2O4-OS-23a-4

対話行為情報を表現可能な音声合成の検討

北条 伸克 *1　井島 勇祐 *1　杉山 弘晃 *2
Nobukatsu Hojo　Yusuke Ijima　Hiroaki Sugiyama

*1日本電信電話株式会社 NTT メディアインテリジェンス研究所
NTT Media Intelligence Laboratories, NTT Corporation

*2日本電信電話株式会社 NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories, NTT Corporation

This paper describes a novel expressive text-to-speech (TTS) synthesis system for spoken dialogue systems. To this end, this paper focuses on dialogue act information. We first construct speech database with dialogue act information. Then, we construct a DNN-based TTS system can generate speech considering dialogue act information. The results of subjective experiments indicated the proposed technique can synthesize speech with improved naturalness.

1. はじめに

従来のタスク指向の対話システム [Williams 13] に対し，エンタテイメントやカウンセリング等を目的とした，雑談対話システム [Higashinaka 14, Ritter 11, Bessho 12] に注目が集まっている．雑談対話システムが音声で発話する場合，発話様式や感情表現と言ったパラ言語・非言語情報の伝達が重要となるため，単なる発話テキスト情報の伝達では不十分である．しかしながら，多くの雑談対話システムにおいては，読み上げ口調の音声を生成する音声合成システムが使用されている．そのため，パラ言語・非言語情報を表現可能な音声合成を行うことで，音声対話システム全体の自然性をさらに向上させることが可能と考えられる．

一方，近年の統計的音声合成技術 [Zen 09] の進展に伴い，テキストと，パラ言語・非言語情報を指定すれば，指定したパラ言語・非言語情報を持つ合成音声を生成することが可能である [Yamagishi 05]．このような技術を音声対話システム上で活用するためには，パラ言語・非言語情報を別途，指定する必要がある．パラ言語・非言語情報の指定のためには，対話シナリオ内の各システム発話について事前に人手で付与を行うか，推定器を構築し，システム発話から自動的に推定する方法が考えられる．しかし，人手による付与は高コストである上，雑談対話等のオープンドメインな対話システム上では実現することが困難である．また，推定器を用いる場合，推定誤りが生じた場合に，誤ったパラ言語・非言語情報に基づく音声合成を行うことで，音声対話システム全体の自然性を低下させる懸念がある．

このような課題に対し，本研究は，対話処理部から取得する情報として，システム発話のみでなく，対話処理部の内部から得られる対話行為情報を活用し，音声合成を行うアプローチを取る．対話行為情報とは，各発話の意図に相当する情報であり，雑談対話システム [Higashinaka 14] などの対話処理部で利用されている．システム発話の意図に対応した音声を合成することで，より文脈に適合した音声が得られるため，音声対話システムにおける合成音声の自然性が向上することが期待される．

対話行為情報に基づく音声合成を検討した先行研究として，[Syrdal 08] では，コールセンタの応答音声を用いた波形接続音声合成について，対話行為情報を利用することで，合成音声の自然性向上を達成している．[Tsiakoulis 14] では，hidden Markov model (HMM) 音声合成の枠組みにおいて対話行為情報を利用することで，合成音声の自然性が向上することが示されている．一方，近年，統計的音声合成の分野では，deep neural network (DNN) に基づく 音声合成 [Zen 13, Zen 14] が提案され，従来の HMM 音声合成に比べ高品質な音声が合成可能であることが示されている．本研究は，従来よりも高品質で，かつ表現力の高い音声合成を実現することを目指し，DNN 音声合成を用いて対話行為情報をモデル化する方法を検討する．

本研究では，対話行為情報に基づく音声合成を実現するため，まず，対話行為情報付き音声データベースを構築し，各対話行為情報の音声の特徴の分析を行う．分析を踏まえ，DNN 音声合成の枠組みにおいて，対話行為情報を表現可能な音声合成モデルを提案し，その有効性を検証する．

