HMM を用いた音声からの唇画像合成法*

5H－8

山本英里、中村 哲、鹿野清宏†

奈良先端科学技術大学院大学 情報科学研究科‡

1 はじめに

自然な動きの唇画像があれば、聴覚障害者は画像からリップリーディングによって、発話内容を得ることが出来る。また、コンピュータエージェントの口部分を、より人間らしく表現することに役立つ。唇画像を合成する為には、音声から画像へ変換する方法と、テキストから画像へ変換する方法があり、ここでは音声から画像へ変換するシステムについて検討する。従来の方法としては、1フレーム毎に、音声パラメータから画像パラメータへと変換するシステムがある [1][2]。この方法では、フレーム毎に逐次変換する為、1発話全体で見るとひずみが大きくなる問題がある。本研究では、HMM(隠れマルコフモデル) を用いて、1発話全体で最適な変換を行うシステムについて述べ、実験により有効性を示す。

2 HMM を用いた合成アルゴリズム

まず学習方法について説明する。音声と画像を同期させたデータを用意する。音声の HMM を用い、1発話毎に全フレームにわたって、遷移確率と出力確率分布から尤度を計算し、最も尤度の高い遷移パスを選択する (アライメントをとる)。これにより、図1の様にフレーム毎に対応する音素と対応する状態が決まる。

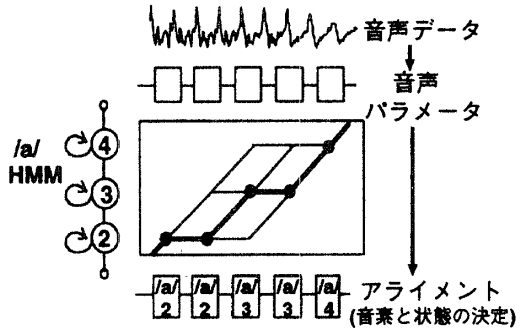

図 1 フレーム毎に音素と状態を決定

学習で使用される全てのフレームのうち、同じ音素、同じ状態をとるフレームを選び、その画像パラメータの平均値をとる。これにより、図2の様に各音素の各状態毎に、画像パラメータの平均値が求まる。

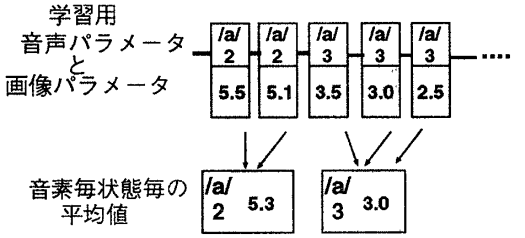

図 2 画像パラメータの学習

次に合成では、テスト用の音声データを分析し、音声の HMM により1発話毎にアライメントをとる。この時、音素毎状態毎に、学習した画像パラメータを対応させる。この画像パラメータを元に、1フレーム毎に3次元の唇画像を合成し、1発話毎に接続したものを画像系列として出力する。

3 実験条件

3次元位置測定装置を用いて、音声と画像のデータを収録する。音声と画像は、125Hz で同期がとられている。分析パラメータは、音声はメルケプストラム係数16次、差分係数16次、差分エネルギーが1次である。画像パラメータは、口の縦の開き x、口の横の開き y、口端の奥行き z、の3変数を使用する [3]。学習とテストに、1名の特定話者 (女性) が発話した316単語を使用する。学習には音韻バランスのとれた216単語を用い、テストにはそれ以外の100単語を使用する。一方、アライメントをとるのに使用した音声 HMM は次の条件で作成している。モデル数は、54音素に発話前 pause と発話後 pause を加えた56個。状態数は3状態。出力確率分布は Gaussian Tied- Mixture 型である。音声の周波数は 12kHz のサンプルに、窓長 32msec のハミング窓をかけ、分析周波数を 125Hz とする。このモデルでの音素認識率は、認識率を (正答音素数−湧出し音素数)/全音素数で定義する場合、学習内単語で 91.8%、テスト単語で 69.4%の結果となる。

*Speech-to-Lip Movement Synthesis by HMM

†Eli Yamamoto, Satoshi Nakamura and Kiyohiro Shikano

‡Graduate School of Information Science, Nara Institute of Science and Technology(NAIST)

