Vol.2012-SLP-93 No.7
2012/10/26

音声合成の多様性向上の取り組み

籠嶋 岳彦，橘 健太郎，布目 光生，森田 眞弘[†1]

本稿で言う音声合成の多様性とは，様々な話者や発話スタイルの音声合成システムが簡単に作れる能力を指している．話者の多様性向上のため，収録音声からその話者性を再現する音声合成モデルを自動生成するシステムを開発し，サービスを一般公開した．本サービスにより，有名人の声だけでなく，開発コストがかけられない一般ユーザの声での音声合成が可能となった．音声合成の応用を拡大していくためには，発話スタイルがそれぞれの応用に適していることが必要である．これまでに，音声インターフェース応用で有用な「対話調」などを開発してきた．さらに，感情をこめた電子書籍の朗読を実現するために，セリフの感情を自動判別するシステムを試作した．

Improvements on Variety of TTS

TAKEHIKO KAGOSHIMA, KENTARO TACHIBANA, KOSEI FUME and MASAHIRO MORITA[†1]

This paper shows our development enabling to build a variety of voices and speaking styles easily. To realize a wide variety of speakers, we developed an automatic voice building system and opened a custom voice service. Any user can build his/her own voice using the service. In order to expand applications of TTS, its speaking style should be adjusted to each application. We have developed some valuable speaking styles such as "dialogue style" for speech interface applications and "emotion styles" for e-book reading. An emotion estimation method from input text was also developed to enable TTS to read novels with suitable emotions.

1. はじめに

フォルマント合成器に代表される初期の音声合成では，音韻性の再現，言い換えれば了解性の向上が最大の課題であった．その後，実際の音声信号に基づいて，音素や音節，CV/VC などの単位のパラメータや波形（音声素片）を接続する素片接続型の音声合成により，了解性は実用上の問題ではなくなった．しかしながら，韻律の変形や接続のひずみによる音質の劣化や，ルールベースの韻律制御による機械的な抑揚により，合成音声の自然性は十分ではなく，音声合成の用途は限定的であった．応用を拡大するためには，自然性の向上が次の課題となった．この自然性向上の原動力となったのが，コーパスベース音声合成の発展である．大量の音声データを用いた統計処理により，合成音声の韻律の自然性や肉声感の向上を実現し，カーナビの音声案内や，コールセンターの音声応答など，様々な用途で音声合成が利用されるようになってきた．

コーパスベース音声合成は，基になる音声コーパスを手本とし，そのスペクトルの変化や基本周波数のパターンを再現する音声合成モデルを学習することで，自然な合成音声を実現しようとするものである．そのため，日本語として自然な音韻や抑揚というだけでなく，音声コーパスの話者性や発話スタイルをも再現することになる．特定個人の話者性の再現が可能になると，タレントや有名人の音声合成のニーズが，エンターテインメント応用で顕在化するようになってきた．基本的に，音声コーパスのサイズ（収録時間）が大きくなるほど合成音声の自然性は向上し，話者の再現性も改善するが，タレントを何週間もスタジオに拘束するというのでは，ビジネスとしては成立しない．例え収録ができたとしても，音声データへの音素ラベリングや F0 抽出などを大量のデータに精度良く行うには，開発コストと期間の問題がある．

同じ話者でも状況によって話し方を変えたり，感情によって話し方が変わったりする．このような発話スタイルのバリエーションもコーパスベース音声合成によってある程度再現が可能である．しかし，話者を追加する場合と同様に開発コストと期間の問題があるため，どのようなスタイルを用意し，どうやって使い分けるかが課題である．

本稿では，上述した話者性や発話スタイルの多様性向上を実現するための取り組みについて述べる．話者性については，一般ユーザの音声から，その話者の音声合成モデルを全自動で作成するサービスについて紹介する．また，発話スタイルについては，電子書籍を感情のこもった音声で朗読するシステムの試作について概説する．

2. 音声合成の多様性

2.1 多様性とは

本稿で言う多様性とは，様々な話者や発話スタイルの音声合成システムが簡単に作れる能力を指している．上述したとおり，コーパスベース音声合成は，コーパスの話者性や発話スタイルを再現しようとするものであるから，多様性向上の１つのアプローチとして，再現性を維持・向上しつつ開発コストや期間を削減するという方向性がある．本稿では，このような話者や発話スタイルの再現による多様