連絡先: 北条伸克，NTT メディアインテリジェンス研究所，神奈川県横須賀市光の丘 1-1，hojo.nobukatsu@lab.ntt.co.jp

2. 対話行為情報付き音声データベースの構築と分析

2.1 音声データベースの構築

本研究では，対話行為情報を表現可能な音声合成システムを構築するため，まず，各発話に対話行為情報が付与された音声データベースを構築する．音声合成用の音声データベースの構築では，発話者が音素バランス文を発声した音声を収録することが一般的である．しかし，パラ言語・非言語情報が含まれる自発性の高い音声を収録する場合，音素バランス文ではなく，収録対象のパラ言語・非言語情報に応じた文章を用いることが望ましいことが知られている．さらに，本研究で対象とする対話行為情報では，前後の対話の文脈等も発話者が発声する際に必要となると考えられる．そこで本研究では以下の条件を満たす音声データベースの構築を行う．

- 複数話者の対話形式の収録文章であり，意図や文脈を考慮した音声が収録されること
- 各対話行為情報の発話の出現回数のバランスが取られていること
- 各対話行為情報の発話における，各音素の出現回数のバランスが取られていること

収録テキストの元となるテキストデータとして，本研究では，テキストチャットコーパス [目黒 12] を使用する．テキストチャットコーパスには，テキスト対話が収録されており，そ

[34]

番号	話者	分類	発話
	A		スポーツにはちょっと疎くて、小学生から中学に上がる頃くらいまでテニスを習っていたきりです。
	A		Bさんは何かされますか?
	B		そうなんですか。
	B		私はバスケをしていました。
	B		テニスも趣味程度ですができます。
3401	A	自己開示_経験	チームプレイが求められるスポーツの経験がほとんどないので、
3402	A	承認	バスケが出来るというのは尊敬します。
3403	A	自己開示_事実	スラムダンクも読んでましたし。

図 1: 音声データベース構築に使用した収録文章例．赤枠で示された最終 1 発話を音声収録に用いる．

のテーマは 12 個で，食べ物，旅行，映画などである．2 名による対話形式であり，全体で 12 名の実験参加者による対話が収集されている．また，各発話に対して，専門家 2 名により，対話行為情報が付与されている．コーパス中の 1 対話には 30〜40 発話が含まれるが，本研究では対話の一部を，話者 A の発話 → 話者 B の発話 → 話者 A の発話のように，3 発話で切り出し，1 対話の単位として使用する．切り出された 1 対話の例を図 1 に示す．次に，切り出された対話について各対話行為情報の発話の出現回数，各対話行為情報の発話における各音素の出現回数のバランスを考慮して並べかえることにより，収録テキストを設計する．各対話行為情報の発話の出現回数，各対話行為情報の発話に含まれる音素の出現回数のバランスの尺度については，エントロピーを使用する．

本研究では，対話行為情報として，全 33 種類からなる対話行為情報 [目黒 12] から,「その他」などの対話行為情報を除いた全 29 種類 (挨拶:Grt，情報提供:Info，自己開示事実:Sd_f，自己開示経験:Sd_e，自己開示習慣:Sd_h，自己開示評価＋:Sd_+，自己開示評価-:Sd_-，自己開示評価 0:Sd_0，自己開示欲求:Sd_d，自己開示予定:Sd_p，相槌:Ack，質問情報提供要求:Q_i，質問事実:Q_f，質問経験:Q_e，質問習慣:Q_h，質問評価:Q_pr，質問欲求:Q_d，質問予定:Q_pl，共感・同意:Sym，非共感・非同意:N-Sym，確認:Conf，提案:Prop，繰り返し:Rept，言い換え:Para，承認:Apro，感謝:Thks，謝罪:Apol，フィラー:Fill，感嘆:Admn) を対象とする．

音声収録時には，図 1 中の赤枠で示された最終発話のみが発声され，それ以前の 2 発話については，音声収録時に，対話の流れを理解し，文脈に合わせた自然な発話を行うために収録文章に掲載される．音声収録時には，プロの女性声優 1 名による発話を収録した．対話の文脈と対話行為情報に応じ，自然な音声を発声するように教示を行った．以上の手続きにより，本研究では，音声長の合計約 180 分，合計 5177 文，各対話行為情報について 99 文以上からなる音声データベースを構築し，以下の音声分析，および音声合成手法の検討に使用する．

