実践報告

日本語の合成音声を用いた講義動画の作成

―調整方法の報告と「聞きやすさ」に関する考察―

劉　羅麟・伊藤　茉莉奈・小林　美希
中川　彩野・渡邉　咲・福島　青史

要　旨

本稿では、オンデマンド科目で使用する講義動画を作成する過程における、1）合成音声の不自然な箇所を調整する方法と、2）調整を行うか否かを判断する中で行われた音声の「聞きやすさ」に関する考察について報告する。ソフトウェアが生成した合成音声には、不自然な読み方・ポーズ・アクセントがあったため、文字表記と句読点の変更・コマンドの挿入・同音異義語への置換などの方法で調整した。こうした調整を行うか否かを判断する中で、①読み方の間違い、②テーマとなる語句の後におけるポーズの欠如、③語句の途中に挿入された不要なポーズ、④頻出する語におけるアクセントの崩れ、⑤ひとまとまりであるはずの語句におけるアクセントの分断や、長い漢字語彙のような難しい語彙におけるアクセントの崩れ、⑥音声上の問題点の多発、という6点が講義動画としての「聞きやすさ」に影響を及ぼす要因として浮かび上がった。

キーワード

読み方　ポーズ　アクセント　Amazon Polly　AI技術

1．はじめに

オンデマンド科目で使用する講義動画を作成するために、筆者らは合成音声を用いて作業を行った。合成音声には不自然だと感じる箇所も多いが、作業時間の関係上、それらを全て修正するのではなく、許容できる不自然さはそのまま残した。ただ、この選別においては、講義動画が履修者にとって聞きやすいかどうかが基準となった。合成音声の不自然さは技術の発展により早晩解決されるものだが、技術の発展途上において露呈したこの不自然さは逆に、講義動画としての「聞きやすさ」を考えるうえで重要なデータとなった。

本稿では、合成音声を用い講義動画を作成する過程において、不自然な箇所を調整する方法と、調整を行うか否かを判断する中で行われた音声の「聞きやすさ」に関する考察について報告する。

2．合成音声を導入した経緯とその試用で遭遇した問題

本稿で取り上げる講義動画は、2023年度より早稲田大学大学院日本語教育研究科が提供

するオンデマンド科目「日本語教育学入門」で使用されるものである。この科目が開設された2020年度には、アナウンサー経験のある人物に音声の吹き込みを依頼したため、講義動画が自然で聞きやすいものとなっていた。一方で、人による音声の吹き込みは、講義内容の一部を修正する場合、一部だけ声が変わると不自然なため、再度同じ人に依頼する必要がある。それが難しければ講義内容を全て吹き込み直すことになり、時間と費用がかかる。しかし、日本語教育を巡る教育的・社会的環境は急速に変わるため、講義内容の修正が簡単にできる体制が必要であった。そこで、音声の自然さを多少犠牲にしても、内容の更新が重要という判断から、合成音声による録音が導入された。

文字データから合成音声を生成できるソフトウェアとして、筆者らは無料利用枠のあるAmazon Polly（以下、Polly）[1]を選択した。Pollyでは、テキストボックスに文字を入力し「音声を聴く」ボタンを押すと、合成音声が即座に生成され確認できるようになる。試用の段階では、スタンダードとニューラルという二つの音声エンジンを比較し、ニューラルのほうが相対的に自然だと判断した。しかし、調整なしの状態では不自然に聞こえる箇所がまだ多く、今回は主に読み方・ポーズ・アクセントという３種類の問題に遭遇した。

３．講義動画の聞きやすさを考慮した合成音声の調整

本章では、読み方・ポーズ・アクセントの問題を解決するために筆者らが取った、合成音声を調整する方法について報告する。そのうえで、調整の作業を捉え直すことで、講義動画としての「聞きやすさ」について考察する。

3.1　読み方の調整

一つ目の問題は、意図しない読み方になる場合である。これは英数字にも起こるが、漢字の場合に特に起こりやすいようである。例えば、「留学生の黄です」という文では、人名の「黄」は「コウ」ではなく「キ」と読まれる。このような読み方の軽微なずれであれば、講義動画の視聴に大きな影響を与えることはなかろう。しかし、耳で得た聴覚情報（講義内容の合成音声）が、目で得た視覚情報（文字やイラストなど）、または聞き手の既有知識（ある語の通常の読み方）がずれる度に、履修者の注意がその不一致に向けられ、講義内容を理解するうえで妨げとなりかねない。

このような音読み・訓読みの間違いは基本、漢字表記を仮名表記に変更することで解決できる（例：留学生の黄です→留学生のコウです）。ほかには、フリガナで読み方を指定する方法もある。例えば、「研究所」が「ケンキュウショ」と読まれるように連濁の有無が意図に沿わない場合は、「研究所（ケンキュウジョ）」のように括弧でフリガナを指定する。ところが、読み方を調整してもアクセントの問題が残ったり、読み方を調整することによりアクセントの問題が生じたりする場合がある。この点については3.3で後述する。

