Vol.2015-MUS-107 No.72
2015/5/24

多様な対話音声合成のための話し言葉音声コーパスの構築と評価

山田 修平[1,a)]　能勢 隆[1]　伊藤 彰則[1]

概要：本稿では，表情豊かな音声対話システムを実現するため，多様な口調や感情を持った話し言葉音声コーパスを構築し，HMM 音声合成において評価した結果を報告する．これまでに大規模な話し言葉音声のコーパスとして日本語話し言葉コーパスが提供されているが，主に音声認識や対話の分析を目的としているため，高品質な音声合成のために利用することは困難であった．また，これまで，感情表現や発話様式を伴う音声を用いた研究は数多く報告されているが，特定の話者を想定し，複数の感情でかつ多様な口調を含む音声コーパスを構築した例は少ない．そこで本研究では，ナレーションのような朗読調の音声と対話における話し言葉調の両方の口調を含みかつ複数の感情を表現可能な特定話者の音声コーパスを構築し，それを用いて学習した場合の合成音声について客観評価を行い，その有効性を示す．

1. はじめに

近年，タスク指向型対話だけでなく，雑談が可能な対話エージェントの研究が盛んに行われている [1], [2]．これらのうち，音声を用いた対話エージェントの利用例としては，NTT ドコモのアプリ「しゃべってコンシェル」がある．本アプリは，スマートフォンを操作するためだけでなく，雑談相手としても用いているユーザも多いという報告がある [3]．他にも，対話ロボットが高齢者の雑談相手となることにより，高齢者同士の交流や外出が増えるなどの効果があると報告されており [4]，高齢者の介護予防などの場面にも，ロボットとしての対話エージェントが活用されていくことが予想される．これまでの研究において，対話エージェントの発話音声に適切に感情を付与することにより，ユーザのエージェントへの印象が上昇することが分かっている [5]．このように，人間同士に近い対話を人間とエージェントが行うためには人間と同様に感情を含みかつ話し言葉口調の音声を合成することが必要となる．

音声合成の手法としては，近年 HMM に基づくアプローチが広く研究されており，その有効性が示されている [6]．その理由としては，高品質で自然性の高い音声が合成できることや，学習用の音声に含まれる話者性や感情表現・発話様式がよく反映されることが挙げられる [7], [8]．一方，HMM 音声合成はコーパスベースの音声合成手法であり，学習データとしてどのようなコーパスを構築するかが非常に重要である．これまでの研究として，特定のキャラクタ (人物) を対象とし，学習用の文コーパスの作成法の研究が行われている [9]．文献 [9] において，ATR 音素バランス文の文末部分を変更することや，キャラクタ性のある文をコーパスに用いることで，自然性・明瞭性が向上することや，キャラクタ性の再現が可能になることが示されている．

本研究では，文献 [9] をさらに発展させ，複数の感情を持つ対話エージェントを実現するため，音声合成を目的とした特定話者の話し言葉音声コーパスの構築について検討し，HMM 音声合成により評価を行う．対話エージェントは話し言葉口調だけでなく，場面によってナレーションなどの朗読調の音声を用いるべきであると考えられるため，朗読調の音声と対話調の音声のどちらも合成できる対話エージェントのための音声コーパスの構築を目標とする．音声の収録には各感情に依存した話し言葉文や朗読調の文が必要となり，このような文を自動で生成・収集するには大きなコストがかかる．そこで我々は文献 [9] を参考に，ATR 音素バランス文を利用し，その文末表現を話し言葉調に変更したものや，比較的少量の対話文を新たに作成し，これらを用いて実験を行う．

2. コーパスの構築

2.1 文コーパスの作成

HMM 音声合成のための学習用音声は，音韻や韻律についてバランスがとれている必要がある．また生成したい音声の感情や口調を含んでいる必要がある．ATR 音素バラ

[1] 東北大学 大学院工学研究科
Graduate School of Engineering, Tohoku University
[a)] s.yamada@dc.tohoku.ac.jp

