DNN 音声合成における少数話者の感情表現を用いて任意話者の感情音声を合成する方式の評価

Evaluation of the method to synthesize many speaker's emotional speech
by transplanting a few speaker's emotional expressions in DNN-based TTS synthesis

井上 勝喜
Katsuki Inoue

岡山大学 阿部研究室
Abe Laboratory, Okayama University

概要 本研究では Deep Neural Network（DNN）を用いた音声合成において，目標話者の感情音声を使用することなく，目標話者の平常音声から所望の感情音声を合成することを目指す．DNN で音声から話者性と感情表現を分離し，これらを組み合わせることで所望の音声を合成可能であると考える．この方式を実現するために，DNN のモデル構造に着目する．本報告では 7 種類の DNN のモデル構造を提案し，提案方式の性能を主観評価を用いて評価した．

1 はじめに

DNN 音声合成とは文章から抽出した言語特徴量を音声から分析した音響特徴量へのマッピング関数として DNN を用いた音声合成方式である [1]．DNN 音声合成の概要図を図 1 に示す．DNN 音声合成に関する研究テーマのひとつは合成音声の多様化であり，話者や感情の多様化 [2, 3] が試みられている．感情の多様化の問題として，感情音声データの収集が困難である点がある．一般的な話者にとって発話の感情の適切な制御や，長時間に亘る感情の維持は困難であり，システム構築に十分な感情音声データの収録は難しい．プロのナレーターに依頼することで感情音声の収集は可能である．しかし，従来の多様化手法では DNN 学習に使用した特定話者の感情音声しか生成できず，一般話者の感情音声を生成できない問題の解決にはならない．また，HMM 音声合成では学習データに含まれていない感情音声の生成が可能な方式が提案されている [4]．

本研究では，HMM 音声合成における上記の方式を参考にして，ある話者の収録していない感情音声の生成を DNN に基づく音声合成で実現することを目指す．本報告では音響モデルと継続時間長モデルを用いた感情付与方式の全体を評価する．

2 提案法

「音声から話者性と感情表現を個別に学習させ，それらの組み合わせにより目標感情音声を生成する方式」を提案する．提案方式の概要図を図 2 に示す．モデル構造は全音声共通の構造と各話者専用・各感情専用の構造からなる．複数話者・複数感情の音声をモデル構造に学習させ，各話者専用の構造では話者性を獲得させ，各感情専用の構造では感情表現を獲得させる．生

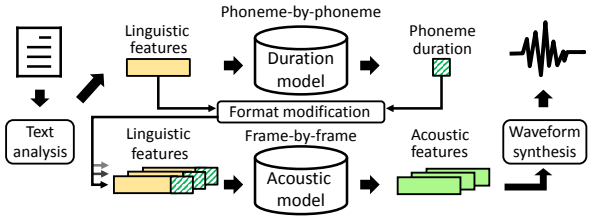

図 1: Deep Neural Network-based Text-to-speech synthesis system

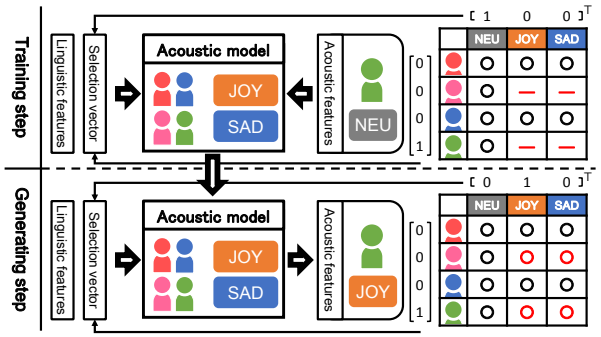

図 2: Proposed method

成時に話者と感情の構造の選択により，多様な感情音声を生成可能とする．他者の感情表現を付与した音声を借用感情音声，自身の感情表現を付与した音声を学習感情音声と定義する．話者と感情を制御するための特徴量として，話者選択ベクトルと感情選択ベクトルと呼ぶ 2 種類の one-hot ベクトルを使用する．

提案方式に用いるモデル構造として，parallel model（PM），serial model（SM），auxiliary input model（AIM）を提案する．PM と SM は話者性と感情表現に対応した構造に持ち，話者性と感情表現の明示的なモデル化を目的とする．PM は両構造を並列に配置し，SM は両構造を直列に配置する．SM として，話者・感情の順に配置した SM_{se} と感情・話者の順に配置した SM_{es} を用いる．また，AIM は話者性と感情表現を示す選択ベクトルを補助入力特徴量とすることで，話者性と感情表現の暗黙的なモデル化を目的とする．

3 評価実験

3.1 実験条件

本実験では，2 種類の日本語音声コーパス（男女各 2 名のコーパス α，男女各 6 名のコーパス β）を使用した．コーパス α では男性 1 名は平静音声（NEU）のみを発声し，残りの 3 名は平静音声と 2 つの感情音声（喜

