Vol.2017-SLP-118 No.5
2017/10/13

音声合成エンジン AITalk を用いた事業展開と最近の取り組みについて

平井 啓之[1,a)]

概要：（株）エーアイは、日本語音声合成の専業メーカです。弊社がこれまで行ったきた、日本語合成エンジン（AITalk) をもちいたいくつかの事業および、関連する技術開発について、その概要を述べる。また、最近の音声合成市場の傾向や、弊社の今後の取り組みについても、簡単に述べる。

Recent approaches and business development with Text to Speech engine AITalk

Hirai Hiroyuki[1,a)]

1. はじめに

(株) エーアイ [1] は、2003 年創業の日本語音声合成の専業メーカである。弊社は、これまで日本語音声合成エンジン (AITalk) を用いて、様々な分野に音声合成の商品やサービスを提供してきた。近年の深層学習の進歩に伴い、音声を用いたマンマシンインタフェースも身近な技術として普及が進んでいる。それに伴い弊社も広い分野へ音声合成エンジンの提供が可能になってきている。本発表では、はじめに AITalk の概要、つぎに弊社がこれまで行ってきた幾つかの事業の紹介、最後に今後の取り組みについて説明する。

2. AITalk の概要

AITalk は弊社で開発された、コーパスベースの波形接続型エンジンである。現在では、WebAPI、SDK、サーバ、組み込み SDK など様々な形で様々な分野に提供を行っている。AITalk の特徴は、自然性の高い高品質な音声を生成するだけでなく、誰の声でも再現できることが挙げられる。弊社では、この技術をカスタム音声と呼んでいる。コーパスベースの音声合成では、原理的に作成に用いたコーパスの音声を再現することが可能である。しかし、カスタム音声をサービスとして成功させるには、どのような話者に対しても少ない収録で個性を再現する必要がある。AITalk は音質より個人性を重視したエンジンではあるが、最低２００文章からカスタム音声の作成が可能である。約２００文は収録時間で約１５分、話者の拘束時間は約２時間である。この程度の収録量から話者性を再現するには、HMM の話者適合技術を利用することが多い。しかし、AITalk では極端な個性を持った音声でも収録された文章と同様の文書であれば、生々しく再現することを目指し、波形接続型合成方式を用いている。詳細な検証は行っていないが、これを実現可能にしているのは、以下の２つが大きな要素だと推定される。第１は、韻律の物理量と対応が良い中間言語の利用と、その中間言語で記述されるアクセントやイントネーション情報を丁寧に手動でラベリングしている点である。個人性を再現するには、スペクトル情報と、韻律（F0, パワー、継続時間長）情報を再現する必要があるが、コーパスのサイズが小さい場合は、特に韻律の再現が重要であると考えている。よって、AITalk では、アクセント句間の相対的な F0 の変化や、文末の F0 の形状を中間言語で定義し、それを正しく手動でラベリングすることで、コンテキストと物理量との対応を良くし、少ないデータから安定して韻律が機械学習できるようにしている。第２は、日本語に特化した韻律のモデルである。韻律モデルは、収

[1] （株）エーアイ
京都府相楽郡精華町光台 2-2-2 ATR 内 W 棟２階
[a)] hirai@ai-j.jp

録量を増やせないことから、できるだけ推定するパラメータを少なくし、コーパスから無駄なく安定して学習できることを目的に作成した。たとえば F0 は、日本語アクセントの特徴を考慮しアクセント句毎に、その概形を表現するために最低限必要な数点のみを予測している。また、韻律の予測器には、アンサンブル予測を用いている。1つの韻律に対し異なる表現の複数のパラメータで予測し、複数のパラメータの尤度が最大になるように韻律を推定する。例えば、継続時間長では、子音と母音の時間長だけでなく、モーラの等時性を考慮しモーラ時間長も予測し、尤度が最大となうように子音と母音の時間長を決定している。その結果、少ないコーパスから安定して韻律が生成可能となった。韻律にかぎると、単純なナレータ読みであれば、５０文章程度から再現が可能である。

3. AITalk の活用事例

これまでに行ってきた AITalk を用いた事業をいくつかを紹介する。

3.1 防災・広域放送システム

弊社の柱となっている事業分野の 1 つである。屋外に設置されたスピーカから、広域に情報を提供するシステムの音声生成に利用されている。J-ALERT(全国瞬時警報システム)[2] の基幹システムに AITalk が採用されたことで各自治体への採用が広まった。J-ALERT は、緊急時の情報を政府が伝えるシステムである。誤りを防ぐため中間言語を各自治体に発信し、端末で音声が合成される。定型文以外にも任意の文章を瞬時に間違いなく作成する必要があるため、中間言語を容易に作成できる GUI を提供している。また、J-ALERT 以外の自治体の広域放送、駅構内のアナウンスなどでも利用が徐々に広がっている。音質としては、ゆっくりとした発話で、かつ明瞭に伝わることが重要である。近年では、全体的な合成音声の品質の向上により、音質より、音声作成の容易さや外国語の作成などへの要望が多い。

3.2 CTI

CTI（Computer Telephony Integration System）における音声自動応答は、古くから音声合成の重要なアプリケーションの 1 つである。一般電話の利用が減少していることから、将来は市場が縮小することも予想されるが、現状では、依然として電話での問い合わせなどは存在するため、CTI での音声合成の利用は減ってはいない。ただし、新規に自社でコールセンタを立てるなどは少なく、外部の Web リソースを利用するなど、コスト削減の傾向にある。チャットボットのような対話のシステムへの移行が期待される分野であるが、音声応答の必要性については懐疑的な意見もある。