2.2 音声分析

本研究で使用する対話行為情報の分類には，共感・同意，非共感・非同意，自己開示_評価+，自己開示_評価-等，発話者の心理的側面を表現すると思われるものが含まれる一方，情報提供，質問，確認，提案等，対話の進行における役割を表現するものも含まれる．したがって，対話行為の中には，その発話が強く音声により表情付けられるものと，そうでないものが含まれると予想される．各対話行為について，その韻律の持つ特徴を明らかにするために，2.1 節で構築された音声データベースを

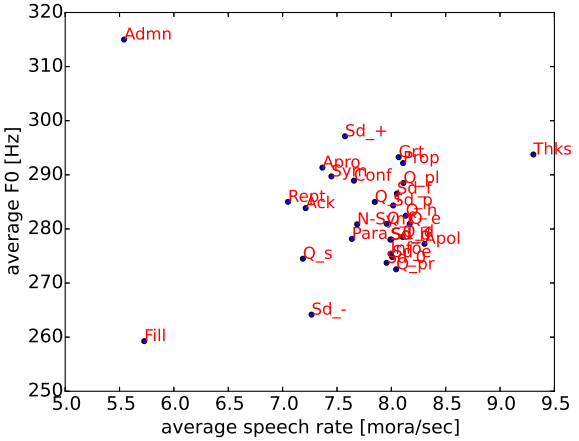

図 2: 対話行為情報別平均 F0，平均話速の散布図

使用した音声分析を行った．

各発話について算出された，話速と平均 F0 を，各対話行為情報ごとに平均したものを図 2 に散布図で示した．感嘆 (Admn)，フィラー (Fil)，感謝 (Thks) の 3 つの対話行為情報では，韻律の傾向が他と大きく異なることが明らかとなった．また，自己開示_評価+(Sd_+)，承認 (Apr)，共感・同意 (Sym) などの対話行為情報では，平均 F0 が他よりも高い傾向が現れた．これらの対話行為情報では，文章が全体として明るい口調で発話されたためであると考えられる．一方，文章が全体として暗い口調で発話された自己開示_評価-(Sd_-) では，平均 F0 が他よりも低い傾向が現れた．

また，図 2 から，29 種類の対話行為情報のうち，幾つかの組みについては，類似した韻律特徴を持つことがわかる．例えば，質問や情報提供などの対話行為情報は，平均 F0 と平均話速ともに類似しており，音声による表情付けは強く行われず，読み上げ口調に近い口調で発声されると考えられる．本研究では，このような対話行為情報の類似性の知見を 3 章の音声合成実験の考察時に活用するため，韻律特徴量に基づき，対話行為情報に関するクラスタリングを行った．クラスタリングの手法としては，各発話に対する対数 F0 の最大，最小，範囲，平均を標準正規化し，各対話行為情報ごとに平均化したものを特徴量として用いた．また，距離尺度としてユークリッド距離を使用し，階層的クラスタリングを行った．

クラスタリングの結果を図 3 に示した．クラスタリング結果から，29 種類の対話行為情報は大きく次の 4 つのサブクラスに分類することが可能であることが明らかとなった．それぞれのサブクラスには，

A. 感嘆
B. 謝罪，自己開示_評価-，フィラー
C. 承認，自己開示_評価+，共感・同意など
D. 読み上げクラス情報提供，質問，言い換え，繰り返しなど

が含まれる．B はネガティブな内容を表現する発話やフィラーなど，全体として F0 の低い発話が多く含まれ，C はポジティブな内容を表現する発話など，全体として F0 の高い発話が多く含まれる傾向にある．D にはあまり強く音声表現のなされない，読み上げらしい発話が多く含まれる．以下では便宜上，それぞれ A を感嘆クラス，B をネガティブクラス，C をポジティブクラス，D を読み上げクラスと呼ぶ．