対話的に変更
ふりあおぐとすぐ頭上を光が走った。
→ふりあおぐとすぐ頭上を光が走っていたわ。
人びとは非難し悩みながら歩み寄っていく。
→人びとは非難し悩みながら歩み寄っていくのよ。
対話的かつ疑問文に変更
私の指には宝石の指輪はもうはめられません。
→私の指には宝石の指輪はもうはめられませんか？
汚れた窓から雨にぬれた街が見える。
→汚れた窓から雨にぬれた街が見えるかな？

図 1 変更した ATR 音素バランス文の例
Fig. 1 Examples of modified ATR phoneme-balanced sentences.

楽しげ
ありがとうございます。
久しぶり。元気だった？
怒り
それ以上言うと怒るわよ。
真面目にやる気があるんですか？
悲しげ
自信なくしちゃうなぁ。
どうして，そんなひどいことを言うんですか？

図 2 対話文の例
Fig. 2 Examples of dialogue sentences.

ンス 503 文は，音素のバランスが取れた文コーパスであるが，その文章は対話的でないため，それのみを対話エージェントの学習に使うことは好ましいとはいえない．一方で，バランスがとれた対話的文章を一から作成することは人的コストが高い．そこで，ATR 音素バランス文の文末部分を 1 名の主観により対話的に変更することにより文を用意した．文末部分を対話的に変更する際には，同じものをなるべく使用せず，様々なパターンを用意するように指示した．変更の方法は以下の通りである．

- ATR 音素バランス文から，200 文 (未変更)
- 200 文のうち，50 文の文末部分を対話的に変更したもの
- 200 文のうち，50 文の文末部分を対話的な疑問文に変更したもの

未変更のものについては，朗読調の音声の合成を目的としたものである．変更した文章の例を図 1 に示す．次に，対話文について述べる．感情に依存した対話文を収集することは容易ではないことから，本研究では SS(ショートストーリー) の執筆経験がある学生 1 名に依頼し，エージェントがユーザと雑談で話すような場面を想定し，各感情に応じた自然な対話文を作成してもらった．ただし，対話文については音素のバランスは考慮しなかった．感情についての内訳は楽しげが 248 文，怒りが 104 文，悲しげが 36 文である．その例を図 2 に示す．

表 1 スタイルと文種ごとの対数 F0 [cent] の平均
Table 1 Average of log F0 for each speaking style and sentence style.

スタイル	original	modified	dialogue
楽しげ	10084	10062	10213
怒り	10308	10299	10293
悲しげ	9412	9411	9479

表 2 スタイルと文種ごとのモーラ長 [ms/mora] の平均
Table 2 Average of speaking speed for each speaking style and sentence style.

スタイル	original	modified	dialogue
楽しげ	136.0	137.6	145.1
怒り	136.9	134.9	154.2
悲しげ	148.6	148.6	169.4

