朗読音声において話速構成要素を独立に制御したときの主観的自然性の分析

竹下隼司　松崎拓也
東京理科大学大学院 理学研究科 応用数学専攻
takeshun1619@gmail.com　matuzaki@rs.tus.ac.jp

概要

朗読音声における話速は，調音速度（AR），ポーズ頻度（F），平均ポーズ長（d）といった複数の要素の組合せによって決定される。本研究では，目標話速に対して自然性が維持される話速構成要素（AR, F，d）の組合せの範囲を明らかにすることを目的として，日本語朗読音声コーパスの分析と，話速構成要素を独立に操作した音声刺激に対する主観評価を行った。その結果，調音速度が自然性に最も強く影響し，特定の調音速度範囲において高い自然性が得られることが示された。一方で，ポーズ頻度および平均ポーズ長は，極端な条件を除き，自然性に与える影響が比較的小さいことが確認された。

1　はじめに

オーディオブックやポッドキャスト，e ラーニング等の音声コンテンツにおいて，再生速度変更は一般的な機能として広く利用されている。近年では，倍速再生を積極的に活用する利用者の増加も報告されている。一方で，現在広く用いられている等比的な時間伸縮による話速変更では，抑揚や声質の歪み，破裂音の立ち上がりの不明瞭化，ポーズの潰れといった問題が生じ，聴取時の認知負荷増大につながる。特に朗読音声では，意味境界や呼吸とポーズが密接に関係しており，ポーズの位置や長さは自然性に大きく寄与する [1]。そのため，話速変更時にポーズ構造を適切に制御することが重要である。

こうしたポーズ構造の制御を体系的に扱うためには，話速を単一の量としてではなく，その構成要素に分解して捉える必要がある。話速は，調音速度（AR），ポーズ頻度（F），平均ポーズ長（d）といった複数要素の組合せによって決定される。ここで，AR は発話のモーラ数をポーズ区間を除く発話時間で割った値，F はポーズ区間数をモーラ数で割った値，d はポーズ区間の平均長と定義され，これらと話速 SR すなわち時間当たりのモーラ数には

$$SR = \frac{AR}{1 + AR \cdot F \cdot d} \tag{1}$$

という関係がある。要素 AR，F，d を個別に操作することで，単純な等比時間伸縮ではない話速制御が可能となる。

本研究では，目標話速に対して自然に聞こえる (AR，F，d) の組合せの領域を同定することを目的として，音声コーパスの分析と主観評価を行った。

2　関連研究

2.1　話速の要素に関する分析

話速を構成する要素に関する分析について，音声学および音声知覚の分野においてこれまでに多数の報告があり，特に調音速度 AR，ポーズ頻度 F，平均ポーズ長 d と話速との関係について，以下のような傾向が示されている。調音速度 AR については，話速の増加に伴い調音速度が増加することが報告されている [2, 3]。また，話速の増加に伴いポーズ頻度 F は低下する傾向が報告されており [2]，ポーズ頻度の低下により話速がより速いと知覚されることが示されている [4, 5]。平均ポーズ長 d についても，話速の増加に伴い短縮される傾向が報告されている [2, 3, 4]。以上の先行研究は，話速と各要素との相関関係や，それらが知覚される速さ感・自然さに与える影響について示唆している。一方で，目標話速に対して自然に聞こえる調音速度，ポーズ頻度，平均ポーズ長の組合せとしての許容領域については，体系的に明らかにされていない。

2.2　TTS における話速制御

テキスト音声合成（TTS）の際に話速を制御する手法が多数提案されている。その多くは，話速（SR）を単一の制御量として扱い，モデルに条件付

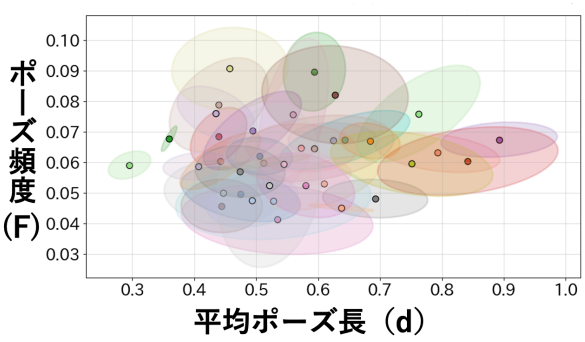

図 1　話者ごとの話速構成要素の分布

けるアプローチを採用している [6, 7, 8, 9]。これらの手法はいずれも話速制御を実現しているが，話速を構成する要素である調音速度，ポーズ頻度，平均ポーズ長の関係性を明示的に扱っていない。Abbas ら [9] は，ポーズ頻度を話速とは独立した制御量として入力しているが，話速と各要素との対応関係，すなわち SR から (AR，F，d) への写像を明示的にモデル化した TTS 研究は見当たらない。