†1 (株)東芝　研究開発センター
Toshiba Corporation Corporate R&D Center

性の向上について述べる．

別のアプローチとして，可制御性の向上という方向性がある．何らかの方法で話者性や発話スタイルを制御可能とすることにより，目標となる話者やスタイルの音声が収録できない場合や存在しない場合でも，所望の話者やスタイルの音声を合成する技術である．

2.2 合成音声の多様性のニーズ

特定の個人の声で音声合成させたいというのが，話者性のカスタマイズの典型的なニーズである．タレントや声優，アナウンサーなどの声の音声合成が，ゲームソフトや Web 上のプロモーションコンテンツなどで利用されている．また，固定メッセージは録音音声を利用し，可変部分を音声合成して接続するような応用では，固定メッセージに話者性を合わせる必要がある．その他に，喉頭摘出などで声を失う人が，自分の声を音声合成で再現できるように残しておきたいというニーズもある．

従来，音声合成の発話スタイルの基本は，文章を淡々と読み上げる調子（読み上げ調）であった．発話の内容を伝えるだけの目的であれば，読み上げ調は聞き取りやすく，特に問題は無い．しかしながら，読み上げ調として自然な音声合成だとしても，これを対話的なユーザインターフェースなどに用いると，たちまち「暗い」「冷たい」「事務的」など，ユーザからネガティブな反応が返ってくる．合成音声がユーザに受け入れられるためには，発話スタイルが用途に合っていることが重要である．話者性についても同様で，デフォルトの成人男声１名，女声１名だけでは，かわいらしいアニメキャラクタやロボットなどの声には対応できず，ビジュアルに合致する話者性が要求される．このように，音声合成の応用を拡大していくためには，用途に応じて話者性や発話スタイルをカスタマイズしていくことは避けて通れない．

3. 話者と発話スタイルの再現性

3.1 話者性

声質の再現性を優先するならば，韻律変形を極力廃した素片選択型の合成器が有利であるが，音声コーパスのサイズが小さくなると，音韻・韻律共に不連続感が増して自然性が低下するという問題がある．これに対して，多様性を向上させるため，小さい音声コーパスでも実用的な自然性・話者性を確保することが可能で，スケーラブルな音声合成方式を筆者らは提案してきた[1][2][3]．複数素片選択融合方式は，合成単位あたり複数の音声素片をコーパスから選択し，素片融合処理によってそれら素片の平均的な特徴を持つ音声素片を生成して接続するものである．融合する素片数を増加させると，肉声感は若干低下するものの，安定した均質な合成音声が得られるようになる傾向がある．この融合素片数をチューニングすることにより，比較的小さい音声コーパスでも自然性と話者性のバランスがとれた合成音声を生成することができる．

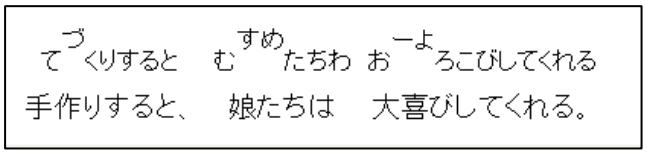

図 1 収録原稿と読み・アクセントの提示

これまでに，声優・タレント・アナウンサーなど，様々な有名人の音声合成を，顧客の求めに応じて開発し，Web 上のキャンペーンやビデオゲーム，放送，組み込み機器などで利用されてきた．収録した音声コーパスのサイズは，用途や予算，開発期間に応じて，10 分程度から数時間と様々である．

3.1.1 一般ユーザ向けカスタム音声合成サービス

有名人のカスタム音声合成は，収録の期間やコストを削減できたとしても，品質管理などを考慮すると，ある程度の開発工数が必要である．もし，音声収録から音声合成モデルの学習までを全自動で行うことが可能となれば，一般のユーザが手軽に自分の声の音声合成モデルを作成して利用することができるようになる．そうなれば，音声合成モデルを家族や友人と共有したり，一般に公開したりするなど，新たな利用方法が考えられる．このような一般ユーザの声による音声合成を実現するため，全自動の音声合成モデル作成システムを開発し[4]，一般ユーザが利用できるサービス[5]を公開した．本サービスでは，音声合成で作成したメッセージをグリーティングカードとしてメールで送ることができる．あらかじめ用意されたナレータやタレントの声に加えて，ユーザが作成して公開した声の中から選んで，合成音声や歌声のグリーティングカードを作成することができる．

