発音情報が未知の言語におけるテキスト音声合成システム構築法の検討 *

☆沢田慶, 橋本佳, 大浦圭一郎, 南角吉彦, 徳田恵一 (名工大)

1 はじめに

世界には 7000 におよぶ言語が存在すると考えられており [1]，あらゆる言語のテキスト音声合成 (text-to-speech; TTS) システムを構築できる手法の確立は，音声合成研究の 1 つのゴールであると言える．TTS システムを構築する一般的な手法は，目的とする言語 (ターゲット言語) に関する専門的な知識が必要であり，言語ごとに高い構築コストがかかる．そこで，本稿ではターゲット言語に関する専門的な知識を利用せずに，ターゲット言語の音声とテキストの対のみから TTS システムを構築する手法について検討する．

2 テキスト音声合成システムの構築

一般的な TTS システムはテキスト処理部と波形生成部から構成される．テキスト処理部は，テキストの読みを発音 (音素) 情報が書かれた辞書を用いて推定する．テキスト処理部により推定された読みに基づいて波形を生成する波形生成部では，単位選択型音声合成や，隠れマルコフモデル (hidden Markov model; HMM) に基づく音声合成に代表される統計的パラメトリック音声合成 (statistical parametric speech synthesis; SPSS) [2] 等が用いられる．

通常，SPSS による TTS システムは音響特徴量を音素単位でモデル化する．そのため，新たな言語の TTS システムを構築するためには音素セットの定義，テキスト処理部で用いるテキストを音素列に変換する辞書またはテキスト音素変換器の構築が必要となる．これらの構築には，ターゲット言語に関する専門的な知識が必要となり，言語ごとに高い構築コストがかかる．そのため，音声合成研究において，言語依存性の低い TTS システムの構築手法の検討は重要な課題である．そこで本稿では，発音情報が未知の言語において，音声とテキストの対からなるデータベースから TTS システムを自動構築する手法について検討する．

3 発音情報が未知の言語におけるテキスト音声合成システムの構築

発音情報が未知のターゲット言語では，音素セットの定義や音声データに対応する音素列を得ることが困難である．そこで，ターゲット言語とは別言語の音声認識器 (例えば英語の音声認識器) を用いて音素列を得る．これにより，ターゲット言語の音素セットには，音声認識器の音素セットが用いられることになる．TTS システムを構築する流れを Fig. 1 に示す．TTS システムは，音声認識器 (speech recognizer; SR)，単語アライメント器 (word aligner; WA)，テキスト音素変換器 (grapheme-to-phoneme converter; G2P)，音声合成器 (speech synthesizer; SS) から構成される．各構成要素の説明を以下で述べる．

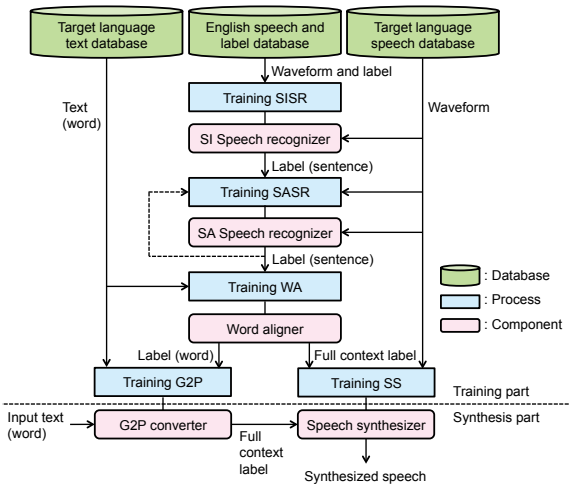

Fig. 1: Overview of TTS system construction.