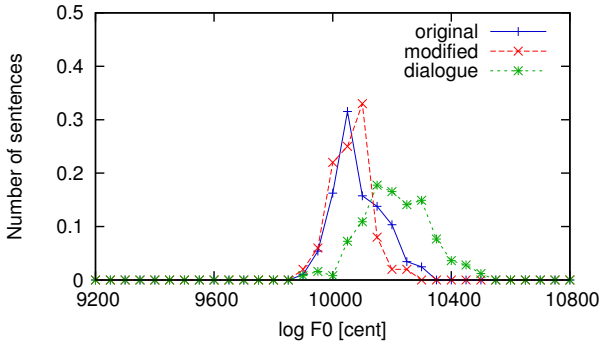
2.2 音声の収録

前述の ATR 音素バランス文セット (original) とそれに変更を加えた文セット (modified)，および感情依存の対話文セット (dialogue) を，プロの女性声優 1 名が楽しげ，悲しげ，怒りの 3 スタイルで発話したものを収録した．音声は，サンプリング周波数 48kHz，量子化ビット数 16bit で，スタジオの防音室にて収録した．これらの文全てを連続して収録することは不可能であるが，同一のスタイルの文を日を空けて収録した場合に，声質や感情の表現方法が異なってしまうことが考えられる．これを抑えるため，声優には感情毎に最初に録音した数発話を適時聞き直してもらい，スタイルをなるべく一定になるように収録を行った．また，文献 [9] の研究で報告されているように，文末部分変更前後や対話文の音声の収録日が異なるために発話様式が異なってしまう現象を防ぐため，音素バランス文，文末を変更した音素バランス文，対話文をそれぞれ別々に収録するのではなく，それらを一定の割合で混合して収録した．これにより，対話文と音素バランス文の間の韻律的な差異を抑えることが期待できる．

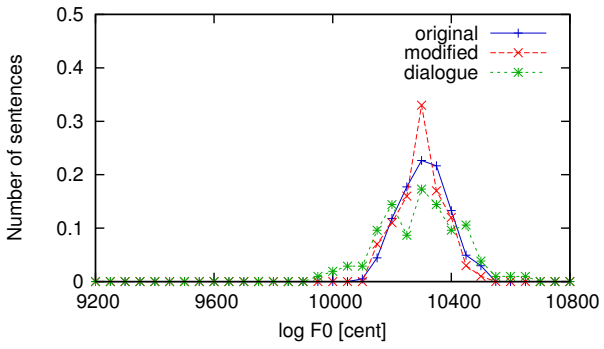
2.3 収録音声の韻律的特徴

各文セットについて収録した音声の韻律的特徴を調べた．各発話の平均対数 F0 の分布をスタイル毎に図 3 に示す．なお，これらの図は各文セットの文章数について正規化されている．各発話の平均対数 F0 の，文セットごとの平均を表 1 に示す． スタイルと文種ごとのモーラ長の平均を表 2 に示す．なお，発話前後の無音区間とポーズは除いて計算している．これらの図および表より，楽しげと比べ悲しげの発話は F0 が低くモーラ長が長い (話速が遅い) ことと，楽しげと比べ怒りの発話は F0 が高いことが分かる．

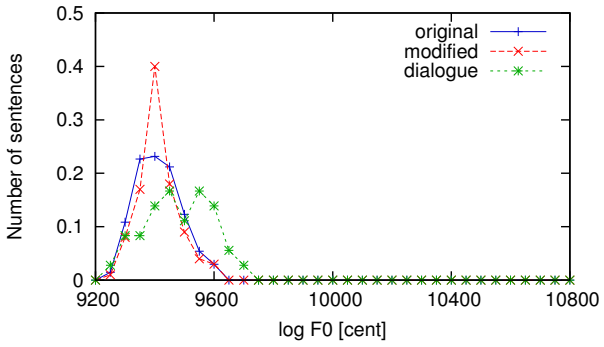
次に，各発話の平均モーラ長と平均対数 F0 の関係をスタイル毎に図 4 の散布図に示す．図から，対話文は音素バランス文に比べ，平均対数 F0 の分散が大きいことが分かる．