3.3 NAVI

車載用のナビゲーションは、古くから音声の利用が重要とされる分野である。しかし、音声を用いた道案内に限ると、収録音声で対応が可能であるため、再収録のコスト削減用途以外では、音声合成を利用しない場合が多い。しかし、各車メーカともに、今後はよりインテリジェントな対話を目指しており、音声合成の利用が広がる可能性が高い。ただし、自然で多様な受け答えなど、品質に対する要望も高く、さらなる改良が必要であると思われる。

3.4 音声対話

上記のチャットボットや次世代のカーナビを含め、機械との音声対話は、今後広まることが予想される分野である。最近では、スマートスピーカの出現で、メディアに露出する機会も多く、一般の方々に広く浸透してきている。しかし、音声合成の市場はそれほど広がってはいない。

3.4.1 スマホ用アプリ

古くは、組み込みエンジンを用いて、携帯電話などに音声合成を組み込んでいたが、現在は、WebAPI を用いてサーバでサービスを提供することが多い。その結果、様々なサービスの提供が可能となっている。AITalk の強みを活かした利用形態の 1 つが、docomo のしゃべってキャラ [3] である。このサービスでは、有名人やアニメのキャラクターなど、数百種類の中から利用者が希望する声で、様々なサービスを得ることできる。

3.4.2 ロボット

カスタム音声の応用例として、AITalk はマツコロイド [6]、ペッパー [5]、漱石アンドロイド [4] などロボットにも広く利用されている。ロボットでは、単純な文字情報の出力だけでなく、キャラクターにあわせた発話スタイルで話す必要がある。現状の AITalk の言語処理は、ナレーション読みを想定したものである。よって、淡々と発話する漱石アンドロイドでは、テキストから自動で作成した音声で問題はないが、ペッパーや、マツコロイドでは、中間言語を手動でチューニングした音声を利用している。また、マツコロイドは TV 番組の中で充分楽しむことのできる漫才を行っているが、中間言語を修正しない音声では全く面白くないものであった。魅力ある発話スタイルを再現する音声合成器として広く利用される為には、表現力豊かな中間言語を、より簡易に作成できるようにする必要がある。

3.5 ウェブキャンペーン

AITalk カスタム音声を一番はじめに多くの方々に利用していただいたのが、企業の Web アクセスを上げるためのキャンペーンである。有名俳優がユーザーの希望する文章を読み上げると言ったサービスであった。Web アクセスの増加には非常に効果があった。現在でも同様の企画を持ち込まれることはあるが、キャラクターが重要で、音声収

録に関わる金銭的な問題で実現することは少ない。

3.6 ゲーム

ゲーム内で TTS を利用したいという要望はある。しかし、要求される品質が高く、限られた分野でしか利用されていない。AITalk は、名前だけの読み上げや、幼い口下手な子供キャラなどに利用されている。

3.7 放送メディア

アナウンサー代わりのニュースの自動読み上げに利用されている。人件費削減が目的で導入されることが多いが、日本語は読み違い無しに完全自動化することが難しく、言語処理結果の確認や修正に手間がかかり、継続利用が少ないのが現状である。言語処理の向上が、市場拡大に繋がる可能性がある分野である。

3.8 教育

社内向けの E ラーニングなどへの利用が増えている。学習用のコンテンツを内製する傾向にあり、ナレーションの代わりとして利用されている。

3.9 パッケージソフト

VOICEROID[7] というコンシューマ向けの製品を OEM で提供している。単価が安いため売上高はそれほど大きくないが、順調に売上は伸びている。主に、ゲーム実況など自分で動画を作成するときのナレーションに利用されている。いわゆる感情音声（喜び、悲しみ、怒りなど）が利用される分野である。GUI で音を作ることが前提なので、言語処理はあまり重要視されないが、「ぎゃーーーっ」など、通常では発声しない独特な読みへの対応が要望されている。

3.10 組み込みエンジン

おもちゃが話すなどの一部の用途を除き、AITalk はこの分野ではあまり広く利用されていない。組み込み用のエンジンでは、言語処理の品質が劣るため、定型文はそのまま読み上げ、自由文は、サーバで中間言語を作成するなど、品質の向上には工夫が必要である。

4. 今後の取り組み

近年、機械学習技術の進歩により、比較的大規模なコーパスがあれば、音声合成の詳細な知識や辞書作成のノウハウが無くても、高品質の合成音声が得られることが明らかになっている。今後も音声合成の事業を発展させていくには、音が良いだけでなく、差別化できる技術・サービスが重要である。

現在、注目しているのは、機器との対話向けの音声合成である。自動で読み間違えなく読み上げる技術や、人間が話し易いように自然に表現豊かに応答する技術が進むことで、市場が拡大することが期待できる。

表現豊かに話すためには、多様な発話スタイルを表現できる中間言語を適切に出力する言語処理と、その中間言語を忠実に再現する 音声合成エンジンが必要である。さらに、弊社のサービスとして考えた場合、カスタム音声での実現が必要である。

現在、専業メーカーの強みであるラベラーによる正確なラベリング技術と DNN などの機械学習を利用することで、これらを実現する技術を開発中である。

参考文献

[1] http://www.ai-j.jp/
[2] https://ja.wikipedia.org/wiki/全国瞬時警報システム
[3] https://www.nttdocomo.co.jp/service/shabette_concier/shabette_chara/
[4] http://naturaleight.co.jp/matsukoroid/
[5] http://www.softbank.jp/robot/special/shiratoke/
[6] http://www.nishogakusha-u.ac.jp/android/index.html
[7] http://www.ah-soft.com/voiceroid/