4 2乗誤差による評価

画像データを学習の結果は、合成画像パラメータ x_s, y_s, z_s と収録画像パラメータ x_o, y_o, z_o 間の2乗誤差 E で評価する。

$$E = \sqrt{(x_s - x_o)^2 + (y_s - y_o)^2 + (z_s - z_o)^2} \ \text{(cm)}$$

表1に学習単語とテスト単語での1フレームあたりの平均2乗誤差を示す。

表1 2乗誤差

	216単語	100単語
正答時 E	1.448	1.503
音素認識 E	1.473	1.530

正答時と音素認識の場合の誤差には大きな差がみられない。具体的な合成画像パラメータを図3に示す。横軸はフレーム番号、縦軸が E である。合成画像は実線、収録画像は破線(全体が滑らか)で描かれている。縦線は、音声区間の区切りを示す。

5 後続音素依存の合成方法

HMMによる合成方法で、誤差が大きく寄与する音素を調べたところ、/h/や促音/Q/, また発話前pauseが顕著であることが分かった。/h/の例を図4に示す。

/h/と/Q/の特徴は、後続音素に依存した口形をとることである。そこで、HMMを用いる過程で、後続音素に依存した合成方法に拡張する。

まず、画像パラメータの学習において、アライメントをとる時に、後続音素に注目する。そして、音素毎状態毎の画像パラメータを平均化する際に、後続音素別に平均値をとる。ただしこれでは、合成画像パラメータのパターン数が多くなりすぎ、学習できないパターンが出てくる。そこで、後続音素の第一状態の画像パラメータを、クラスタリングする。この結果出来るクラスを、ここではvisemeと呼ぶ。画像パラメータの平均値は、音素毎状態毎、更に後続音素のviseme毎に用意することになる。

合成時にも、アライメントをとる時には、後続音素のvisemeに注目する。入力フレーム毎に、音素、状態、後続音素のvisemeを見て、平均値画像パラメータを出力する。

後続音素依存の結果を表2に示す。全体として後続音素依存の方法により、誤差値が改善されることが示されている。

表2 2乗誤差(後続音素依存)

	216単語	100単語
正答時 E	1.113	1.203
音素認識 E	1.155	1.277

また画像パラメータの変化を図5に示す。図では、/h/の部分の大きな誤差が、目立って小さくなっている事が分かる。

6 まとめ

音声を入力し、自然な動きの唇画像を合成する事を目的に、音声パラメータから画像パラメータへ、HMMを用いて変換する方法について述べた。さらに、後続音素を参照する改善方法により改善が可能であることを示した。

今後、視覚的な差異を正しく反映する客観的評価尺度を求めていく必要がある。また、主観的なリップリーディングによる評価も行う予定である。

謝辞

データの使用を許可して頂いたATR人間情報通信研究所の東倉社長、Bateson博士に感謝致します。

参考文献

[1] Lavagetto, F. IEEE Trans. on Rehabilitation Engineering, Vol.3, No 1,1995.

[2] Morishima, S. & Harashima, H. IEEE J. on Selected Area in Communications, Vol.9, No.4.,1991.

[3] Guiard-Marigny, T., Adjoudani, A. & Benoit, C. Proc. of the Second ESCA/IEEE Workshop on Speech Synthesis, 1994.

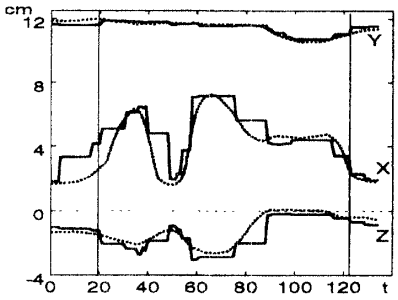

図3 /depaato/

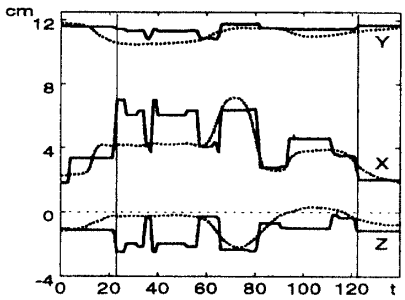

図4 /hohoemu/

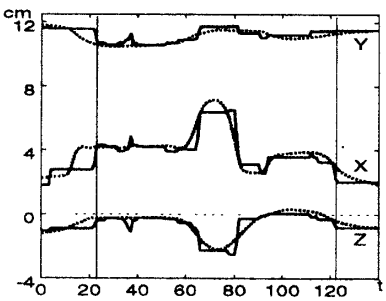

図5 /hohoemu/
(後続音素依存)