(a) 楽しげ

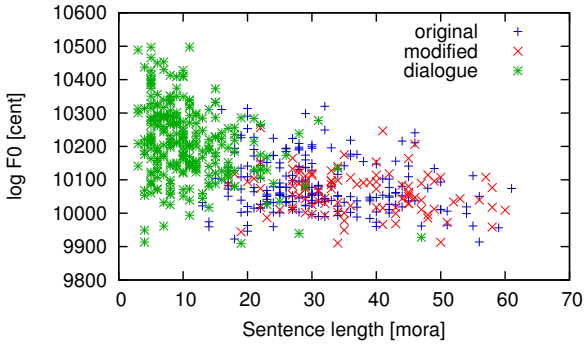


(b) 怒り

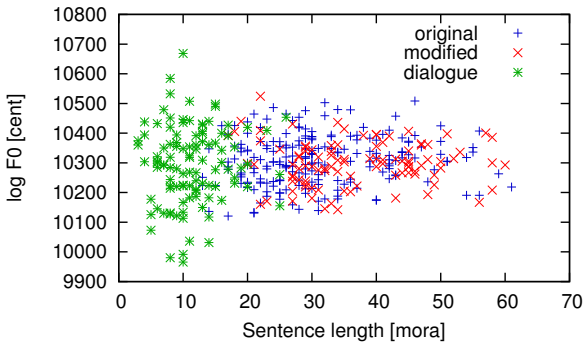


(c) 悲しげ

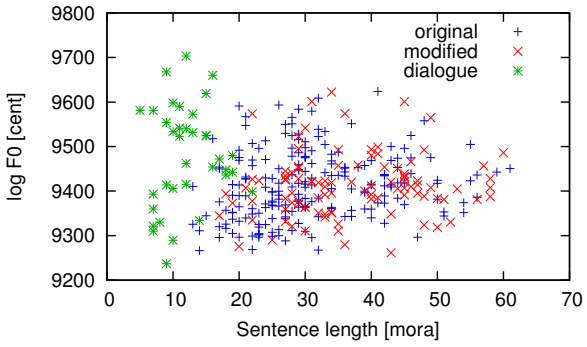
図 3 対数 F0 の分布
Fig. 3 Histogram of log F0.



(a) 楽しげ



(b) 怒り



(c) 悲しげ

図 4 発話の文長と平均対数 F0 の関係
Fig. 4 Relation between length and mean log F0 of utterance.

これは，文章が短いため，発話内容によって話し方が異なり，発話間の違いが大きくなったことが原因だと考えられる．荒生ら [9] の先行研究では，音素バランス文 (文末変更あり・なし) と対話文の対数 F0 の平均値が 173 cent 違うことが問題点として取り上げられていて，その原因に収録時期の違いが挙げられている．本研究では，収録時期の違いに韻律的特徴が影響されないように，2 節で述べた通り収録順序を考慮した．しかしながら，音素バランス文と対話文の間には対数 F0 の差が最大で 151 cent あった．音素バランス文 (文末の変更あり・なし) の間の対数 F0 の差は最大でも楽しげの 22 cent であることと，収録順序を考慮しても音素バランス文と対話文に対数 F0 の差があったことから，これらの差は収録時期の違いの影響ではなく，発話内容の違いに影響された差であると考えられる．またモーラ長に関しては，荒生ら [9] の先行研究では，音素バランス文 (文末変更あり・なし) の間に 12 ms/mora の差があったものの，本研究ではそれらの間には最大でも 2.0 ms/mora の差にとどまっており，収録順序の工夫の効果が現れている．しかし，音素バランス文と対話文の間にはモーラ長の差が 7.5〜19.3 ms/mora あり，発話内容が大きく違うことによりモーラ長にも差が生じたのだと考えられる．

本研究では，音素バランス文 (文末変更あり・なし) はスタイルに関係なく同じ文章を用いて収録したが，音素バランス文および変更した文末は必ずしも各発話スタイルに即しているとは限らず，普通は用いられないと考えられる文章とスタイルの組み合わせも存在する．一方，対話文は各発話スタイルを想定して作成しているため，全て発話スタイルに即している文章である．そのため，音素バランス

文より対話文の方が感情を表現しやすく，感情表出の度合いが違ったことで，音素バランス文と対話文の間での対数F0の平均やモーラ長の平均の違いが起こったのだと考えられる．

3. 客観評価実験

本研究では，構築したコーパスから文を組み合わせていくつかの学習セットを作成し，各学習セットを用いて朗読調の文と対話調の文を合成した．合成した音声と自然発話音声の特徴量の距離を求めることにより，学習セットごとの合成音声の品質を客観的に評価した．また，本研究が目標としている，朗読調の文章も対話調の文章も合成できるコーパスとしてのバランスを確認するため，朗読調の文と対話調の文を合成した時の特徴量の距離の差の絶対値を比較した．

