統計的パラメトリック音声合成のための オーディオブックを用いた学習コーパス自動構築*

☆沢田慶, 伊神和輝, 浅井千明, 佐藤雄介, 橋本佳, 大浦圭一郎, 南角吉彦, 徳田恵一 (名工大)

1 はじめに

統計的パラメトリック音声合成 (statistical parametric speech synthesis; SPSS) システムは，大量の学習データを利用して統計モデル (音響モデル) を学習することにより合成音声の音質が向上する．音響モデル学習のために，利用できるビックデータとしてオーディオブックがある．オーディオブックは，インターネット上に大量に公開されており，音声データとテキストの対が存在することから，音響モデルの学習データに適している．しかし，音声データの発話内容とテキストには不一致が存在し，学習データとして利用するためには，この問題に対応する必要がある．本研究では，オーディオブックを学習データとして SPSS システムを構築するために，音声認識を用いた学習コーパスの自動構築法について検討する．

2 オーディオブックを用いた統計的パラメトリック音声合成システムの構築

2.1 統計的パラメトリック音声合成システム

近年，計算機の処理能力向上によりビックデータが利用可能となり，様々な研究分野においてビックデータを用いた統計モデルの学習が成功を収めている．音声合成においては，隠れマルコフモデル (hidden Markov model; HMM) やディープニューラルネットワーク等の音響モデルを用いた SPSS が成果を挙げている．SPSS は，十分に整備された大量の学習データを用いて音響モデルを学習することにより合成音声の音質が向上することが知られている．そのため，SPSS の研究においてビックデータを用いた音響モデル学習は重要な課題である．しかし，SPSS の学習データには，クリーン (ノイズが少ない) かつ同一環境で収録された音声が適しており，音響モデル学習のために大量の音声を収録することは高いコストを必要とする．そのため，大量の音声データとテキストの対が比較的容易に利用できるオーディオブックを学習データとする SPSS システムの構築が注目を集めている [1].

2.2 オーディオブック

オーディオブックは，書籍 (テキスト) を朗読した音声を収録したコンテンツであり，既に大量のコンテンツがインターネット上に存在する．同一の話者による比較的クリーンかつ同一環境で収録された音声が多く存在することから，音響モデルの学習データに適している．さらに，音響モデルの学習には音声データとテキストの対が必要となるが，オーディオブックには朗読したテキストが存在するため，容易に音声データとテキストを利用することができる．このようなことから，オーディオブックは音響モデル学習のためのビックデータとして非常に有用である．

オーディオブックは，大量の音声データとテキストを学習データとすることを可能とするが，音響モデルの学習データとして問題もある．オーディオブックには，テキストの読み間違いや，テキストには存在しない文章 (ブックの説明や擬音語等) を追加情報として収録する等，音声データの発話内容とテキストに不一致が存在する．この不一致は，音響モデルの学習に悪影響を与える．そこで本研究では，オーディオブックから音響モデルの学習に適する学習コーパスを自動構築する手法について検討する．

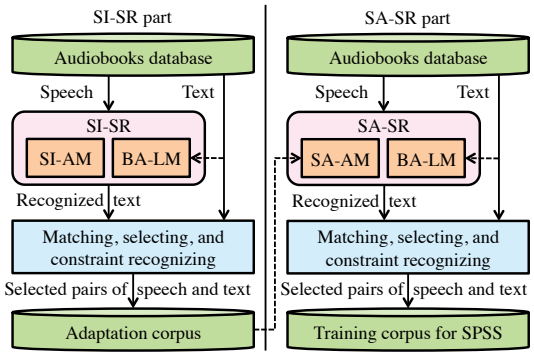

Fig. 1: Overview of training corpus construction (SI: speaker independent, SA: speaker adapted, BA: book adapted, SR: speech recognizer, AM: acoustic model, LM: language model)

