自然対話音声コーパスを用いたDNN音声合成におけるパラ言語情報制御*

☆横山雅季, 森大毅, 永田 智洋 (宇都宮大)

1 はじめに

話者の感情や態度などをパラ言語情報として表現できる音声合成によって，より人間に近いコミュニケーションを実現することができると考えられる。我々は，重回帰 (MR)HSMM に基づく対話音声合成の検討を進めてきた [1]。これにより，快-不快，覚醒-睡眠などの次元によるパラ言語情報の制御が可能となったが，自然性が課題であった。

近年， Deep neural networks (DNN) を用いた音声合成手法の研究 [2] が盛んに行われている。従来の HMM 音声合成よりも高品質な音声を合成できるようになってきたため， 自然対話音声コーパスを用いたパラ言語情報の制御に関しても，自然性の改善とともにパラ言語情報の制御性の向上も期待できる。

本研究では，自然対話音声コーパスを用いた DNN 音声合成におけるパラ言語情報の制御を検討する。

2 パラ言語情報制御手法

DNN 音声合成は，学習データの言語的特徴と音響的特徴の関係のモデル化を DNN で行うものである。今回用いた自然対話音声コーパスは，宇都宮大学パラ言語情報研究向け音声対話データベース (UUDB)[3] である。UUDB に含まれる発話には，ラベラ 3 名による快-不快， 覚醒-睡眠など 6 次元の評価値が 7 段階 (4:中立) で付与されており，今回は 3 名の平均評価値をパラ言語情報ラベルとして用いた。言語的特徴に加えて，各発話の快-不快，覚醒-睡眠の 2 次元の平均評価値を連続量として入力に用いた。

3 パラ言語情報制御実験

学習データには，UUDB の女性話者 1 名 (627 発話) のうち，セッション C002 から C007 に対応する 538 発話を用い，セッション C001 の 89 発話で音声を合成した。サンプリング周波数は 16 kHz である。

音声合成ツールキット Merlin [4] を用いて，学習，合成を行った。DNN の継続長モデルの入力には，言語的特徴ベクトルに快-不快，覚醒-睡眠の 2 次元を加えた 777 次元を，音響モデルの入力には，フレーム内の音素の位置を与える 4 次元をさらに加えた計 781 次元を用いている。音響的特徴は，対数 F0，非周期性指標，メルケプストラム係数 35 次元と，それぞれの動的特徴量を合わせたものに，有声無声判定 1 次元を加えた 112 次元である。

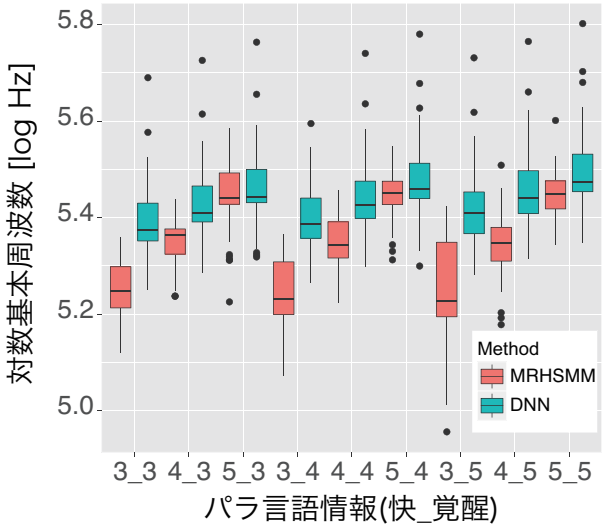

Fig. 1: パラ言語情報の変化に伴う対数 F0 の平均値の変化

評価対象音声の合成において入力するパラ言語情報は，快-不快 3 レベル (3,4,5)，覚醒-睡眠 3 レベル (3,4,5) の全ての組合せとした。

3.1 合成音声の分析

DNN 音声合成によるパラ言語情報の可制御性を，MRHSMM に基づく音声合成 [1] との比較により検討した。MRHSMM の回帰行列は，最大事後確率推定によって得ている。

入力するパラ言語情報を変化させた時の対数 F0 平均値の変化を図 1 に示す。DNN，MRHSMM のいずれも，快-不快 次元の入力を快寄りにすると，対数 F0 の平均値が上昇していることがわかる。

スペクトル傾斜の変化を図 2 に示す。MRHSMM では，覚醒寄りになるにつれてスペクトル傾斜が大きくなる傾向があるが，DNN ではほとんど変化が見られなかった。

3.2 主観評価実験

DNN 音声合成と MRHSMM で合成した音声を自然性とパラ言語情報の知覚実験によって比較検討する。主観評価実験で呈示する音声は，UUDB の話者 FTS のセッション C001 の中から，パラ言語情報を反映しやすいと思われる 10 発話を選んだ。被験者 8 名に DNN と MRHSMM の 2 種類の音声合成手法で，

*Paralinguistic information control in DNN Speech synthesis using natural dialogue speech corpus. by YOKOYAMA, Masaki, MORI, Hiroki, NAGATA, Tomohiro (Utsunomiya University)

