1—7—25

波形接続型 Speech-to-Speech 音声合成の自然対話発話コーパスへの適用*

◎正木敦之 (奈良先端大 情報) △柏岡秀紀 (奈良先端大/ATR)
ニック キャンベル (奈良先端大/ATR/CREST)

1 はじめに

音声合成技術はカーナビやロボットへの搭載, 障害者のコミュニケーション補助の技術として期待されている. 従来の波形接続型合成器は音素ラベル済みの小規模音声コーパスを利用し, 朗読音声の合成を実現したが, 日常会話など表現豊かな音声の合成には至っていない.

このような音声の合成には日常会話そのものを収録した大規模な音声コーパスが必要であると考えられる. 我々はこの大規模なコーパスを簡単に扱うため, Speech-to-Speech と呼ばれる技術を提案している [1]. Speech-to-Speech は入力音声に最も近い音声を, ラベル無し大規模音声コーパスから音響情報をたよりに再構築する技術である. 入力音声として従来の合成器 (たとえば CHATR) の出力を使い, Speech-to-Speech 出力として, 様々な発話スタイルに表れるような韻律のバリエーションを含んだ音声を取得し, 後処理として所望の発話スタイルを絞ることにより, 最終的に我々の望む音声を得ることができると考えられる.

文献 [1] では音声コーパスとして ATR 音素バランス文 (503 文) を使用し Speech-to-Speech 技術を検証した. 本稿では, Speech-to-Speech 合成技術の詳細をおさらいした後, 大規模自然対話発話コーパスを使用した Speech-to-Speech 合成音の品質評価実験について述べる.

2 Speech-to-Speech 合成の詳細

Speech-to-Speech は以下の 3 つのセクションから構成される. 音韻ラベルを使わずに処理するため, 音声コーパスを容易に拡張することが出来る.

2.1 音声単位の切り出し

接続歪みの軽減を狙いエネルギの局所的最小点を音声単位境界とした. 実際には以下のようなチューニングを行い音声単位の長さを制限している. これは 503 文をコーパスとしていたことに由来する.

まず 0 ～ 8000 Hz までの周波数帯域をメルスケールで 16 バンドに等分割し, 各バンドのエネルギを求める. 次に, いわゆる Power と各バンドのエネルギ双方の局所的最小点を Mermelstein らの手法 [2] で求め, 音声単位の境界候補とする. 各境界候補のうち, Power 由来のものと各バンド由来のものとを比べ, 20 msec 以内の差であれば同じ境界であるとし, Power 由来のものにマージする. さらに, 各バンド由来の境界候補のうちマージされずに残った候補同士を比べ, それぞれ 40 msec 以内の候補を時間的に先の候補にマージする. ここで残った候補と Power 由来の候補を最終的な音声単位境界とする.

2.2 音声単位の特徴付け

スペクトル情報, 韻律情報の両面から特徴付けを行う. スペクトル情報として上記切り出し過程で用いた 16 バンドのエネルギを参照する. 各エネルギの時間的な変化は離散コサイン変換により符合化されそれぞれ 5 つのパラメータに落される (計 80 個のパラメータとなる). 韻律情報には F_0 と Dur を用いる. F_0 の時間的な変化は同様の手法で 5 つのパラメータとなり, Dur を併せて計 86 次元のベクトルとしてひとつの音声単位が特徴付けられる.

2.3 単位選択と波形接続

入力音声とコーパス中の音声は前述の方法で音声単位に切り分けられ, それぞれの音声単位は 86 次元のベクトルで特徴付けられる. 単位選択は, 入力音声から切り出された音声単位の特徴ベクトルとコーパス中の特徴ベクトルの重み付きユークリッド距離の計算により実現される. 重みはベクトルの各次元について標準偏差の逆数とする. この距離の一番近い単位を順に接続し, Speech-to-Speech の出力音声とする. 将来的には, 距離の近い順に上位 N 個を候補とし, この候補を後処理で所望の発話スタイルに絞り込むことになる.

3 評価実験

前節で述べた合成技術の評価のため, 3 つの実験を行った. 実験に使用したデータ, 実験内容, 実験結果の順で以下に説明する. 被験者は日本人男子学生 4 名である.

3.1 使用データ

3.1.1 音声コーパス

CREST/ESP プロジェクト [3] により採集された女性話者 FAN の電話での自然対話発話 [4] 約 44 時間のうち, 無音区間を除いた約 7 時間のデータを使用し

*"Application of concatenative Speech-to-Speech synthesis for spontaneous speech corpus" by A. MASAKI (Nara Institute of Science and Technology (NAIST)), H. Kashioka (NAIST/ATR) and N. Campbell (NAIST/ATR/CREST)

表 1: 実験 (1) 書き取り精度実験結果

ターゲット	FAN	JFA	JMA	CHATR
精度平均値 (%)	76.74	50.93	42.56	63.26

表 2: 実験 (2) 了解度実験結果

ターゲット	FAN	CHATR
了解度平均値	3.950	3.025

た. 音声単位の切り出し (2.2 節) の結果, 236923 個の音声単位が切り出され, 音声単位の平均 Dur は 0.102 sec であった.