3 学習コーパスの自動構築

オーディオブックにおいて音声データの発話内容とテキスト (ブックテキスト) には不一致が存在する．学習コーパスのテキストには，発話内容と一致するテキスト (正解テキスト) を用いることが望ましい．しかし，大量の音声データの正解テキストを人手により得ることは高コストである．そこで，音声認識により音声データのテキスト (認識テキスト) を推定する．そして，ブックテキストと認識テキストの単語一致度の高い音声データとテキストの対を学習コーパスとする [2, 3]．本研究では，学習コーパスとして採用する基準である単語一致度の閾値と，学習コーパスのテキストの違いが合成音声へ与える影響を調査する．さらに，ブックテキストと認識テキストから正解テキストに近いテキストを推定する手法を提案する．不特定話者音声認識部と話者適応音声認識部から構成される，オーディオブックを用いた学習コーパス自動構築の流れを Fig. 1 に示す．

3.1 オーディオブックの音声データとテキストを用いた音声認識

本手法では，オーディオブックの音声データとテキストを用いて音声認識精度の向上を図る．

音声データとブックテキストの大部分は一致していると考えられる．そのため，ブックテキストを用いた言語モデルは音声認識に非常に有用である．そこで，大量のテキストデータから構築した汎用言語モデルに，ブックごとに適応を行ったブック適応言語モデルを音声認識に利用する．

不特定話者音声認識部では，不特定話者音響モデルとブック適応言語モデルを用いて音声認識を行う．そして，ブックテキストと認識テキストの単語一致度

*Automatic construction of training corpus using audiobooks for statistical parametric speech synthesis. by SAWADA, Kei, IGAMI, Kazuki, ASAI, Chiaki, SATO, Yusuke, HASHIMOTO, Kei, OURA, Keiichiro, NANKAKU, Yoshihiko, and TOKUDA, Keiichi (Nagoya Institute of Technology)

Correct text: baby dinosaurs broke out of the eggs crack
Book text: baby dinosaurs broke out of the eggs
Recognized text: they dinosaurs broke out of the eggs crack

Fig. 2: Example of correct, book, and recognized texts

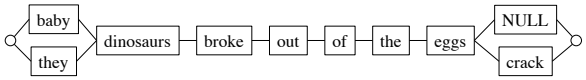

Fig. 3: Example of a word network for decoding

が高い音声データとテキストの対を適応コーパスとして話者適応音響モデルを構築する．話者適応音声認識部では，話者適応音響モデルとブック適応言語モデルを用いて音声認識を行い，単語一致度が高い音声データとテキストの対を音声合成用音響モデルの学習コーパスとする．

3.2 学習データの選択

音響モデルの学習は，大量の学習データを用いるほど音質が向上する．学習データとして採用する基準である単語一致度の閾値を低く設定すると学習データ量は増加する．しかし，単語一致度が低い場合には，ブックテキストと認識テキストは正解テキストとの違いが大きい可能性がある．一方，閾値を高く設定すると，学習データ量は減少するが，正解テキストとの違いが小さいと考えられる．そのため，単語一致度の閾値には適切な設定をする必要がある．

3.3 学習コーパスのテキスト

オーディオブックの音声データには，ブックテキストには存在しない追加情報が収録されることがある．正解 · ブック · 認識テキストの例を Fig. 2 に示す．Fig. 2 の例では，ブックテキストには存在しない擬音語 (crack) が収録されている．そのため，ブックテキストを学習コーパスとして用いると，追加情報に対応するテキストは存在しないため，音響モデルの学習に悪影響を与える．一方，認識テキストを学習データとして用いると，追加情報に対応するテキストには音声認識結果が用いられるが，認識テキストには認識エラー (Fig. 2 中の they) が含まれる可能性がある．

