Vol.2010-SLP-83 No.8
2010/10/29

音声合成の利用シーンと要求される品質との関係

三井康行†

音声合成技術の利用シーン毎に，要求される品質を満たすために重視される評価項目が異なっている．これについて，実際に提供されているサービスのうち，「電話自動応答」と「キャラクタ音声合成サービス」を例にとって紹介する．サービスの提供を通して得た知見を元に，「電話自動応答」では「明瞭性」，「韻律の自然性」，「肉声感」が重要であるのに対し，「キャラクタ音声合成サービス」では「キャラクタ性」と「韻律制御の自由度」が重要な評価項目であった．

Various Quality Measures for Text-to-Speech among Usage Scenes

Yasuyuki Mitsui†

There are various quality measures in evaluating Text-to-Speech engines among its usage scenes. As an example, we introduce two business scenes and their feedbacks by its users. One scene is "Interactive voice response system", and we describe that the important measures in the scene are "clearness of voice", "naturalness of prosody" and "similarity with human voice". The other scene is "Amine character speech synthesis service", and we found that the important measures in the scene are "similarity with Anime character" and "flexibility of prosodic control" in contrast.

1. はじめに

テキスト音声合成（Text-to-Speech：TTS）技術は古くから研究され，進歩を遂げてきた．特に近年は，大規模コーパスを利用した音声合成技術（コーパスベース音声合成）が登場し，音声合成は「人間らしさ」を獲得してきた．

「人間らしさ」を表す 1 つの指標である，テキストに対する「読み上げの自然性」に関して，一定以上の品質を獲得した音声合成は，商用目的での利用シーンを拡大してきた．それにつれて，音声合成はテキスト読み上げの自然性以外にも複数の評価項目があり，利用シーン毎にエンドユーザが重視する評価項目が異なっていることが分かってきた．音声合成が人間に匹敵する表現力や柔軟性を獲得するには，依然多くの技術的課題が存在しているため，1 つの音声合成エンジンで全てのシーンに対応することは困難である．このような現状では，「読み上げの自然性」という項目だけでは，それぞれの利用シーンにおけるエンドユーザの求める品質を満たすことができるかを評価できない．したがって，実際に商用等で利用する場合には，それぞれのシーンで要求される品質を満たすために，どのような評価項目を重視すればよいかを検討することが，エンドユーザの満足度を向上させる上で非常に重要となる．

本稿では，音声合成技術が利用されているシーンにおいて，エンドユーザが要求する品質を満たすために，どのような評価項目が重視されているかについて，実際に提供されているサービスで得られた知見を例にとって紹介する．

2. 音声合成の利用例

音声合成技術は様々なシーンで利用されているが，その中でも音声の収録コストや話者を確保できないリスクを低減させるために，人間の音声を代替する手段として利用されることが多い．例えば，ラジオでの株価読み上げや web 上で提供される新聞記事の読み上げ，音声合成が利用されてきた．近年では，携帯オーディオプレイヤーといった大きな画面を持たない端末上で，音声配信されたブログや SNS 等の記事を音声で聴取するといった新たな楽しみ方が提案されている．電車やバスといった公共交通機関のアナウンスについても，現状では録音音声が主に使われているが，近い将来合成音声に置き換わっていくと考えられる．

また，自動車の運転や工場等での作業のように，視覚を作業に集中させる必要がある環境では，聴覚を用いた作業指示や確認が望ましく，このような場合の音声による情報提供を目的とした音声合成システムの導入も進められている．同様に，視覚障がい者による PC の補助アプリケーションとしても，従来から音声合成による操作内容

† NEC　情報・メディアプロセッシング研究所
Information and Media Processing Labs., NEC Corp.

