音素選択型音声合成方式に用いるデータベースサイズと主観評価の関係分析

Analysis of the relation between the size of speech database which is used for phone-sized unit selection speech synthesis and subjective evaluations

稲井 禎
Inai Tadashi

岡山大学 大学院自然科学研究科
Graduated School of Natural Science, Okayama University

1. はじめに

現在，主流となっている音声合成方式として波形合成方式 [1] と HMM 音声合成方式 [2] が挙げられる．波形合成方式は，一般的に多量の音声データを必要とする．HMM 音声合成方式は，波形合成方式ほどの音声データは必要ないが，合成音声がやや機械的になる．また，上記の 2 つの方式のハイブリッドとして HMM に基づく音素選択型音声合成方式 [3] が提案されており，同程度のデータベースサイズの波形合成方式や HMM 音声合成方式よりも高品質な音声を合成できる可能性がある．

本研究では，音素選択型音声合成方式における少量のデータベースによる合成音声の高品質化を目標としている．そこで本稿では，その第一段階としてデータベースサイズと合成音声の品質との関係を知るために複数種類のデータベースサイズによる音声合成を行い，実施した主観評価実験をもとにデータベースサイズと主観評価との関係を分析する．

2. 音素選択型音声合成方式

音素選択型音声合成方式の概要を図 1 に示す．この方式は図 1 に示すように，学習部と合成部からなる．

学習部では，音声分析により特徴量を抽出したのち，それらとラベルファイルを学習データとして HMM 学習を行う．その後，音素継続時間長が再設定されたラベルファイルをもとに，抽出した特徴量および音声波形を音素ごとに分割し，トライフォン単位でグループ分けをすることにより音素データベースを構築する．

合成部では，HMM により音声合成に必要なパラメータを生成する．その後，音素データベースに対して音素ごとに前後の音韻環境を考慮した事前選択を行い，適合コスト計算・接続コスト計算を経て HMM によって生成されたパラメータに類似した音素波形系列が選択される．最後に，選択された音素波形系列を滑らかに接続することで音声を出力する．

3. 評価実験

音素選択型音声合成方式に用いるデータベースサイズと合成音声の品質の関係を見るために主観評価実験を行った．使用したデータベースは女性話者 1 名の音声 20 分から 1200 分の 6 種類である．主観評価実験では，被験者 10 名による「1:非常に悪い」から「5:非常に良い」の 5 段階評価を行った．データベースサイズごとの評価結果を図 2 に示す．図 2 のエラーバーは MOS の分散を表す．また，音素選択の際に重視する音声パラメータに応じて合成音声の品質が変化すると考え，表 1 に示す計算順序を用意し，音声合成を行った．♣ はメルケプストラム，対数 F_0 の順に計算することを表し，♢ は ♣ の場合とは逆の計算順序になることを表す．図 2 より，計算順序 3,4 の MOS が計算順序 1,2 よりも高いことから，合成音声の高品質化には F_0 の接続性が重要であることがわかる．

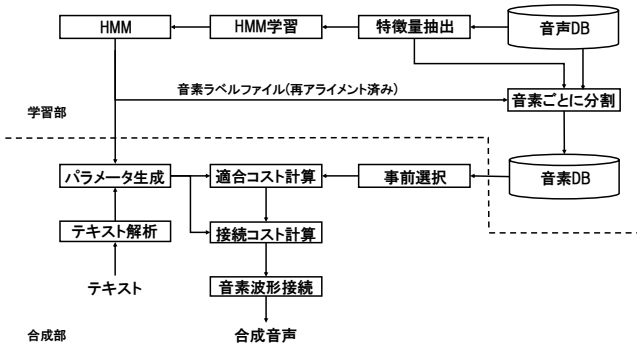

図 1　提案方式の全体像

表 1　各コスト計算に用いる特徴量の使用順序

計算順序	1	2	3	4
適合コスト	♣	♢	♣	♢
接続コスト	♣	♣	♢	♢

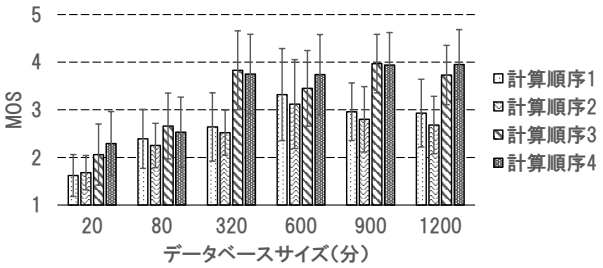

図 2　主観評価結果

4. まとめと今後の課題

本稿では，音素選択型音声合成方式に用いるデータベースと主観評価の関係を分析した．

今後は，本稿によって得られた知見をもとに音素選択型音声合成方式における最適な音素選択方式を検討する予定である．

参考文献

[1] ニック・キャンベルら，信学技報 SP96-7, pp. 45–52, 1996.
[2] 徳田 恵一，信学技報 SP 2000-74，pp. 43–50，2000.
[3] Z. Ling *et al.*，ICASSP，Vol.4，pp. 1245–1248，2007.