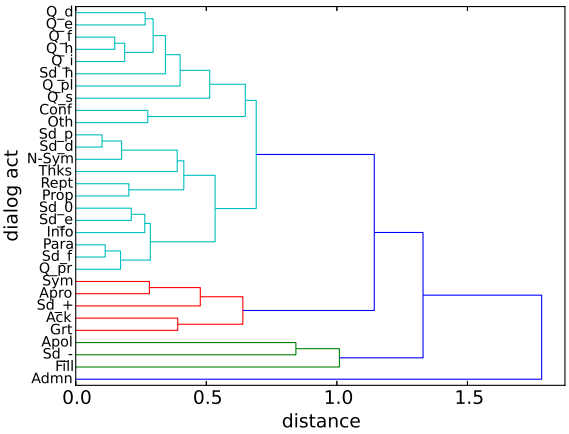

図 3: 対話行為情報のクラスタリング結果

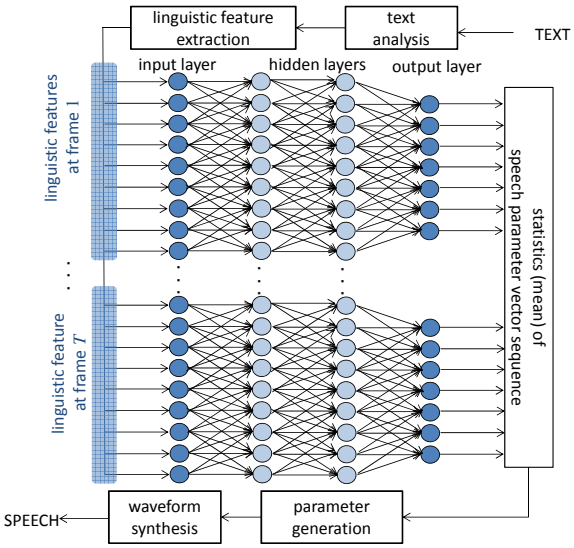

図 4: DNN 音声合成 (従来法) の概略図

3. 音声合成手法

本研究では，音声合成手法として，DNN 音声合成を使用する．DNN 音声合成は，従来の HMM 音声合成 [Yoshimura 99] と比べ，自然性の高い音声が合成されることが示されている．しかしながら，十分な品質の合成音声を得るためには，大量の音声データが必要であることが知られている．本研究で構築した音声データベースに含まれる各対話行為情報の音声データは少量であるため，29 種類の対話行為情報ごとに DNN を学習することは難しい．一方，我々はこれまで DNN 音声合成で他の話者の音声データを活用することで，学習対象となる話者の音声データ量が少量の場合でも高品質な合成音声が生成可能であることを示している [北条 15]．そこで，本研究では各対話行為情報の音声データ量が少量である場合にも，十分な品質の合成音声を生成するため，[北条 15] を対話行為情報へ拡張したモデルを提案する．

3.1 DNN 音声合成 (従来法)

従来法である，DNN 音声合成の概略を図 4 に示す．DNN に基づく音声合成方式 [Zen 13] では，各時刻 t における読みやアクセント等の言語情報に対応する言語特徴量ベクトル $\boldsymbol{x}^t$ と，音響特徴量ベクトル $\boldsymbol{y}^t$ 間のマッピングに，DNN が利用される．言語特徴量ベクトル $\boldsymbol{x}^t$ は，入力テキストをテキスト解析することにより得られ，各時刻 t における音素やアクセント型等を表現する．音響パラメータ $\boldsymbol{y}^t$ は，時刻 t におけるスペクトル特徴量や基本周波数 (F0)，およびその時間差分を表す．DNN の学習は，学習用音声の音響特徴量と，出力ベクトルとの平均二乗誤差最小化を基準とし，誤差逆伝播により行う．

3.2 対話行為情報を表現可能な音声合成 (提案法)

提案法では，複数話者の音声をモデル化し，音声合成を行う手法 [北条 15] を，複数の対話行為情報の音声のモデル化・合成のために拡張する．提案法では，対話行為情報をベクトルで表現し，ベースライン手法の言語特徴量ベクトルと連結し，ニューラルネットワークに入力する．