3.1 実験条件

音声波形のサンプリング周波数が16 kHzの場合，メルケプストラムは0次〜39次のものが使われることが多い．しかし，本研究での音声波形のサンプリング周波数は48 kHzである．Adriana [10] らはメルケプストラムが50，55，60，65，70次元の場合について分析合成音を主観的に比較し，60次元のものを用いている．本研究ではそれに従い，メルケプストラムは0次〜59次のものを用いた．音声特徴量は，STRAIGHT [11] によって抽出したメルケプストラム (0次〜59次)，対数F0，非周期性指標 (5次元) と，それぞれ1次，2次の動的特徴量，計198次元の特徴量ベクトルを使用した．フレームシフトは5 msとした．音声合成にはHTS [12] を使用し，音響モデルは，5状態 left-to-right HSMMを用いた．また，出力分布は単混合ガウス分布，共分散行列は対角を仮定した．

3.2 学習セットの作成

本研究では，収録した音声を次の通り組み合わせ，3つの学習セットを作成した．

ATR 文末部分変更前のATR音素バランス文150文，計5040モーラ．

mod 文末部分変更前のATR音素バランス文50文に加え，文末を変更した文100文 (50文は平叙文，50文は疑問文)．計150文，5220モーラ．なお，文末部分以外はATRと同一の文である．

mix ATRセットよりランダムに抽出したATR音素バランス文126文 (4215モーラ) に加え，対話文92文 (1009モーラ)．計218文，5224モーラ．

ATRセットおよびmodセットは楽しげ，怒り，悲しげの3スタイルについて作成した．mixセットは，十分に対話文の量を確保できた楽しげのスタイルについてのみ作成した．

3.3 文末部分を変更する手法の検討

まず，音素バランス文の文末を対話的に変更した学習セットを使用した場合の効果について検討する．ATRセットとmodセットで学習し合成した音声の各特徴量について，自然発話音声との距離を求め，学習セットの違いによる差を比較した．ここで比較した特徴量は，メルケプストラム距離 [dB] (mcep)，対数F0のRMSE [cent] (lf0)，音素継続長のRMSE [ms] (dur) である．

評価セットは，以下の2つを作成した．

read 文末部分変更前のATR音素バランス文50文，計1223モーラ．

dialogue 用意した対話文．楽しげが248文 (2648モーラ)，怒りが104文 (1187モーラ)，悲しげが36文 (435モーラ)．スタイルにより文数が異なるが，全ての文を使用し比較．

文全体を用いて評価した場合の結果を表3に示す．readセットを合成した場合は，ATRセットで学習した方が誤差がわずかに小さくなることが多かった．modセットで学習した方が若干誤差が小さくなった項目は悲しげのメルケプストラム距離 (0.038 dB)，怒りの音素継続長 (0.8 ms) であるが，大きな差とは言えないと考えられる．一方，対話文を合成した場合は，文末を変更した音素バランス文であるmodセットで学習した全てのスタイルおよび特徴量について誤差が小さくなった．

さらに，文末を対話調に変更することの文末部分への効果を検討するために，各発話の最終モーラについて同様に客観評価を行った．結果を表4に示す．readセットを合成した場合は，怒りの音素継続長以外全て，ATRセットで学習した方が誤差が小さくなった．readセットを合成した場合においても，modセットで学習した方が怒りの音素継続長の誤差が11.4 ms/mora小さくなったことの原因は，今後検討が必要である．dialogueセットを合成した場合は，文全体で評価をした場合と同様に，全てのスタイルおよび特徴量についてmodセットで学習した方が誤差が小さくなった．文全体を評価した時に比べ，最終モーラのみ評価した時の方がより誤差が小さくなったため，対話文の文末周辺は特に自然発話音声に近づいたと言える．

readセットを合成した時の距離とdialogueセットを合成した時の距離の差の絶対値については，文章全体で評価した時も，最終1モーラのみ評価した時も，全てのスタイルについて2つのセットの距離の差の絶対値がmodセットの方が小さくなったことから，ATRセットよりmodセットの方が，朗読調の文章も対話調の文章も合成できるコーパスとしてバランスが取れていると言える．