3.1.2 ターゲット

FAN そのもの (ターゲットはコーパスから隠しておく), CREST/ESP プロジェクト で作成中の電話対話データベース [5] 中より女性話者 JFA と男性話者 JMA , そして話者 FAN の ATR 音素バランス文 (503 文) を用いた CHATR 合成音の 4 つを使用した.

3.2 実験

以下の 3 つの実験により, スペクトルに関する評価 (書き取り, 了解度) と韻律に関する評価 (F_0 高低の類似性) を行った.

3.2.1 実験 (1) 書き取り

被験者に合成音を試聴してもらい発話内容を書き取ってもらった. 問題は FAN, JFA, JMA, CHATR をターゲットとした合成音各 20 発話分を用意した. 正解の揺れを抑えるため, 各被験者互い違いに 20 問中 5 問ずつターゲットそのものを試聴してもらい, 正解セットを作成した. 評価は正解セットと被験者の回答とを比較し音素単位の正解精度を計算した.

3.2.2 実験 (2) 了解度

被験者に発話の書き起こしテキストを見てもらい, 合成音がそのテキストを読んだものとして 5(完全に了解できる), 4(おおむね了解できる), 3(かろうじて了解できる), 2(ほとんど了解できない), 1(全く了解できない) の 5 段階で評価してもらった. 問題は実験 (1) で精度の良かった FAN と CHATR をターゲットとし合成音各 20 発話分を用意した.

3.2.3 実験 (3) F_0 高低の類似性

F_0 の高低を再現しているかを評価するため, 合成音 1 つに対し A,B の 2 種類のターゲット候補を試聴してもらい, 合成音がどちらのターゲットから合成されたかを判断してもらった. 回答は「ターゲット A」「ターゲット B」「わからない」の 3 種類の中から選択してもらった. また, ターゲット音声は FAN の発話から, 音韻の影響を抑えるため発話「うん」のみに絞った. 問題は合計 20 問用意した.

3.3 結果

各実験の結果を表 1, 表 2, 図 1 に示す.

実験 (1) では, コーパスとターゲットの話者が同じ FAN をターゲットとした場合が一番精度が良く,

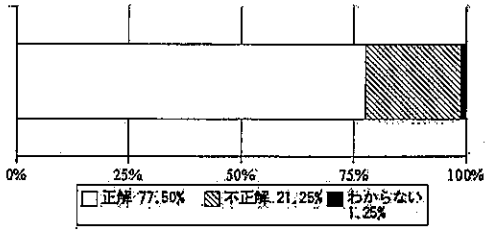

図 1: 実験 (3)F_0 高低の類似性実験結果

FAN の小規模コーパス利用した CHATR 合成音, 同性の JFA, 異性の JMA の場合と続いた.

実験 (2) ではターゲットが FAN の場合 3.950 ポイントで「おおむね了解できる」, CHATR の場合 3.025 ポイントで「かろうじて了解できる」ことが確認できた. 実験 (1) と併せると, FAN あるいは FAN の小規模コーパスを利用した CHATR 合成音がターゲットの場合に, ある程度良好にスペクトルが再現されることがわかった.

実験 (3) では正解率が 77.50 % で不正解を大きく上回り, F_0 の再現性が確認できた.

4 まとめ

表現豊かな音声の合成を目標として, ラベル無し大規模コーパスを簡単に利用できる Speech-to-Speech 音声合成技術を提案し, 大規模自然発話コーパスへの適用を行った.

コーパスとして女性話者 FAN の自然発話音声を, ターゲットとして FAN そのもの, 同性異話者 JFA, 異性話者 JMA, FAN の小規模コーパスを利用した CHATR 合成音の 4 種を用い Speech-to-Speech 合成音を作成, 品質評価実験を行った結果, コーパスと同話者, あるいは同話者の小規模コーパスで作成した CHATR 合成音をターゲットとした場合に, ある程度良好にスペクトルが再現できること, F_0 が 8 割近い確率で再現できることが確認できた.

今後, 将来的に狙う CHATR をターゲットとした場合の品質を向上させるとともに, 発話スタイルによる出力音声の絞り込みを行う必要がある.

謝辞 本研究の一部は科学技術振興機構戦略的基礎研究推進事業 (JST/CREST) の援助により行われた.

参考文献

[1] 正木, 柏岡, キャンベル, "波形接続型 Speech-to-Speech 音声合成のための可変長音声単位による単位選択手法", 信学技報, SP2003-82, pp.49-54, August, 2003.

[2] P. Mermelstein, "Automatic segmentation of speech into syllabic units", J. Acoust. Soc. Am. 58(4), pp.880-883, 1975.

[3] http://feast.his.atr.co.jp/

[4] 石井, キャンベル, 音講論, Vol.1, pp.277-278, 2002-10.

[5] 芦村, キャンベル, "JST/CREST 発話様式プロジェクトの電話対話データベース", 人工知能学会全国大会論文集 3C5-11(2002).