具体的には，対話行為情報の全種類数を $M(=29)$ としたとき，各対話行為情報 $c(c = 1, \cdots, M)$ に対応する，対話行為コード $\boldsymbol{z} = [z_1, \cdots, z_M]$ は次式で表現される．

$$z_m = \begin{cases} 1 & (m = c) \\ 0 & (m \neq c) \end{cases} \tag{1}$$

対話行為コードを各時刻の言語特徴量ベクトル $\boldsymbol{x}^t = \{x_1^t, \cdots, x_N^t\}$ に連結し，音響モデルの入力ベクトル $\boldsymbol{v}^t$ とする．

$$\boldsymbol{v}^t = [x_1^t, \cdots, x_N^t, z_1, \cdots, z_M] \tag{2}$$

モデル学習時には，学習データに対応する $\boldsymbol{v}^t$ と $\boldsymbol{y}^t$ から，平均二乗誤差最小化基準でネットワーク全体のパラメータを学習する．音声合成時には，同様に対話行為コードを言語特徴量ベクトルに連結したものを DNN へ入力し，順伝播を行うことで音響特徴量が得られる．

4. 主観評価実験

本章では，提案法を用いて合成された音声の，自然性の向上について検証する．

4.1 実験条件

提案法により合成される音声が対話音声として自然であるかを評価するため，上記の 2 つの音声合成手法により合成される音声について，対比較による主観評価実験を行った．

提案法の DNN 音響モデルの学習のためには，2 章において構築した対話行為情報付き音声データベースの全発話 (5177 発話) のうち，音響モデル学習のために必要なラベルが整備された 3769 発話を使用した．うち，3519 発話がモデル学習用のデータとして使用され，250 発話が評価用データとして使用された．一方，従来法については，韻律の傾向が大きく異なる発話が混在する学習データから音響モデルを学習した場合，合成音声の韻律が不安定になる懸念がある．このため，本稿では，従来法の学習データとして，2.2 節のクラスタリングの結果，読み上げクラスに分類された対話行為情報の発話を使用した．読み上げクラスに分類された対話行為情報の発話は，3519 発話のうち，1881 発話であった．

音声分析のサンプリング周波数は 22.05kHz，特徴量抽出の際のフレームシフトは 5ms とし，スペクトル特徴量として，STRAIGHT 分析 [Kawahara 99] から得られる 40 次メルケ

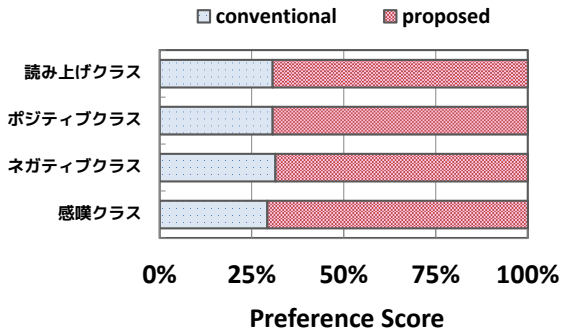

図 5: 合成音声の対比較結果.

プストラムを使用した．また，5 次元の非周期性指標を使用した．DNN の隠れ層は提案法，従来法ともに 4 層，ユニット数は 512 とした．本実験では，音素継続長は収録音声のものを利用した．音素継続長のモデルについても，提案法と同様のモデル化が可能であると考えられる．

合成音声の評価手順として，まず被験者は図 1 で例示される対話文章を読み，対話の流れを理解した上で，赤枠で囲まれた最終発話に対する 2 つの合成音声を比較し，文脈と当該文章の発話意図に対応する音声として，より自然である音声を選択した．被験者は 5 名とした．評価用文章として，2.2 節で得られた各対話行為クラスについて，20 文をそれぞれ使用した．

4.2 実験結果

実験結果を図 5 に示した．結果から，全対話行為クラスについて，提案法がより高いスコアを示し，有意な差が存在することが確認された (有意差水準 5%)．このことから，対話行為情報を活用する提案法により，各対話行為クラスについて，合成音声の自然性が向上することが示された．提案法により，対話の流れや対応する意図を反映した音声を合成することが可能となり，いずれの対話行為情報についても，その合成音声の自然性が向上したためであると考えられる．