本サービス実現における技術的な課題は主に以下の 3 点である．

- 収録時間の削減
- データ処理の精度向上と全自動化
- 収録環境に対するロバスト性向上

一般ユーザがそれほどの負担を感じることなく試すことができるようにするため，収録に要する時間は，30 分から長くても 1 時間程度までに抑える必要がある．収録時間は，主に収録原稿の長さに依存する．収録時間と合成音声品質のバランスを考慮し，予備実験を行って約 90 文の音韻バランス文セットを作成した．本文セットで収録した音声データの長さは，平均的な発話速度で 10 分間程度である．

収録した音声に対して，音素ラベリングやコンテキスト情報の作成を行う場合，音声から音素列やアクセント型を自動抽出するのは困難であるため，精度を上げるためには目視による確認と修正が必要である．今回は，全自動化が必須であることから，あらかじめ用意した音素列やアクセント型のデータに，ユーザの発声を合わせてもらうことで解決を図った．具体的には，(１) 読み・アクセントの表示，

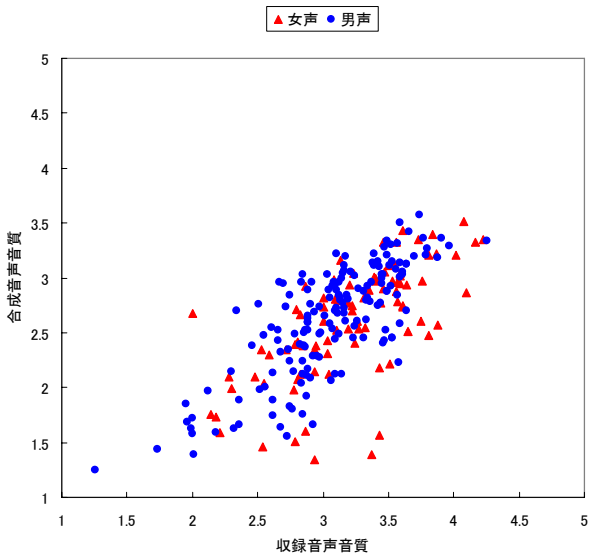

図 3 収録音声音質と合成音声音質の関係

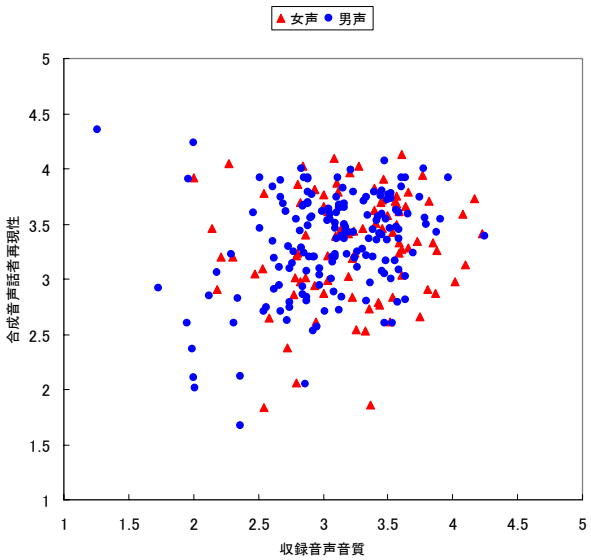

図 2 収録音声音質と合成音声話者再現性の関係

（2）お手本音声の提示，（3）収録音声のセルフチェックをサービスに組み込んだ．読みとアクセントは，図 1 に示すように，読みを表す文字の位置で高低アクセントを表示した．さらに，ナレータのお手本音声を，収録の前に 1 文毎に自動再生し，視覚的・聴覚的に正解を意識させるようにした．また，録音停止後に収録音声を自動再生し，読み誤りやノイズの混入などを自らチェックして，問題があれば再収録できるようにした．

本サービスでは，ユーザが各自のＰＣから音声収録を行うため，マイクなどの収録機器や周囲の環境などを統制することができず，収録音声の品質のバラツキが大きいことが予想された．事前に小規模な実験を行った結果，雑音除去と振幅の正規化処理により，多くの場合で収録音声の品質が改善され，副作用は比較的小さいことが分かった．そこで，収録環境に対するロバスト性向上のため，これらの処理を収録音声の前処理として導入することとした．