3.1 音声認識器 (SR)

SPSS のモデルを音素単位で構築するためには，音素セットと音声データに対応する音素列が必要となる．そこで，ターゲット言語とは別言語の不特定話者音声認識器 (speaker independent SR; SISR) を用いて音声認識 (トライフォン音素タイプライタ) を行うことで音素列を得る．より高精度に音素列の推定を行うために，音声データと SISR により得られた音素列を適応データとして，SISR に話者適応手法を適用し話者適応音声認識器 (speaker adapted SR; SASR) を構築する．そして，SASR を用いた音素列の推定と話者適応手法の適用を繰り返すことで音素列の精度を向上させる．

SS では音素継続長をモデル化する必要があるため，音素継続長を考慮した音素列を推定することで，SS に適した音素列が得られると考えられる．しかし，通常の音声認識では音素継続長を考慮した認識は困難である．そこで，明示的に継続長を扱うことができる隠れセミマルコフモデル (hidden semi-Markov model; HSMM) による音素アライメント尤度を用いた音素列の選択を行う．音声認識結果の上位仮説の中で，HSMM を用いた音素アライメント尤度が最も高くなる音素列を最終的な音素列とする．

3.2 単語アライメント器 (WA)

多くの言語は単語単位で分かち書きされているため，単語単位の G2P を構築することが望ましい．また，単語境界情報を用いたフルコンテキストラベルは SS の構築に有用である．そのため，単語単位のテキストに対応する音素列が必要となるが，SR により得られる音素列は文単位であり単語境界が存在しない．そこで，joint multigram モデルを用いた WA を構築する [3]．単語境界を含むテキストと音素列の対を学習データとして WA を学習し，Viterbi デコーディングにより音素列の単語境界を得る．ここで，単語境界の推定精度向上のため，音声認識結果のポーズは単

* Investigation of text-to-speech system construction in unknown-pronunciation language.
by SAWADA, Kei, HASHIMOTO, Kei, OURA, Keiichiro, NANKAKU, Yoshihiko, and TOKUDA, Keiichi
(Nagoya Institute of Technology)

語境界であるという制約を課す．

3.3 テキスト音素変換器 (G2P)

任意のテキストの音声合成を可能とするためには，ターゲット言語のテキストを音素列に変換する必要がある．しかし，発音情報が未知の言語ではテキストを音素に変換する辞書を構築することは困難である．そこで，WA と同様に joint multigram モデルによる単語単位の G2P を構築する [3]．単語単位のテキストと音素列の対を学習データとして G2P を学習し，任意のテキストの音素列を Viterbi デコーディングにより得る．

3.4 音声合成器 (SS)

SR と WA により得られた単語境界付き音素列より，ターゲット言語の音声データに対応するフルコンテキストラベルを作成する．音声データとフルコンテキストラベルの対により，通常の SS と同様の構築手順で SS を構築する．

4 評価実験

提案法の有効性を評価するために，主観評価実験を行った．SISR の言語は英語とし，日本人による評価実験を可能とするためにターゲット言語は日本語とした．SISR の学習データには TIMIT, WSJ0, WSJ1 を用いた．日本語のデータベースとして ATR 音声データベースセット B の MHT 話者を用い，モデルの学習データとして 450 文，評価データとして学習に用いていない 53 文を使用した．日本語のテキスト (書記素) 表記はひらがな，カタカナ，漢字と多岐に渡っており G2P を構築するためには膨大なデータが必要となる．そこで，本実験では日本語テキストはひらがなのみであると仮定した．さらに，分かち書きされた言語を想定して，日本語テキストの文節を分かち書き境界であると仮定した (日本語テキスト例: てれびげえむや␣ぱそこんで␣げえむお␣して␣あそぶ)．SR にはサンプリング周波数 16kHz，フレーム長 25ms，フレームシフト 10ms とした 12 次の MFCC とその 1 次 2 次動的特徴量を音響特徴量として用いた．HMM の構造は 3 状態のスキップなし left-to-right 型とし，SASR における音素列の推定と話者適応の繰り返しは 3 回とした．また，音声認識結果の上位 50 仮説を出力し，HSMM によるアライメント尤度により音素列を選択した．SS にはサンプリング周波数 16kHz，フレームシフト 5ms とした STRAIGHT 分析によって得られた 39 次のメルケプストラム，19 次の非周期成分，対数基本周波数とそれらの 1 次 2 次動的特徴量を音響特徴量として用いた．HMM の構造は 5 状態のスキップなし left-to-right 型の MSD-HSMM とした．また，発話内変動を考慮したパラメータ系列の生成を行った．