の読み上げが利用されてきた．

さらに，コミュニケーションロボットに代表されるような，対話型ユーザインターフェース等においても，従来は録音音声によるガイダンスが主であったが，音声合成の導入によりコストを抑えて導入することができるようになった．近年では，デジタルサイネージ等において，広告を視聴する人に合わせて内容をリアルタイムに変更できる広告のスタイルが注目されており，ここでも柔軟に文言変更ができる音声合成への期待が高まっている．

3. 実際のサービス例

実際に提供されているサービスのうち，「自動電話応答（Interactive Voice Response：IVR）」および「キャラクタ音声合成サービス」を紹介する．

3.1 自動電話応答

IVR とは，企業等のカスタマーサポートや商品予約販売等，電話を窓口とした業務を，音声による自動応答で行うシステムである．従来は収録音声によるガイダンスが主だったが，近年は音声の新規登録にかかる音声収録コストを低減させるため，合成音声に置き換わりつつある．我々は実際に音声合成機能を搭載した IVR システムを提供しており [1]，商品予約販売等のサービスに利用されている．

3.2 キャラクタ音声合成

キャラクタ音声合成サービス [2), 3)] とは，特定のアニメキャラクタの音声をエンドユーザが自由に作成できるサービスである．このサービスでは，実際にアニメキャラクタを演じる声優の音声を収録し，データベースを作成することで，キャラクタの声質や韻律の特徴を音声合成システムで再現する．エンドユーザは任意のテキストを入力でき，あたかもアニメキャラクタに好きなセリフを喋らせている感覚を得る．本サービスは web 上で提供されており，エンドユーザは web ブラウザを介して，プラットフォームに依存せずにサービスを享受できる．また，漢字の読みやアクセント情報，ローカルなテンポやピッチ周波数パタンを編集できる機能を備え，エンドユーザはこれらを自由に変更できる．具体的には，図 1 に示すフローのように，エンドユーザは試聴と編集を繰り返しながら，希望通りの音声を生成する．さらに，図 2 に示すような，直感的な編集が可能となる GUI をエンドユーザに提供した．

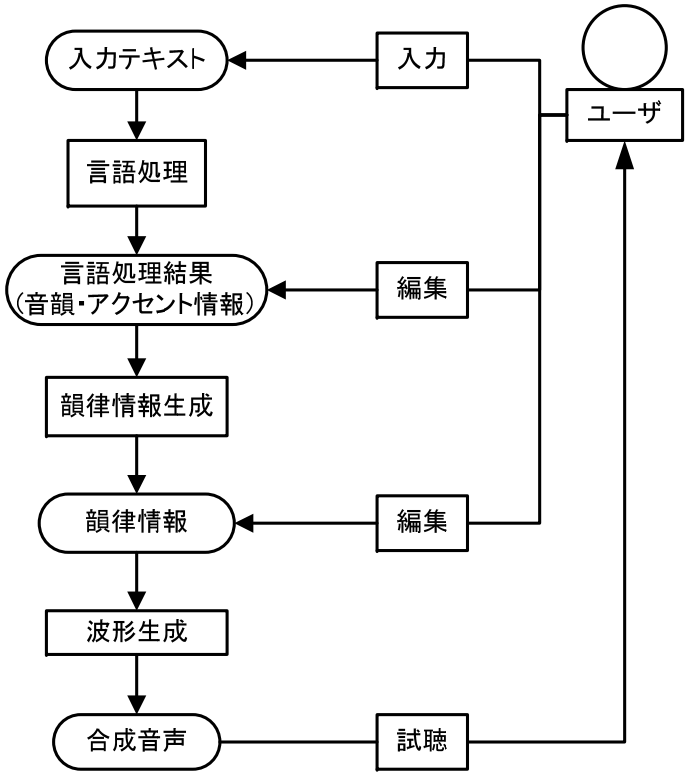

図 1．キャラクタ音声合成サービスにおける処理フロー

Vol.2010-SLP-83 No.8
2010/10/29

図 2．編集用 GUI

4. 得られた知見

本章では，前章で紹介した 2 つのサービスについて，実際に提供して得られた知見と，それを元に重視される評価項目について説明する．IVR は従来からサービスとして提供しているため，重視される評価項目は概ね判明している．一方，キャラクタ音声合成サービスは新しいサービスであるため，一般的には知見が得られておらず，重視される評価項目も分かっていなかった．そこで，商品予約販売を目的とした IVR とキャラクタ音声合成サービスにおいて重視される評価項目の違いについて検討した．それぞれのサービスにおいて重視されるべき評価項目とその重要度を表 1 に示す．表 1 では，「◎」は最も重要な評価項目，「○」は重要な評価項目，「△」は重要度の低い評価項目を示す．

