# 対話文脈エンコーダを利用した対話音声合成*

◎永田智洋, 森大毅 (宇都宮大)

## 1 はじめに

対話において発話は独立したものではなく文脈を持つ．この文脈は韻律などの音響的特徴に影響を与えることが知られている [1].

現在，統計モデルに基づく音声合成では音声の音響的特徴の変動要因を言語的特徴によって表現している．この言語的特徴は発話内で完結した情報のみが利用されていることがほとんどであり，対話における文脈は利用されていない．したがって，対話システムにおける合成音声は，1 発話を単独で評価した際の自然性は高くても，文脈にふさわしくないパラ言語的特徴を持つために不自然となる可能性がある．

そこで，本研究では対話文脈を考慮した対話音声合成を行う．主観評価実験により，合成音声の自然性および文脈における適切性を評価し，提案法の有効性を検討する．

## 2 対話文脈を考慮した音声合成

### 2.1 対話文脈エンコーダ

Deep Neural Network (DNN) に基づく音声合成では，言語的特徴と音響パラメータの関係を多層のニューラルネットワークでモデル化する [2]．本研究では言語的特徴に加えて，対話文脈を表現するための特徴 (対話文脈ベクトル) を利用することを提案する．すなわち，言語的特徴と対話文脈ベクトルを入力とし，対応する音響パラメータを出力するネットワークを構築する．

ニューラルネットワークの構成を Fig. 1 に示す．これは，(1) 2.2 節で説明する特徴を入力，対話文脈ベクトルを出力とする再帰ニューラルネットワーク (RNN)，(2) 言語的特徴と対話文脈ベクトルを入力，音響パラメータを出力とする RNN で構成される．本研究では (1) の RNN を対話文脈エンコーダ (DCE) と呼ぶ．対話文脈ベクトルはフレーム単位の言語的特徴とともに音響パラメータを出力する RNN に入力される．このとき，同一発話区間では，その発話に対応する対話文脈ベクトルが入力され続ける．

### 2.2 エンコーダ入力

本研究では以下の発話情報を DCE へ入力する特徴として使用する．

- 話者交替／話者継続
- 笑いの有無
- 感情状態 (快-不快, 覚醒-睡眠)

話者交替／継続は発話の韻律的特徴と関連があることが報告されている他，話速の発話間における加減速傾向とも関係があることが示唆されている [3]．話者交代／継続情報の利用はこれらの特徴を捉えることを期待している．

笑いは話者の感情などを伝達する社会的シグナルであり，対話において重要な役割を果たしている．笑いの有無の情報は当該発話だけではなく，後続発話に影響を与える (例: 笑いの誘発 [4]) と予想されるため，対話文脈として定義された．

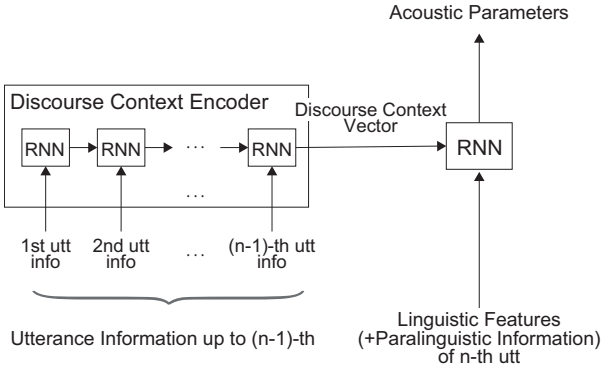


Fig. 1 The constitution of neural network.

本研究では次元説に基づいて表現された感情を発話情報として定義した．発話から知覚される感情が結果として同じであっても，その表出方法が一意であるとは限らない．当該発話に至るまでの感情系列を利用することによって，より対話文脈に適した感情表現が実現されることを期待する．

以上の情報を各発話に対して付与した．DCE へ入力する際には，当該発話までの発話情報を入力する．

### 2.3 合成音声の作成

本研究では，親しい友人同士の自然対話が収録されている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)[5] の女性話者 1 名の発話を用いて音声合成を行った．音響モデルおよび継続長モデルの学習には 586 発話 (約 16 分) が使用された．

言語的特徴は tri-phone やアクセント句，アクセント型などを表現する 564 次元のバイナリベクトルとした．更に，本研究では言語的特徴に加え，パラ言語情報を特徴としてネットワークに入力した．パラ言語情報は UUDB の各発話に付与されている「快-不快」，「覚醒-睡眠」の評価値とし，42 次元のバイナリベクトルで表現した．また，音響モデル学習時は更にフレーム特徴 4 次元を追加した．

DCE への入力は 2.2 節の情報を表現する 38 次元のバイナリベクトルとし，64 次元の対話文脈ベクトルに符号化した．

音響パラメータはサンプリング周波数 16 kHz の音声から抽出された 40 次元のメルケプストラム係数，対数基本周波数，1 次元の帯域非周期性指標，それぞれの $\boldsymbol{\Delta}$，$\boldsymbol{\Delta\Delta}$ パラメータ，および有声／無声情報である．

DCE には単方向 LSTM を使用し，入力層のユニット数は 38，中間層を 1 層，出力層のユニット数を 64 とした．音声合成用の RNN の入力層ユニット数は

*Dialogue speech synthesis with discourse context encoder. by NAGATA, Tomohiro, MORI, Hiroki (Utsunomiya University)

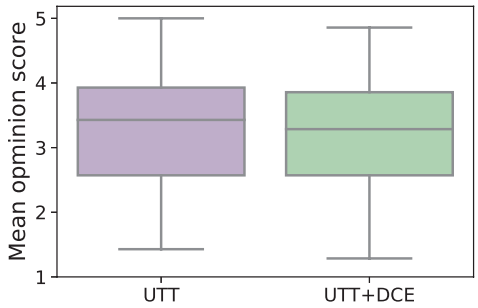


Fig. 2　The result of naturalness evaluation.