3　朗読音声コーパスの分析

以上の先行研究を踏まえ，本研究では日本語多話者朗読音声コーパス J-MAC [10] を用い，話者ごとに話速 SR と調音速度 AR，ポーズ頻度 F，平均ポーズ長 d との関係を分析した。その結果，SR と AR の間には正の相関（$r = 0.77$）が，SR と F および d の間には負の相関（それぞれ $r = -0.62$，$r = -0.42$）が確認され，これまでの音声学的研究で報告されている傾向と整合的な結果が得られた。対応する図は付録 A の図 5 に示す。一方で，平均ポーズ長とポーズ頻度の関係を示す d–F 平面における分布は，話者（図 1）や作品（付録 B，図 6）ごとにばらつきが見られ，「自然に聞こえる話速構成要素の範囲」には個人差があることが示唆される。J-MAC に含まれる朗読音声はいずれもプロ話者によるものであり，これらの分布は自然な発話を実現するパラメータ領域（以下，これを「自然領域」と呼ぶ）の近似として解釈できる。

4　ポーズ予測モデル

朗読音声では，朗読者や作品，会話文か地の文か等の違いによって読み方が変化し [11]，ポーズの位置や長さも大きく異なる。また，適切なポーズ挿入は，自然な朗読音声合成において重要である [12]。このような多様性をもつ朗読音声に対して，ポーズ頻度や平均ポーズ長を制御した評価音声刺激を作成する際に，自然なポーズ位置および長さを人手で一つ一つ設定することは困難である。

そのため，文内容および文脈を考慮してポーズ位置および長さを自動的に推定する予測モデルを用いる。具体的には，形態素列を入力として，各形態素後のポーズの有無（分類）およびポーズ長（回帰）を予測する [11]。モデルは BERT [13] と BiLSTM を組み合わせた構成であり，文脈情報を考慮した系列予測を行う。学習データには，J-MAC に含まれる全 74 作品の朗読音声を用いた。

5　制御の方法

FastSpeech2 [14] の Duration Predictor により予測された音素ごとの持続時間および，前節で説明したモデルによるポーズ位置・ポーズ長の予測結果を用いて，話速構成要素である調音速度 AR，ポーズ頻度 F，平均ポーズ長 d を明示的に操作する。

まず，形態素列に対するポーズ位置およびポーズ長をモデルを用いて予測する。ポーズ位置予測に関しては，予測器が出力する確率 $P($ 位置 i にポーズが存在する $)$ に対する閾値を調整することでポーズ頻度を制御し，予測された形態素間のポーズ位置にポーズトークンを挿入する。挿入後の形態素列を Open JTalk [15, 16] により音素列へ変換し，FastSpeech2 の Duration Predictor により各音素の持続時間を予測した後，文章全体の調音速度が目標値に一致するよう一律に伸縮する。一方，ポーズ長については，文章全体の平均ポーズ長が目標値と一致するよう，予測モデルにより推定されたポーズ長を全体的に伸縮する。これらの操作により，調音速度および平均ポーズ長を独立に制御した持続時間列を得る。最後に，修正後の持続時間列を FastSpeech2 に入力し，JSUT コーパス [17] で学習したモデルを用いて音声を合成する。制御手法により，文章ごとに自然なポーズ位置およびポーズ長の特徴を保持しつつ，持続時間を介した話速構成要素の明示的な制御を実現する。

評価に用いる音声刺激は，J-MAC に含まれる 5 つの朗読作品それぞれについて，連続する 2 文からなる文章を 2 つずつ，計 10 種類抽出した。抽出にあたっては，ポーズ位置予測モデルの閾値を段階的に変更した際に，ポーズ頻度が各段階で変化する文章を対象としたうえで，各文章におけるポーズ頻度の変化が，当該朗読作品全体における平均的なポーズ頻度の変化と大きく乖離しないものを選定した。