表 1．実際のサービスで重視される評価項目

評価項目	重要度	
	IVR	キャラクタ音声合成サービス
明瞭性	◎	△
韻律の自然性	○	—
肉声感	○	—
キャラクタ性	—	◎
韻律制御の自由度	—	○

4.1 自動電話応答

同目的における IVR において，特に重視される項目は，「明瞭性」，「韻律の自然性」，「肉声感」である．この中でも，最も重視されるのは「明瞭性」である．同目的で IVR を利用するエンドユーザにとって最も重要な情報は，購入を希望する商品名や数量，エンドユーザ自身が入力した電話番号，住所といった情報である．これらの情報は，オペレーターが対応する場合と同様に，合成音声により復唱され，エンドユーザが確認する．したがって，これらの情報について，内容がはっきりと聞き取れることを評価する「明瞭性」が最も重要な評価項目となる．

次に重要となるのが，「韻律の自然性」と「肉声感」である．同目的における IVR では数分～10 数分といったある程度長い時間のやり取りがされる．このため，日本語として不自然な発声や機械的な音声がガイダンスに含まれると，内容が聞き取れたとしても，エンドユーザが不快に感じ，疲れてしまうという報告があった．したがって，「韻律の自然性」と「肉声感」が重要となる．

4.2 キャラクタ音声合成サービス

キャラクタ音声合成サービスにおいて，エンドユーザが最も重視するのは対象となるキャラクタの特徴が正しく再現されているか否かである．本サービスでは，キャラクタ性（キャラクタの声質および韻律に関する特徴の再現性）を再現することを重視し，二段単位選択を用いた音声合成エンジン [3), 4)] を採用している．これにより，キャラクタ性について高い評価を得て，多くのエンドユーザを獲得した．この結果から，「キャラクタ性」が最も重視される評価項目であることが分かる．

逆に，キャラクタの特徴が合成音声に反映されていなければ，エンドユーザにとって価値のないサービスとなってしまう．このため，キャラクタ性の確保のためには，他の評価項目を犠牲にすることも検討する必要がある．実際，本サービスで提供しているキャラクタの中には，スペクトルのランダム成分を抑える処理をすると声質がオリジナルと大きく変わってしまうキャラクタがあった．一般的にランダム成分が強調

されると明瞭性が低下するため，我々はランダム成分を抑える方式を採用している．そこで，本サービスとして「キャラクタ性」と「明瞭性」のどちらを重視すべきかを検討するために，通常通りランダム成分を抑える方式（方式 A）と，抑えない方式（方式 B）の 2 種類を作成して，聴取による主観評価を行った．被験者は，研究員とサービス開発者の併せて 4 名とした．その結果，方式 A の合成音声は，明瞭性は高いもののキャラクタ性が確保できていないという結論を得た．それに対し，方式 B は明瞭性で劣るものの，キャラクタ性が十分に出ていたため，本サービスにおいては，方式 B を採用した．

このように，IVR で重視された「明瞭性」は，本サービスでは「キャラクタ性」に依存するものであり，必ずしも重視されない．さらに極端な話をすれば，再現対象となるキャラクタが，滑舌が悪く，明瞭性が極端に低いという特徴を持っている場合は，生成される合成音声の明瞭性も低いことが求められる．つまり，キャラクタ音声合成サービスにおいては，IVR と同様に明瞭性を最も重視するような音声合成システムでは，必ずしもエンドユーザに訴求することはできないと言える．