これらの結果より，音素バランス文の文末を対話的に変更することによって，対話文の客観評価結果が改善することが分かった．改善は特に文末の対数F0で顕著であった．これは，ATR音素バランス文に含まれない，対話的な文

表 3　mod セットで学習した時のスタイルごとの客観評価結果
Table 3　Result of objective evaluation for each speaking emotion when learned with mod set.

評価セット	スタイル	学習セット	mcep	lf0	dur
read	楽しげ	ATR	5.69	248.8	29.3
		mod	5.73	249.4	29.8
	怒り	ATR	6.12	245.1	32.9
		mod	6.14	248.7	32.1
	悲しげ	ATR	5.49	181.9	31.4
		mod	5.45	184.4	34.1
	平均	ATR	5.77	225.3	31.2
		mod	5.78	227.5	32.0
dialogue	楽しげ	ATR	6.04	323.0	38.7
		mod	6.01	288.2	34.9
	怒り	ATR	6.29	349.2	38.6
		mod	6.25	337.3	36.9
	悲しげ	ATR	5.67	254.9	46.0
		mod	5.45	246.5	41.5
	平均	ATR	6.00	309.1	41.1
		mod	5.90	290.7	37.8
read, dialogue の差の絶対値	楽しげ	ATR	0.35	74.2	9.4
		mod	0.27	38.8	5.1
	怒り	ATR	0.17	104.2	5.7
		mod	0.10	88.6	4.9
	悲しげ	ATR	0.18	73.0	14.6
		mod	0.00	62.1	7.4
	平均	ATR	0.23	83.8	9.9
		mod	0.13	63.2	5.8

表 4　mod セットで学習した時のスタイルごとの最終モーラの客観評価結果
Table 4　Result of objective evaluation of the last mora for each speaking emotion when learned with mod set.

評価セット	スタイル	学習セット	mcep	lf0	dur
read	楽しげ	ATR	5.37	324.7	24.7
		mod	5.91	351.9	30.1
	怒り	ATR	5.96	341.1	66.7
		mod	6.29	430.1	55.3
	悲しげ	ATR	5.31	196.4	27.3
		mod	5.79	248.4	50.5
	平均	ATR	5.55	287.4	39.6
		mod	6.00	343.5	45.3
dialogue	楽しげ	ATR	6.51	562.5	69.1
		mod	6.21	401.1	44.2
	怒り	ATR	6.42	553.4	58.1
		mod	6.27	470.2	52.6
	悲しげ	ATR	6.71	419.2	73.0
		mod	5.81	337.1	61.5
	平均	ATR	6.55	511.7	66.7
		mod	6.10	402.8	52.8
read, dialogue の差の絶対値	楽しげ	ATR	1.14	237.9	44.4
		mod	0.30	49.1	14.1
	怒り	ATR	0.47	212.2	8.6
		mod	0.03	40.0	2.6
	悲しげ	ATR	1.40	222.8	45.7
		mod	0.02	88.7	11.1
	平均	ATR	1.00	224.3	27.1
		mod	0.10	59.3	7.5

末表現や疑問調が表せるようになったことによる効果だと考えられる．また，楽しげ，怒り，悲しげのいずれのスタイルについても，対話文らしい音声を表現できるようになり，複数のスタイルを含むコーパスに関しても，このような手法を取ることによりバランスのよいコーパスが作成できることが分かった．