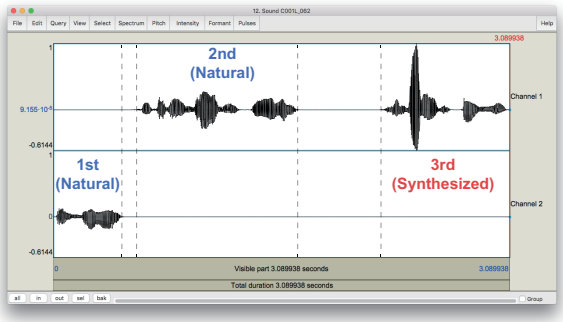


Fig. 3　An example of presented stimuli.

674(言語的／パラ言語的特徴 610+対話文脈 64) とした．中間層にはユニット数 512 の双方向 LSTM を利用し，出力層のユニット数は 127(音響モデル)，1(継続長モデル) である．

ネットワークから出力された音響パラメータからの波形合成には STRAIGHT ボコーダを用いた．

## 3　自然性評価実験

対話文脈を考慮することの有効性を検討するために，対話文脈を考慮せずに当該発話の情報だけから合成した音声 (UTT) と対話文脈を考慮して合成した音声 (UTT+DCE) を比較する．両者は以下の 2 つの条件で評価した．

- 合成音声を単独呈示
- 合成音声を対話の文脈で呈示

実験には，男子大学生 3 名および男子大学院生 4 名が被験者として参加した．実験は静かな実験室内による両耳聴取によって行われた．

### 3.1　単独呈示評価

単独の合成音声から知覚される自然性を 5 段階 (1:不自然, 5:自然) で評価するように指示した．作成した合成音声を 2 つのセットにわけて被験者に呈示した．各条件で合成された音声はセット間でシャッフルされた．1 セットの刺激の数は 83 個であり，刺激の総数は 166 個である．

自然性の平均評価値 (MOS) の分布を Fig. 2 に示す．UTT および UTT+DCE の平均 MOS はそれぞれ 3.29, 3.24 であり，両手法において有意な差は見られなかった ($t(82) = 0.73$, $p = 0.47$)．

### 3.2　文脈呈示評価

本実験では刺激として対話を利用する．各条件で合成された音声と，その音声が発せられるまでに出現した発話から呈示刺激を作成する．本実験では対話の長さを 3 発話とし，最後の発話を合成音声，それまでの発話を自然音声の分析合成音によって刺激を作成した．作成した刺激の例を Fig. 3 に示す．

作成した音声を 2 つのセットにわけて被験者に呈示した．各条件で合成された音声はセット間でシャッフルされた．被験者には呈示された音声の対話における適切性を 5 段階 (1:不適切, 5:適切) で評価するように指示した．

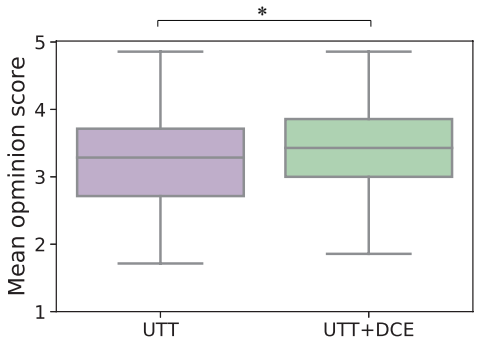


Fig. 4　The result of appropriateness evaluation.

被験者による MOS の分布を Fig. 4 に示す．図より，UTT+DCE の適切性が UTT よりも高いことがわかる ($t(82) = -2.03$, $p < .05$)．また，UTT および UTT+DCE の平均 MOS はそれぞれ 3.29, 3.44 である．これは UTT+DCE が，UTT よりも文脈に対して適切なパラ言語的特徴を合成音声に反映可能であったと考えることができる．以上の結果から，提案法の有効性が示された．

## 4　おわりに

本研究では対話文脈を考慮するために DNN 音声合成に対話文脈エンコーダを導入することを提案した．また，対話文脈を考慮することの評価法として合成音声だけではなく対話文脈も呈示する主観評価実験を行った．自然性評価実験により，提案法は合成音声の自然性を保ちつつ，文脈に対する適切性が改善することを示した．今後の課題として，対話文脈と音響的特徴との関係の調査が挙げられる．

## 参考文献

[1] 小磯 他, 人工知能学会研究会資料, 37, 139–144, 2003.
[2] Zen *et al.*, Proc. ICASSP, 7962–7966, 2013.
[3] 大須賀 他, 人工知能学会論文誌, 21 (1), 1–8, 2006.
[4] Trouvain and Truong, Proc. WASSS, 1–5, 2013.
[5] Mori *et al.*, Speech Communication, 53, 36–50, 2011.