3.2　ポーズの調整

二つ目の問題は、ポーズの有無、またはポーズの長さが不自然に聞こえる場合である。ポーズの有無については、Pollyでは句読点にポーズを置く仕様になっていると考えられ

るため、句読点を入れるかどうかで調整できる場合が多い。ただし、句読点に付与されるポーズの既定の長さは、必ずしも意図どおりの長さではない。ここでは「留学生の日本語について考えてみましょう。」という文を例に述べる。「について」の後に読点がないとポーズが置かれないため、講義動画を視聴する履修者にとって速すぎて理解が追いつかない可能性や、落ち着きがないような印象を与える恐れがある。なぜなら、「について」で示されるテーマは聞き手に注目してほしい箇所だが、ポーズがないと文のフォーカス（ひいては講義内容の要点）が掴みにくくなる。しかし、「について、」のように読点を入れると、必要以上に長いポーズになってしまう。

ポーズの問題を解決するために、Pollyのブレイクというコマンドを使用した。上記の例で言えば、「留学生の日本語について<break strength="weak"/>考えてみましょう。」のようにコマンドを挿入し、短めのポーズを入れた。ほかには、中程度のポーズを入れるコマンド（<break strength="medium"/>）や、長めのポーズを入れるコマンド（<break strength="strong"/>）もある。また、<break time="X.Xs"/>のようなポーズの長さを指定するコマンドもある。望ましい長さのポーズを句読点で指定しづらい場合は、講義動画を視聴する履修者のことを想定しながら、前述のコマンドを使い分けて調整する。

一方で、句読点もポーズのコマンドも置かれていないにも関わらず、意図しない箇所にPollyが自らポーズを入れる場合がある。特に、助詞の後にはこの傾向が顕著である。その結果、一文が二つに分かれているように聞こえたり、文の後半が重要であるような印象を与えたりしてしまう。今後、不要なポーズを削除するコマンドの開発が待たれる。

3.3　アクセントの調整

三つ目の問題は、アクセントが不自然に聞こえる場合である[2]。特に複合語の場合、「日本語教育」が「ニ↑ホ↓ンゴキョ→ウイク」のように読まれるなど、アクセントが崩れやすいようである。前述した読み方とポーズと比べ、アクセントの調整は格段に困難だが、軽視できない問題だと考える。なぜなら、講義動画に頻出する語のアクセントの崩れが、聞き手である履修者に持続的に違和感を与え、講義動画の視聴を妨げることに繋がる可能性があるからである。それだけでなく、例えば、「○○の14業種で」のアクセントが分断され「○○の14」と「業種で」と別々の語に聞こえる場合や、「自国第一主義」が「ジ↑コク↓ダイ→イチ↑シュ↓ギ」のように読まれることにより意味の特定すら困難になる場合も多い。さらに、聞き手が正しい意味を推測できたとしても、講義内容の信憑性に対する不信感に繋がる恐れもある。

アクセントを調整するために、Pollyの発音仮名というコマンドを使用した。このコマンドでは、<phoneme alphabet="x-amazon-pron-kana" ph="YYY">XXX</phoneme>のように、「XXX」に調整したい語を入れ、「YYY」に読み方を指定する。起伏式（頭高・中高・尾高型）の語はアクセント核の位置にアポストロフィ（'）を入れ、平板型の場合は入れない。例えば、前述した「日本語教育」を<phoneme alphabet="x-amazon-pron-kana" ph="ニホンゴキョ'ウイク">日本語教育</phoneme>に置き換えると、正しく「ニ↑ホンゴキョ↓ウイク」と発音される。なお、二語以上を一度に調整したい場合は、語をスペースで区切る。3.1で述べた、語の読み方を調整したことによりアクセントが崩れる場合には、この

コマンドで読み方とアクセントを同時に指定する方法が有効である。ただし、コマンド自体が煩雑であり、作業時間を短縮するために後述する調整方法を併用した。

一つ目は3.1で述べた文字表記の変更を援用する方法である。例えば、日本語能力試験（JLPT）のレベルを示す「N1」を、合成音声が「エ↓ヌ イ↑チ」のように読む。その原因は、合成音声のAIが「N1」を「N」と「１」の二語として認識しているからだと考えられる。そこで、「N1」を仮名表記の「エヌイチ」に変更することにより、一語であることをAIに認識させ、意図どおりに平板型の「エ↑ヌイチ」に調整できた。

二つ目は3.2で述べたポーズのコマンドを援用する方法である。例えば、「講義２は（以下略）」という文では、「コ↓ウギニハ」のように、数字「２」の部分にアクセント核が付与されない。これは、合成音声が人間の音声と違い、コマンドなしではポーズやプロミネンス（強調）のない中立発話になりやすいからだと推測される。そこで、「講義<break strength="weak"/>２は」のように間に強制的にポーズを挿入した。その結果、「２」のピッチの高さがリセットされ、「コ↓ウギ ニ↓ハ」のように自然なアクセントになった。