5. 結論

本研究では，各システム発話の意図を表現する音声を合成することにより，より自然な音声を合成可能な音声対話システムを構築することを目指し，システム発話の対話行為情報を使用した音声合成を提案した．本研究では，テキストチャットコーパスを基にした収録文章を使用し，対話行為情報付き音声データベースを構築した．構築された音声データベースを用いた音声分析により，感嘆，フィラーなどの対話行為情報で特に，他と異なる韻律を持つ音声が発話されることを明らかにした．また，韻律特徴量を用いた対話行為情報のクラスタリング結果から，本研究で利用した全 29 種の対話行為情報は，4 種の対話行為クラスに分類可能であることが示された．音声合成実験では，従来の DNN 音声合成の枠組みにおいて，新たに対話行為情報を入力ベクトルとして活用する音声合成方式を提案した．主観評価実験により，提案法の合成音声が，対話の文脈と当該文章の対話行為情報に対応する音声として，より自然であることを確認した．今後の課題として，対話行為情報を利用した音素継続長のモデル化，話者の多様化，音声対話システム上で提案法を利用した場合の効果の評価に取り組む予定である．

参考文献

[Bessho 12] Bessho, F., Harada, T., and Kuniyoshi, Y.: Dialog system using real-time crowdsourcing and Twitter large-scale corpus, in *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 227–231Association for Computational Linguistics (2012)

[Higashinaka 14] Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., and Matsuo, Y.: Towards an open-domain conversational system fully based on natural language processing., in *COLING*, pp. 928–939 (2014)

[Kawahara 99] Kawahara, H., Masuda-Katsuse, I., and De Cheveigné, A.: Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds, *Speech communication*, Vol. 27, No. 3, pp. 187–207 (1999)

[Ritter 11] Ritter, A., Cherry, C., and Dolan, W. B.: Data-driven response generation in social media, in *Proceedings of the conference on empirical methods in natural language processing*, pp. 583–593Association for Computational Linguistics (2011)

[Syrdal 08] Syrdal, A. K. and Kim, Y.-J.: Dialog speech acts and prosody: Considerations for TTS, in *Proceedings of Speech Prosody*, pp. 661–665 (2008)

[Tsiakoulis 14] Tsiakoulis, P., Breslin, C., Gasic, M., Henderson, M., Kim, D., Szummer, M., Thomson, B., and Young, S.: Dialogue context sensitive HMM-based speech synthesis, in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 2554–2558IEEE (2014)

[Williams 13] Williams, J., Raux, A., Ramachandran, D., and Black, A.: The dialog state tracking challenge, in *Proceedings of the SIGDIAL 2013 Conference*, pp. 404–413 (2013)

[Yamagishi 05] Yamagishi, J., Onishi, K., Masuko, T., and Kobayashi, T.: Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis, *IEICE TRANSACTIONS on Information and Systems*, Vol. 88, No. 3, pp. 502–509 (2005)

[Yoshimura 99] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T.: Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis, in *In Proc. EUROSPEECH 1999*, Vol. 6, pp. 2347–2350 (1999)

[Zen 09] Zen, H., Tokuda, K., and Black, A. W.: Statistical parametric speech synthesis, *Speech Communication*, Vol. 51, No. 11, pp. 1039–1064 (2009)

[Zen 13] Zen, H., Senior, A., and Schuster, M.: Statistical parametric speech synthesis using deep neural networks, in *In Proc. ICASSP 2013*, pp. 7962–7966 (2013)

[Zen 14] Zen, H. and Senior, A.: Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis, in *In Proc. ICASSP 2014*, pp. 3844–3848 (2014)

[北条 15] 北条 伸克, 井島勇祐, 宮崎昇：話者コードを用いた DNN 音声合成の検討, 日本音響学会秋季講演論文集, No. 2–1–1, pp. 215–218 (2015)

[目黒 12] 目黒 豊美, 東中竜一郎, 堂坂浩二, 南泰浩：聞き役対話の分析および分析に基づいた対話制御部の構築, 情報処理学会論文誌, Vol. 53, No. 12, pp. 2787–2801 (2012)