本研究では，正解テキストに近いテキストを得るために，学習データを選択した後に学習コーパスとして用いるテキストを推定する．Fig. 3 のように，ブックテキストと認識テキストで違いがある単語を選択できる単語ネットワークを構築し，この単語ネットワークに制限した音声認識を行う．単語ネットワークには，ブックテキストが NULL の場合には認識テキストが選ばれやすくなるペナルティ，それ以外にはブックテキストが選ばれやすくなるペナルティを課す．また，言語モデルは利用せずに，単語ネットワークのペナルティを考慮した音響モデルのスコアによりテキスト (制限認識テキスト) を推定する．この制限認識テキストは，認識エラーを削減しつつ，追加情報に対応するテキストを得ることが期待される．

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．オーディオブックには，同一話者により収録された英語の児童書 22 冊を用いた [4]．音響モデルの学習データとして 19 冊，評価データとして学習に用いていない 3 冊を使用した．このオーディオブックは児童書を朗読した音声であるため，表現豊かな音声やブックテキストには存在しない擬音語等が多数収録されている．音声ファイルは，ブックのページごとに分かれており，音声ファイルの長さは平均 20 秒である．本実験では，音声ファイルを文ごとに分割することはせず，ページ単位の音声ファイルを用いて学習を行うこととした．

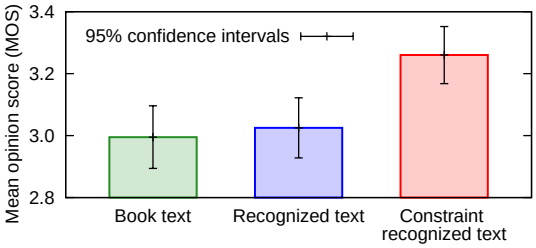

Fig. 4: Results of MOS test

音声認識は，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, 2 次動的特徴量を音響特徴量とした．不特定話者音響モデルと汎用言語モデルの学習には TIMIT と WSJ を用いた．音響モデルは GMM-HMM，言語モデルは単語 bi-gram とした．適応コーパスとする単語一致度の閾値は 80%，学習コーパスとする閾値は 60, 70, 80%として実験を行った．音声合成は，サンプリング周波数 44.1kHz，フレームシフト 5ms とした STRAIGHT 分析によって得られた 49 次のメルケプストラム，24 次の非周期成分，対数基本周波数とそれらの 1, 2 次動的特徴量を音響特徴量とした．音響モデルは MSD-HSMM とし，発話内変動を考慮したパラメータ系列の生成を行った．

学習コーパスのテキストとして，ブック · 認識 · 制限認識テキストを用いた 3 つのシステムの比較実験を行った．各システムにおいて，学習コーパスとする閾値には 60, 70, 80%の中で最も低いメルケプストラム歪が得られた閾値 (ブック: 60%, 認識: 70%, 制限認識: 60%) を用いた．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からランダムに再生される 20 ページについて評価した．主観評価実験の結果を Fig. 4 に示す．実験結果より，制限認識テキストを学習コーパスのテキストとして用いたシステムが最も高い主観評価値を得た．

5 むすび

本研究では，オーディオブックを学習データとして SPSS システムを構築するために，音声認識を用いた学習コーパスの自動構築法について検討した．実験結果より，単語ネットワークに制限を設けた音声認識により得られた制限認識テキストは，学習コーパスのテキストとして有用であった．今後の課題として，より大量の学習データを用いた音響モデルの学習や母語話者による主観評価実験が挙げられる．

謝辞 本研究の一部は，JST CREST の支援を受けた．

参考文献

[1] King *et al.*, “The Blizzard Challenge 2013,” Blizzard Challenge 2013, 2013.
[2] Braunschweiler *et al.*, “Lightly supervised recognition for automatic alignment of large coherent speech recordings,” Interspeech 2010, pp. 2222–2225, 2010.
[3] Takaki *et al.*, “Overview of NITECH HMM-based speech synthesis system for Blizzard Challenge 2013,” Blizzard Challenge 2013, 2013.
[4] Usborne Publishing Ltd release of audiobook recordings for Blizzard 2015, http://www.cstr.ed.ac.uk/projects/blizzard/2015/usborne_blizzard2015