3.1.2 音質と話者再現性の評価

本サービスによって，2011 年 12 月のオープン以来約 9 ヶ月で 700 名以上の一般ユーザが自分の音声合成モデルを作成した．収録に要する時間は，ほとんどの場合で当初想定した 30 分から 1 時間程度の範囲となっている．本サービスのユーザ 256 人分について，収録音声と本サービスで生成された合成音声[†2]とを用いて以下の 3 種類の主観評価実験を行った．

- ◆ 合成音声音質：合成音声を提示してその音質・自然性を評価する 5 段階ＭＯＳ評価．
- ◆ 合成音声話者再現性：収録音声と合成音声（発話内容は異なる）を対で提示し，話者性の類似度を評価する 5 段階ＭＯＳ評価．
- ◆ 収録音声音質：収録音声を提示して，その音質・自然性を評価する 5 段階ＭＯＳ評価．

いずれの評価試験でも，話者あたりの評価文章は 10 文章，1 文章あたりの評定者は 25 名とした．なお，上述した 3 種類の評価は独立に実施しており，異なる種類の評価値（例えば収録音声音質と合成音声音質）を比較することはできない．図 2 に，収録音声音質と合成音声音質の関係を示す．相関係数は 0.75 であり，比較的強い相関があることが分かる．男声・女声では，顕著な傾向の違いは見られない．この結果からは，今後さらに音質を向上させるためには，収録音声の品質を上げることが必要であると考えられる．また，図 3 に収録音声音質と合成音声話者再現性の関係を示す．合成音声音質の場合と異なり，話者再現性と録音音声音質との相関は 0.21 と低い値である．録音音声音質よりは，声の個性の強さが支配的な要因と推測される．

3.2 発話スタイル

同じ話者でも，状況によって話し方が異なる．そのような話し方（発話スタイル）のバリエーションは，主に以下のような要因の影響を受けると考えられる．

- ◆ 話し手と聞き手の関係
- ◆ 発話の目的
- ◆ 話者の感情

例えば，音声合成のベースラインとなっている「読み上げ調」の発話スタイルは，感情は平静，意図は無く，聞き手は存在しないか不特定多数のモノローグであるとみなすことができる．発話スタイルのバリエーションは無限にありうるが，発話スタイルの拡張にはコーパス収集などのコストがかかることから，優先順位をつけて有用なスタイルから取り組む必要がある．

3.2.1 話者と聞き手の関係

読み上げ調のスタイルが適切でない応用の一つに，上述

†2 評価に用いた収録音声および音声合成モデルは，個人を特定できない形で抽出して使用した．

した対話インターフェースがある．特定の相手に対して話しかける調子は，明らかにモノローグとは異なっている．エンターテインメント的な使い方を除けば，ユーザビリティの向上を目的に導入される音声インターフェースにおいては，一般に特定の相手に礼儀正しく話しかける調子が望まれる．話し手である機械と聞き手であるユーザの関係は，人間同士に例えると，店員と顧客の関係に類似している．そこで，店員から顧客への話しかけを想定した音声コーパスを収集して「対話調」の音声合成を作成した．読み上げ調では不自然に感じられる挨拶や問いかけも，対話調では違和感が小さく，インターフェース用途に適している．

3.2.2 発話の目的

発話の目的と発話スタイルが整合していることも重要である．例えば，もし謝罪を目的に読み上げ調で発声したならば，目的が達成されないどころか逆効果になるかもしれない．このように，言語情報とパラ言語情報が矛盾している場合，言語情報で伝えたいメッセージが正しく伝わらない危険性がある．音声合成の応用としては，例えば防災無線放送などで，発話スタイルの制御が必要になる可能性がある．このような応用を想定して，緊急性の高いメッセージを伝えるための発話スタイル（警告調）の音声合成を試作した．直ちに避難を促すようなメッセージの場合，読み上げ調では緊急性が伝わりにくいが，警告調であれば，伝えたいメッセージを直感的に理解することが可能である．