3.4　対話文を混合する手法

次に，音素バランス文と対話文を混合した学習セット (mix) を用いた手法の効果を検討する．この手法については，怒りや悲しげのスタイルについて用意できた対話文が少なかったため，楽しげのスタイルについてのみ評価を行った．なお，mix セットについては，学習と合成に用いるラベルに朗読調の文 (音素バランス文) と対話文のどちらであるかの情報を付与した．評価セットは以下の 2 つを作成し，mod セットのみを用いた実験と同じ特徴量について評価を行った．

read　文末部分変更前の ATR 音素バランス文 50 文，計 1223 モーラ (mod セットのみを用いた実験と同じ)．

dialogue　用意した対話文 (楽しげのみ)．収録データから mix の学習に用いたものを除き，156 文 (1639 モーラ) を使用．

文全体での評価結果を表 5 に示す．read セットを合成した場合は，mix セットで学習した場合の対数 F0 を除いて，ATR セットで学習した方が誤差がわずかに小さくなった．mix セットで学習した場合の対数 F0 は，ATR セットで学習した時に比べ 1.3 cent 誤差が小さい結果となった．一方，dialogue セットを合成した場合は，mod セットと同様に mix セットで学習しても誤差が小さくなった．mod，mix を比較すると，メルケプストラム距離と音素継続長には大きな差はないものの，対数 F0 は mix セットで学習した方が誤差が 18.5 cent 小さくなった．

また，最終モーラの評価結果を表 6 に示す．文全体で評価した場合と同様に，read セットを合成した時は ATR セットで学習した方が誤差は小さくなったが，dialogue セットを合成した時は，mod セットと同様に mix セットで学習しても誤差が小さくなることが分かった．したがって，mix セットも mod セットと同様に対話文の特徴をよりよく表せることが分かった．

mod セットより mix セットを使用した時の方が，文全体で比較した場合も，最終モーラのみ比較した場合も，対数 F0 の誤差が小さくなった．これは，ラベルに朗読調の文であるか対話調の文であるかの情報を含めたことにより，

表 5　各学習セットについての客観評価結果
Table 5　Result of objective evaluation for each learn set.

評価セット	学習セット	mcep	lf0	dur
read	ATR	5.69	248.8	29.3
	mod	5.73	249.4	29.8
	mix	5.78	247.5	30.3
dialogue	ATR	6.04	323.4	40.4
	mod	5.99	283.3	36.4
	mix	6.00	264.8	36.8
read, dialogue の差の絶対値	ATR	0.34	74.6	11.1
	mod	0.26	33.9	6.6
	mix	0.22	17.3	6.5

表 6　各学習セットについての最終モーラの客観評価結果
Table 6　Result of objective evaluation of the last mora for each learn set.

評価セット	学習セット	mcep	lf0	dur
read	ATR	5.37	324.7	24.7
	mod	5.91	351.9	30.1
	mix	5.60	331.7	24.4
dialogue	ATR	6.39	556.9	71.2
	mod	6.06	375.7	45.0
	mix	6.11	348.1	43.7
read, dialogue の差の絶対値	ATR	1.02	232.2	46.5
	mod	0.15	23.7	14.9
	mix	0.51	16.4	19.3

合成したい文の口調を区別して音声を合成することが可能になったためだと考えられる．

2 つのセットについての距離の差の絶対値は mix セットの方が概ね小さかったが，最終モーラのメルケプストラム距離や音素継続長の差の絶対値は mod セットより大きかった．しかし，dialogue セットを合成した場合の文全体で評価した時のメルケプストラム距離，音素継続長はほぼ差異はない範囲だと考えられるため，これは read セットのメルケプストラム距離や音素継続長が mod セットより小さくなり，朗読調の文の再現性がよりよくなったことが原因だと考えられる．コーパスとしてのバランスは，コーパス中の対話文の割合を多くすることにより，read セットの評価は下がるものの，dialogue セットの評価が改善しバランスがよくなり，mod セットと比較した時の dialogue セットの評価も改善すると推測される．したがって，mod セットより mix セットの方が，朗読調の文も対話文も表現できるコーパスとしてよい構築法だと考えられる．音素バランス文と対話文の混合割合については，今後の検討課題である．

