつネイティブスピーカーにかなり近い英語音声を作成できる、そしてわずかな課金があるものの作成された音声ファイルを自由に配布、活用できるという点で、これらクラウドベースの TTS サービスは私たちにとって福音である。しかしながら、SSML タグを巧みに打ち込み、試行錯誤を繰り返しつつ、それぞれの教員シーンや学習者のニーズにマッチした韻律をもつ合成音を作成することは、思ったほど容易ではない。

今後の英語教育での TTS 合成音声の利用研究の重点は、同じサービスの同じボイスを使ったとして、いかに教員のニーズにマッチした韻律をもつ音声をそのボイスから作り出すか、つまりいかに SSML タグを駆使するのか、そしてさらにその次には、いかに SSML を駆使して教育シーンにマッチした TTS 合成音を作成できる「TTS 合成音プロデューサー」を養成することができるかといった点に移行するであろう。

参考文献

- Azuma, J. (2008) “Applying TTS technology to foreign language teaching,” *Handbook of Research on Computer-Enhanced Language Acquisition and Learning*, Zhang, F. and Barber, B. eds, IGI Global, Hershey, 497-506.
- Azuma, J. (2010). “A quantitative evaluation of the quality of Text-To-Speech (TTS) engines: toward the application of TTS technology to TEFL,” *Language Education and Technology (LET 50) Conference Proceedings*, Yokohama, 232-233.
- 東淳一 (2018) 最新 TTS 合成音声の外国語教育現場での活用. 2018 年度外国語教育メディア学会全国研究大会. 2018.8.9.