3.2.3 話者の感情

上述した発話スタイルのバリエーションは，音声合成の利用シーンと対応付けられるため，適切な発話スタイルを選択して使い分けることができる．これに対して，怒り・悲しみ・喜びなどの感情のバリエーションは，１つの利用シーンの中で，入力テキストの１発話毎，あるいはより短い単位で使い分ける必要がある．例えば，アニメーションの動画にセリフをつける場合のように，ある程度工数をかけて作りこむことが可能であれば，手作業で適切な感情を選択することができる．しかしながら，電子書籍を読み上げるような利用シーンの場合は，手作業による感情ラベル付けは非常にコストがかかり，現実的ではない．

音声合成による電子書籍の読み上げは，既に実用化されているが，発話スタイルは読み上げ調が一般的である．ビジネス本では，発話スタイルの点での問題は少ないが，小説ではセリフ部分での違和感は否めない．そこで，小説のセリフ部分に自動で感情のラベル付けを行って，感情を込めて読み上げる電子書籍ビューアを試作した[6].

感情のラベルは，平静に加えて，怒り・悲しみ・喜びの 4 種とした．感情ラベルの推定は文単位で行った．入力文に対して，各感情の推定モデルを用いてスコアを算出し，最もスコアの高い感情を当該文の感情ラベルとした．感情推定モデルとしては，メンテナンスの容易さや拡張性を考慮して，ナイーブベイズに基づく手法を採用した．

提案手法では，ある文 s が与えられた場合に，感情 c のスコア $E_c(s)$を次式で求める．

$$E_c(s) = P(c)P(w_1|c)P(w_2|c) \ldots P(w_n|c)$$

P(c)：感情 c の出現確率

$w_1 \ldots w_n$：文 s を構成する各単語

n：文 s に含まれる単語数

P(w|c)：感情 c が与えられたときの単語 w の出現確率

ここで，P(c)および P(w|c)は，文単位の感情ラベル付けを手作業で行ったテキストコーパスから学習するものとする．書籍数十種類から抜粋したセリフ約 3000 文を対象に，モデル学習と精度評価実験を行った．その結果，全体の約 2/3 のデータを用いて学習し，全データで評価したところ，適合率は約 0.9 であった．

試作した電子書籍ビューアは，地の文とセリフで話者を切り替え，さらにセリフの部分は推定結果に基づいて感情を切り替えて音声を合成する．これらの機能により，文面を見ずに音声を聞くだけで内容が理解しやすく，違和感を軽減した読み上げが可能となっている．

4. おわりに

本稿では，音声合成の話者や発話スタイルの多様性向上のための取り組みについて述べた．比較的小さい音声コーパスでも自然で話者性を保持した音声の合成が可能な複数素片選択融合方式を開発し，データ処理の自動化などにより，一般ユーザの声で音声合成が可能なサービスを開発した．発話スタイルのバリエーションとしては，実用性を考慮して，対話調や警告調などの音声合成モデルを開発した．また，電子書籍の感情を込めた朗読を実現するため，テキストから感情を自動判定する方式を開発して電子書籍ビューアを試作した．

今後は，話者性や感情の種類と程度を制御することで多様性を実現する，可制御性の向上のアプローチについても取り組んでいきたい．

参考文献

[1] 籠嶋 岳彦，赤嶺 政巳: 閉ループ学習に基づく最適な素片選択の解析的生成，電子情報通信学会論文誌 D-II, Vol.J83-D-II, No.6, pp.1405-1411 (2000).

[2] T. Mizutani and T. Kagoshima: Concatenative Speech Synthesis Based on the Plural Unit Selection and Fusion Method, IEICE Trans. Vol.E88-D, No.11, pp.2565-2572 (2005).

[3] M. Tamura, T. Mizutani and T. Kagoshima: Fast concatenative speech synthesis using pre-fused speech units based on the plural unit selection and fusion method, IEICE Trans. Vol.E90-D, No.2, pp.544-553 (2007).

[4] 橘 健太郎，平林 剛，水谷 伸晃，籠嶋 岳彦: 個人声の合成音作成フレームワークの開発，日本音響学会講演論文集, Mar. 2011.

[5] みんなの声でグリーティングカードを送ろう！, http://tospeak.ivc.toshiba.co.jp/grcd/

[6] 布目光生，鈴木優，森田眞弘: 自然で聞きやすい電子書籍読み上げのための文書構造解析技術，東芝レビュー Vol.66 No.9, pp.32-35 (2011).