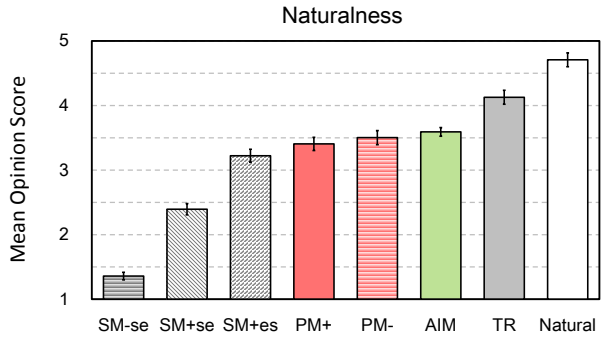

図 3: The results of MOS test with their 95% confidence interval

び：JOY，悲しみ：SAD）を発声している．平静音声と感情音声は同一のテキスト 500 文（約 35 分）を発声している．コーパス β では 12 名は同一のテキスト 130 文の平静音声（約 40 分）を発声している．音声信号は 16 bit，22.05 kHz サンプリングである．音声データの 90% を学習，5% を開発，残りの 5% を評価に用いる．

借用感情音声では借用する側の女性話者の感情音声を学習せず，残りの男女各 1 名の感情音声を学習させた．学習感情音声（Trained emotion : TR）では目標女性話者の感情音声も含め，PM+ に学習させた．

評価実験には 8 種類のモデル構造（PM±，SM±$_{se}$，SM±$_{es}$，AIM，TR）を使用した．継続時間長モデルの構築において，PM，SM，TR の入力特徴量ベクトルは 289 次元の離散的な言語コンテキストと 9 次元の数値的な言語コンテキストからなる．AIM の入力特徴量ベクトルは，話者選択ベクトルと感情選択ベクトルを加えて 316 次元となる．出力特徴量はフレーム数単位の継続時間長を表す整数のスカラ値である．

音響モデルの構築において，PM，SM，TR の入力特徴量ベクトルは継続時間長モデルと同じ入力特徴量に 7 次元の時間情報を加えている．AIM の入力特徴量ベクトルは 323 次元である．出力特徴量は対数 F0，40 次元のメルケプストラム係数，10 次元の帯域非周期性指標とそれらの Δ と $\Delta\Delta$ 特徴量，有声無声を示すフラグからなり，合計で 154 次元である．これらは STRAIGHT 分析（5 ms のフレームシフト）により，得られた特徴量から算出した．

3.2 主観評価実験

借用感情音声の性能評価のために，自然性と感情の認識率に関する主観評価実験を実施した．参加者は 10 名であり，暗騒音レベル 20 dB の防音室で実施した．借用感情音声（PM±，SM±$_{se}$，SM+$_{es}$，AIM），学習感情音声（TR），目標話者・目標感情の原音声の分析合成音声（Natural）を使用した．評価音声はモデルごとに 36 発話である．自然性では女性話者 A・B の JOY・SAD を 9 発話ずつ，感情の認識率では女性話者 A・B の NEU・JOY・SAD を 6 発話ずつ評価させた．

図 3 に自然性に関する主観評価実験結果を示す．借用感情音声において，PM と AIM が高い性能である．これらに比べ SM の性能は劣るため，SM は学習データ中の NEU の割合が高くなる不均衡の影響を受けやすいと考えられる．SM+$_{es}$ は SM+$_{se}$ より高い性能であるため，（感情，話者）の順にすることでデータの不均衡の影響を低減できると考えられる．借用感情音声（PM，AIM）は学習感情音声（TR）に比べ，自然性が低下している．音声の感情表現には個人差があり，借用感情音声では目標話者以外の感情を使用する．このため，借用感情音声と目標音声の誤差が学習感情音声に比べ大きくなったと考える．

表 1: Subjective emotional classification results (The value indicates the accuracy of classification.)

Model	NEU	JOY	SAD
Natural	**0.75**	**1.00**	**0.63**
TR	**0.92**	0.81	0.59
AIM	0.89	**0.62**	**0.43**
PM−	0.93	**0.74**	**0.58**
PM+	0.85	**0.68**	**0.52**
SM−$_{se}$	0.89	0.13	0.68
SM+$_{se}$	0.89	0.43	0.68
SM+$_{es}$	0.94	0.34	0.84

表 1 に感情の分類に関する主観評価結果を示す．Natural に注目すると，JOY の正解率は 100% であり，NEU や SAD に比べ明確に異なる音声であることが分かる．また，SAD の正解率は約 60% と低い値である．この結果を詳細に分析すると，SAD は NEU と誤分類されやすい傾向にあった．SAD の音声は JOY の音声に比べ，NEU の音声との差が小さく，分類が困難であったと考える．借用感情音声では PM と AIM が Natural に近い傾向である．これより，PM と AIM では他者の感情を付与した感情音声を生成できていると考えられる．また，SM では JOY の正解率がチャンスレートである 0.33 と同程度である．これより，SM では任意の感情表現を付与した感情音声の生成は困難であり，提案方式に適していないことが分かる．

4 まとめ

本報告では，音声合成における感情付与方式とモデル構築のための DNN のモデル構造を提案した．主観評価実験より，提案構造のうち PM と AIM は借用感情音声の生成において有効であることが示された．

今後の課題として，時間的に感情表現を徐々に変化させるための音声モーフィングなどを検討したい．

参考文献

[1] H. Zen *et al.*, *ICASSP*, pp. 7962–7966, 2013.
[2] N. Hojo *et al.*, *INTERSPEECH*, pp. 2278–2282, 2016.
[3] J. Trueba, *et al.*, Speech Communication, vol. 99, pp. 135–143, 2018.
[4] Y. Ohtani *et al.*, *INTERSPEECH*, pp. 274–278, 2015.