4. まとめ

本研究では，多様な対話音声合成を行うためのコーパスの構築方法を検討するため，コーパスを音素バランス文の文末を一部変更したものにすることや，コーパスに対話文を含めることの効果について客観評価を行った．その結果，客観評価実験においては，どのスタイルについても文末を対話調に変更することで，対話文の誤差が小さくなり，対話的な文末表現や疑問調が表せるようになることが確認できた．また，楽しげのスタイルについて，音素バランス文と対話文を混合する手法についても，対話文の誤差が小さくなり，同様の効果があることが分かった．朗読調の文と対話調の文の合成品質のバランスの面では，文末を変更する手法より音素バランス文と対話文を混合する手法の方がより優れていることが推測された．

今後の検討課題としては，音素バランス文と対話文を混合する手法について，楽しげ以外のスタイルについても評価をするとともに，それらの文の混合割合を検討する必要がある．また，主観評価実験を行い，聴感上の効果を確認する必要がある．

参考文献

[1] 大村祐司，川端　豪：雑談可能な目的達成型音声対話システム (オーガナイズドセッション)，電子情報通信学会技術研究報告. SP, 音声， Vol. 112, No. 369, pp. 47–51 (2012).
[2] 稲葉通将，平井尚樹，鳥海不二夫，石井健一郎：非タスク指向型対話エージェントのための統計的応答手法 (自然言語処理)，電子情報通信学会論文誌. D, 情報・システム，Vol. 95, No. 6, pp. 1390–1400 (2012).
[3] 大西可奈子，吉村　健：コンピュータとの自然な会話を実現する雑談対話技術，NTT DoCoMo テクニカル・ジャーナル， Vol. 21, No. 4, pp. 17–21 (2014).
[4] 日経産業新聞：対話できるロボ，高齢者交流促す，NEC，実験成果．, 2011/02/09, p.7.
[5] 加瀬嵩人，能勢　隆，伊藤彰則：対話シナリオに応じた感情音声合成を利用した音声対話システムの評価，日本音響学会春季講演論文集，pp. 201–202 (2015).
[6] Zen, H., Tokuda, K. and Black, A.: Statistical parametric speech synthesis, *Speech Communication*, Vol. 51, No. 11, pp. 1039–1064 (2009).
[7] 徳田恵一：HMM による音声合成の基礎，電子情報通信学会技術研究報告. SP, 音声，Vol. 100, No. 392, pp. 43–50 (2000).
[8] Yamagishi, J., Onishi, K., Masuko, T. and Kobayashi, T.: Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis, *IEICE Trans. Inf. & Syst.*, Vol. E88-D, No. 3, pp. 503–509 (2005).
[9] 荒生侑介，能勢　隆，篠崎隆宏，小林隆夫：複数ドメインコーパスからの文選択に基づくキャラクター音声合成の検討，日本音響学会秋季講演論文集，pp. 351–352 (2013).
[10] Stan, A., Yamagishi, J., King, S. and Aylett, M.: The Romanian Speech Synthesis (RSS) corpus: building a high quality HMM-based speech synthesis system using a high sampling rate, *Speech Communication*, Vol. 53, No. 3, pp. 442–450 (2011).
[11] 河原英紀，森勢将雅，高橋　徹，入野俊夫，坂野秀樹，藤村　靖：STRAIGHT スペクトルに基づく音源信号の抽出と非周期成分の評価について (一般)，電子情報通信学会技術研究報告. SP, 音声， Vol. 106, No. 333, pp. 43–48 (2006).
[12] The HTS working group: HMM-based Speech Synthesis System (HTS) (online),
入手先 ⟨http://hts.sp.nitech.ac.jp/⟩ (2015.04.18).