三つ目はアクセントが異なる同音異義語に置き換える方法である。例えば、前述した「日本語教育学」のアクセントが崩れる場合は、「日本」を「二本」にすることで自然なアクセントとなった。他にも、「日本語能力試験」の「試験」を「私見」にしたり、「中国人」の「国」を「語句」にしたりすることで自然なアクセントになるケースがあった。

3.4　講義動画としての「聞きやすさ」を考える

3.1から3.3までは、読み方・ポーズ・アクセントを中心に合成音声を調整する方法について報告し、なぜそれを調整する必要があったのかについて述べてきた。１章でも述べたように、今回筆者らが作成したのはオンデマンド科目に使用する講義動画である。不自然な合成音声を調整するか否かという判断の背後には、講義動画としての「聞きやすさ」という基準があった。合成音声を調整する作業を捉え直すことで、講義動画としての「聞きやすさ」に影響を及ぼす要因として、以下の６点が浮かび上がった。

１）読み方の間違いは、聴覚情報と視覚情報または既有知識とのずれを生じさせ、履修者の注意をその不一致に向けさせることで講義内容の理解を妨げる。
２）テーマとなる語句の後におけるポーズの欠如は、文のフォーカス、ひいては講義内容の要点を掴みにくくする。
３）語句の途中に挿入された不要なポーズは、あたかも後ろの部分が重要であるかのような印象を与え、講義内容において重要となる箇所に対する誤判断に繋がる。
４）頻出する語におけるアクセントの崩れは、聞き手に持続的に違和感を与え、講義動画の視聴を妨げる。
５）ひとまとまりであるはずの語句におけるアクセントの分断や、長い漢字語彙のような難しい語彙におけるアクセントの崩れは、意味の判断や特定を困難にさせる。
６）意味が正しく理解された場合でも、多発する音声上の問題点は、講義内容の信憑性に対する不信感に繋がりかねない。

上記の要因を排除するために行った合成音声の調整は、つまり、誤解の可能性を軽減させ、履修者の注意の分散や逸脱を抑制し、講義内容の信憑性を維持する作業とも言える。

4．振り返りと今後の展望

本稿では、読み方・ポーズ・アクセントを中心に合成音声の調整方法と、調整するか否かを判断する中で行われた音声の「聞きやすさ」に関する考察を述べた。本章では、合成音声を用い講義動画を作成してきた過程を振り返り、今後の展望を述べる。

今回の作業において筆者らは終始、自分自身を講義動画を視聴する履修者の立場に置き、講義内容に対する理解が妨げられるかどうかについて考えていた。そして、その恐れがあると判断した場合は、様々な方法を試行錯誤しながら合成音声の調整を行っていた。つまり、履修者にとっての「聞きやすさ」を追求したのである。これは、現在の合成音声のAI技術では、対応が十分になされていない点である。

日本語教育の分野において、合成音声を用いた講義動画の作成というようなAI技術の活用は、これからも益々発展していくと予測される。そのため、本稿で述べた読み方・ポーズ・アクセントの調整方法は、無論、合成音声を運用する際の参考になる。ただし、１章でも述べたように、合成音声自体の問題点は技術の発展により早晩解決されるものだと考えられる。むしろ、合成音声における不自然な箇所を調整するか否かという判断基準に関する考察が、日本語教育関係者にとって「聞きやすさ」について再考する契機になろう。今後も、合成音声などのAI技術の活用に関する知見と、その活用の過程で生まれる研究者・教育者の思考が共有されることが期待される。

注

1　本稿で述べる内容は2023年３月末時点の情報である。参照：https://aws.amazon.com/polly/
2　合成音声におけるアクセントの問題には、文のイントネーションが関与していると推測される場合もあるが、本稿ではアクセントに着目して述べる。

参考文献

Amazon Web Services ブログ「Amazon Polly を使用した日本語テキスト読み上げの最適化」<https://aws.amazon.com/jp/blogs/news/optimizing-japanese-text-to-speech-with-amazon-polly/>（最終閲覧：2023年３月24日）

Amazon Polly Developer Guide, Supported SSML Tags <https://docs.aws.amazon.com/polly/latest/dg/supportedtags.html>（最終閲覧：2023年３月24日）

（りゅう　ろうりん　東京大学大学院工学系研究科）
（いとう　まりな　早稲田大学大学院日本語教育研究科・研究生）
（こばやし　みき　早稲田大学大学院日本語教育研究科・博士後期課程）
（なかがわ　あやの　早稲田大学大学院経済学研究科・研究生）
（わたなべ　さき　宝塚医療大学留学生別科）
（ふくしま　せいじ　早稲田大学大学院日本語教育研究科）