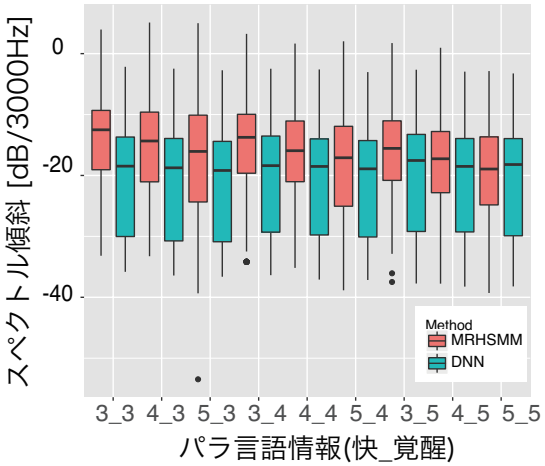

Fig. 2: パラ言語情報の変化に伴うスペクトル傾斜の変化

快-不快 3 レベル (3,4,5)，覚醒-睡眠 3 レベル (3,4,5) の 9 種類のパラ言語情報を操作することによって作成した音声計 180 発話をヘッドホンにより聞かせた。評価者は，音声を 1 度のみ聴取し，合成音声の自然性を 1~5 の 5 段階で評価させた。その後もう一度だけ音声を聴取させ，知覚されたパラ言語情報の評価値を快-不快，覚醒-睡眠に関してそれぞれ 1~7 の 7 段階で評価させた。

3.3 実験結果及び考察

入力したパラ言語情報ごとの自然性の MOS を図 3 に示した。与えたパラ言語情報が快-不快，覚醒-睡眠ともに中立のときの MRHSMM と DNN の MOS の中央値は，3.13, 3.69 であった。特に，パラ言語情報を不快寄りに操作すると MRHSMM の MOS 評価値が DNN に比べて大きく低下していることが読み取れる。DNN での MOS は，パラ言語情報を操作してもほとんど変わっていない。このことから，DNN 音声合成でのパラ言語情報制御は可能であると結論できる。

パラ言語情報の知覚実験の結果を図 4 に示した。入力したパラ言語情報の値と知覚された値の相関を調べた結果，DNN の方が相関係数が大きいことが分かった。MRHSMM の相関係数が過去の研究 [1] に比べて極端に低いが，これは図 3 が示すように，今回の検討では与えるパラ言語情報によっては，音声の自然性が極端に悪化しており，パラ言語情報の評価が困難だったためだと考えられる。DNN 音声合成の方は，客観評価ではパラ言語情報を操作した時の音響特徴量の変化は，MRHSMM に比べて小さかったが，主観評価では違いが明確に知覚できている。

4 おわりに

本稿では，対話音声合成におけるパラ言語情報制御手法として，MRHSMM による方法と DNN による

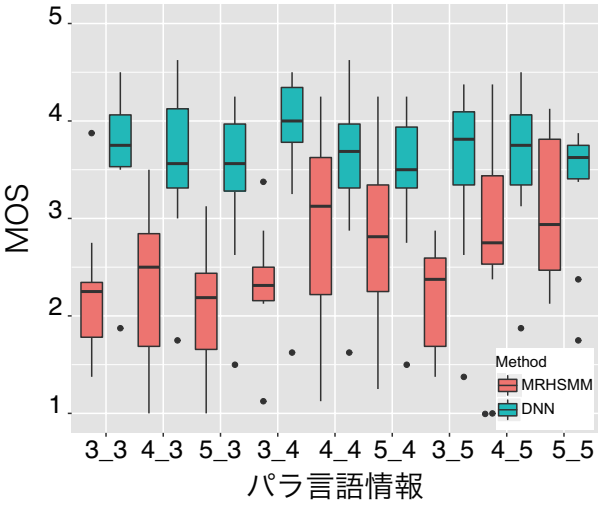

Fig. 3: 自然性評価実験結果

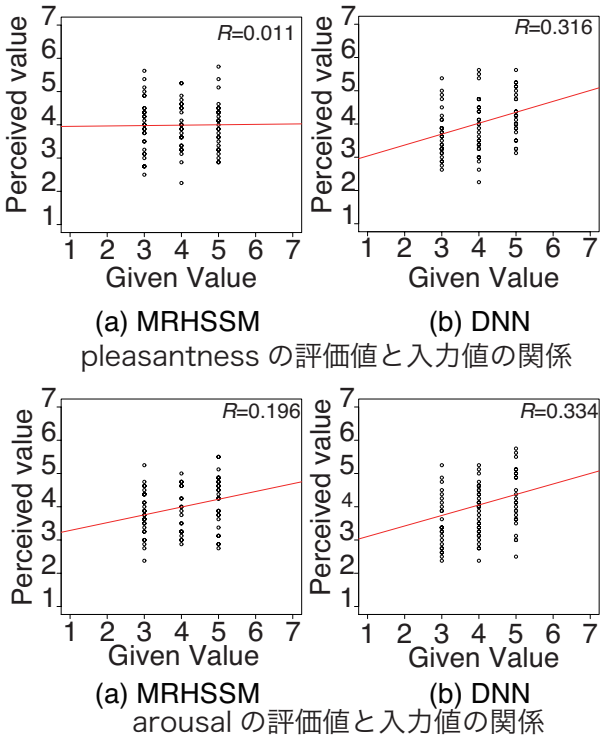

Fig. 4: パラ言語情報制御の知覚実験結果

方法を比較した。客観的な分析では，MRHSMM の方がパラ言語情報の変化により敏感な傾向にあることが分かった。一方で，DNN の方でも主観評価において違いが明確に知覚できていることが分かった。

参考文献

[1] Nagata *et al.*, Speech Commun., **88**(13), 137–148, 2017.
[2] Zen *et al.*, Proc. ICASSP 2013, 7962–7966, 2013.
[3] Mori *et al.*, Speech Commun., **53**, 36-50, 2011.
[4] Wu *et al.*, Proc. 9th ISCA Speech Synthesis Workshop (SSW9), 2016.
[5] M. Morise, Speech Commun., **84** ,57–65 ,2016.