加えて，キャラクタ音声合成サービスでは，「韻律制御の自由度」も重視されることが分かった．アニメ作品に登場するキャラクタの場合，エンドユーザが想定するシチュエーションによって，ローカルなテンポやピッチ周波数パタン等の韻律情報が異なる．そのため，入力されたテキスト情報だけでは，期待した音声が生成されない可能性がある．例えば，「おはよう」という音声を生成する場合に，明るく元気な雰囲気とするか，あるいは暗く寝ぼけたような雰囲気とするかはシチュエーションに依存する．このような課題に対応するには，韻律情報を変更して，エンドユーザが望む韻律へと変更できる機能が必要である．具体的には，明るい雰囲気にしたい場合は，グローバルなピッチを高めに変更し，暗い雰囲気にしたい場合は，グローバルなピッチを低めに，テンポを遅めに変更する．そこで，前述した GUI（図 2）をエンドユーザに提供して，生成された合成音声の韻律情報を編集できるようにした．

さらに，ローカルなテンポおよびピッチ周波数のパラメータは，通常の発声ではありえない帯域まで変更できるように，十分なダイナミックレンジを設定した．これは，エンドユーザに幅広い楽しみを見出してもらうことを目的としている．例えば，作品の中では落ち着いている性格を持つキャラクタに，慌てて裏返った声を出させるといった楽しみ方が，これによって可能となっている．実際，我々が集計した結果によると，エンドユーザの大半が提供した GUI を使って自動生成された合成音声に対して編集を加えていることが判明している．また，エンドユーザは自由に編集することにより，通常の読み上げにはない様々な楽しみ方を試みていた．したがって，エンドユーザの満足度を向上するために，「韻律制御の自由度」の確保が重要であることが分かる．

また，サービスを開始した後には，実際に使ってみたエンドユーザから，「テンポをもっと細かく設定したい」「韻律の制御点を増やしてほしい」といった要望が寄せられた．図 2 に示した GUI はこれらの要望に応えた結果を反映したものであり，韻律制御の自由度をさらに高めることにより満足度を向上させることができた．

5. 考察

本稿では，2 つのサービスに着目し，対比することで重視される評価項目を明らかにしたが，さらに多様な音声合成の利用シーンを想定すると，他にも重視すべき評価項目が存在すると考えられる．さらに，より詳細な分析を行うためには，例えば「明瞭性」や「キャラクタ性」等の評価項目を細分化する必要がある．これらを明らかにすれば，ユーザやサービス提供者がどんな判断基準で音声合成の採用/不採用を決定しているかを，定性的あるいは定量的に判断できるようになる．その結果，エンジン開発者にとっては，音声合成エンジンの改良の方向性が明確になるため，必要な項目を満たすエンジンを開発，提供することができる．また，サービス提供者にとっても，評価項目毎の詳細なベンチマークが可能となるため，製品やサービスに最適な音声合成エンジンを選択することができるようになる．

6. おわりに

音声合成技術の利用シーン毎に，要求される品質を満たすために重視される評価項目が異なっている．これについて，実際に提供されているサービスのうち，「電話自動応答」と「キャラクタ音声合成サービス」を例にとって紹介した．サービスの提供を通して得た知見を元に，「電話自動応答」では「明瞭性」，「韻律の自然性」，「肉声感」，「電話帯域制限の影響」が，「キャラクタ音声合成サービス」では「キャラクタ性」と「韻律制御の自由度」が重要な評価項目であることが分かった．

今後は，評価項目の詳細化や，他の利用シーンにおける評価項目の検討等を行っていく予定である．

参考文献

1) Voice Operator, http://www.nec.co.jp/middle/VoiceOperator/
2) コエラボ, http://voice.biglobe.ne.jp/
3) 三井康行, 近藤玲史, 加藤正徳, 杉浦淳: 二段単位選択による音声合成のキャラクタ表現への適用, 音講論(春), 1-22-21, (2008).
4) 加藤正徳, 近藤玲史, 三井康行: 二段単位選択を用いた高品質音声合成, 音講論(春), 1-11-22, (2008).