はじめに，音声認識時の挿入ペナルティ 0(P0) と −10(P−10)，HSMM による音素列の選択ありなしの組み合わせの計 4 手法の比較実験を行った．実験では合成音声を自然性に関する 5 段階 MOS 試験によって評価した．被験者は 10 人であり，各被験者は評価データの 53 文からランダムに選択された 20 文について評価した．主観評価実験の結果を Fig. 2 に示す．**P0HSMM** が最も高いスコアを得られたことから，HSMM による音素列の選択は有効であると考えれらる．また，**P − 10HSMM** は **P0** と **P − 10** と同等のスコアであることから，適切な挿入ペナルティの設定が重要であると言える．

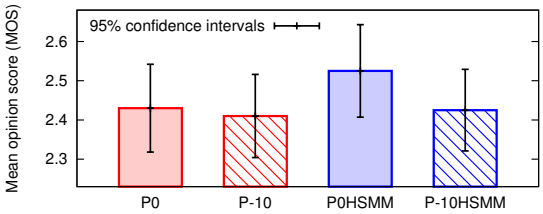

Fig. 2: Results of MOS test comparing four systems.

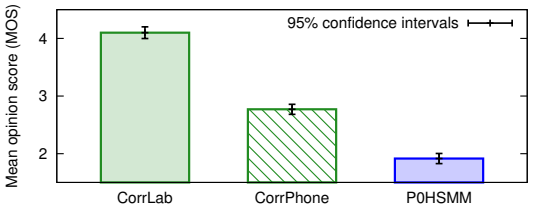

Fig. 3: Results of MOS test comparing three systems.

次に，正解のフルコンテキストラベルから SS を構築した手法 (**CorrLab**)，正解の音素列から WA, G2P, SS を構築した手法 (**CorrPhone**)，**P0HSMM** の 3 手法による比較実験を行った．主観評価実験の結果を Fig. 3 に示す．実験結果より，**CorrLab** と **CorrPhone** のスコアに大きな差があることから，WA と G2P の精度は自然性に大きく影響すると言える．また，**CorrPhone** と **P0HSMM** のスコアにも大きな差がある．これは，SISR に用いた英語と日本語の音素の種類が異なることから，正確にモデル化できない日本語の音素 (促音等) が存在するためだと考えられる．そのため，ターゲット言語に適する音素セットの利用が重要であると言える．

5 むすび

本稿では発音情報が未知の言語の音声とテキストの対のみから TTS システムを自動構築する手法について検討した．本手法により，ターゲット言語の専門的な知識を利用せずに TTS システムを構築できた．主観評価実験により，HSMM による音素列の選択の有効性と，各構成要素の精度が合成音声の自然性に影響があることが示された．今後の課題として，国際音声記号による多言語 SISR の構築，音声データに基づいた音素セットの決定手法の検討 [4, 5] などが挙げられる．

謝辞 本研究の一部は，JST CREST の助成を受けた．

参考文献

[1] Ethnologue: https://www.ethnologue.com

[2] Tokuda *et al.*, “Speech synthesis based on hidden Markov models,” Proceedings of the IEEE, vol. 101, no. 5, pp. 1234–1252, 2013.

[3] Bisani *et al.*, “Joint-sequence models for grapheme-to-phoneme conversion,” Speech Communication, vol. 50, issue 5, pp. 434–451, 2008.

[4] Fukada *et al.*, “Speech recognition based on acoustically derived segment units,” Proc. of ICSLP 96, vol. 2, pp. 1077–1080, 1996.

[5] Muthukumar *et al.*, “Automatic discovery of a phonetic inventory for unwritten languages for statistical speech synthesis,” Proc. of ICASSP 2014, pp. 2594–2598, 2014.