文間のポーズ長については，各朗読作品における文間ポーズ長の平均値に対し，文中ポーズ制御で用いた「予測モデルによるポーズ長から制御後のポーズ長への伸縮率」を適用したものを用いた。これにより，文中および文間のポーズ長制御が整合的に行われるようにした。最終的に，文間ポーズ長を持つ無音区間を挿入した上で，各文に対し個別に合成した音声を結合し，評価用音声刺激を作成した。

6 評価データの作成

3 節のコーパス分析結果をもとに，予測される自然領域の内外を含むように話速構成要素（調音速度 AR，ポーズ頻度 F，平均ポーズ長 d）の制御点を設定した。具体的には，J-MAC における (AR，F，d) の三次元分布を参照し，分布内部には格子点状に制御点を配置するとともに，分布境界および分布外側にも制御点を配置することで，自然性が維持される条件から逸脱する領域を含めた横断的な評価が可能となる制御点集合を構成した。その結果，合計 60 点の制御点を設定した。

各制御点に対して，朗読文 10 文章それぞれについて音声刺激を合成した。評価対象となる比較対は，制御を行わない音声に対する比較（制御点 vs ベースライン）を中心に構成するとともに，自然領域内外の相対関係を補足的に捉える目的で，一部に制御点同士の比較を含めた。最終的に，評価対象となる比較対は 200 組（制御点 vs ベースライン：60 組，制御点 vs 制御点：140 組）とした。

主観評価は，自然さに基づく一対比較（AB テスト）として実施した。クラウドソーシングを用い，各比較対について 100 件の回答を収集し，総回答数は 20,000 件（1,000 人 × 20 回答）となった。回答者には，同一朗読文を異なる制御点条件で合成した 2 音声を提示し，より自然に聞こえる方または「同程度である」を選択させた。回答負荷と信頼性を考慮して，1 回のタスクあたりの回答数を 21 比較対とし，うち 1 対は注意確認用のチェック問題とした。

7 分析方法

一対比較によって得られた評価結果に対して Bradley–Terry（BT）モデル [18] を適用し，各制御点の潜在的な自然性スコアを推定した。BT モデルは，比較対の間の勝敗に基づいて項目間の相対的な尺度を推定する確率モデルであり，本研究では制御点間の自然性の相対関係を定量化する目的で用いる。

各比較対 A/B について，「同程度である」の回答数の半分を A, B それぞれの勝利数に算入し，全 200 比較対（制御点 vs ベースラインおよび制御点 vs 制御点）を対象として BT モデルを適用した。推定されたスコアは，ベースラインのスコアを 0 とした相対的な log スコアとして示す。

また，自然領域の境界を同定するために，ベースライン音声に対する勝率に基づき，明らかに不自然であるような制御点のスコアの範囲を調べた。具体的には，各制御点についてベースラインとの比較における勝率（勝利数／比較回数）を算出する。これらと自然性スコアの関係を調べ，勝率が 0.25 以上となる制御点のうち，自然性スコアが最小となる制御点のスコアを，境界として採用する。

なお，分析を行うにあたり，ポーズ位置予測モデルにおける各閾値設定ごとに 10 種類の音声刺激でのポーズ頻度の実測値の平均値を算出し，その値をポーズ頻度 F として用いる。設計段階での課題として，ポーズ頻度に関して自然領域の境界近傍に対応する制御点を十分に設定できていない。そのため，境界付近におけるポーズ頻度の影響については十分に評価できておらず，今後の課題である。

8 分析結果

8.1 自然領域の境界の決定

図 3 に，各制御点の潜在的自然性スコアと，ベースラインに対する勝率の関係を示す。各点は制御点を表しており，縦方向のバーは，「同程度である」とされた回答をすべて非勝利とみなした場合から，すべて勝利とみなした場合までの勝率の範囲を表し，その中央の点は BT モデルの推定に用いた勝利数に対応する。

7 節で述べた基準を図 3 のデータに適用することで，自然性スコア −0.46 を自然領域の境界と判断した。また，図より，スコアがこれを超える制御点は，ベースラインと「同程度である」という回答が多い傾向が分かる。

8.2 潜在的自然性スコアの分布

図 4 に，(AR，F，d) の三次元空間における各制御点の潜在的自然性スコア分布，図 2 にその調音速度ごとの断面図を示す。色は自然性スコアを，大きさはベースラインからのスコア差を表している。

本実験で設定した制御点の範囲においては，

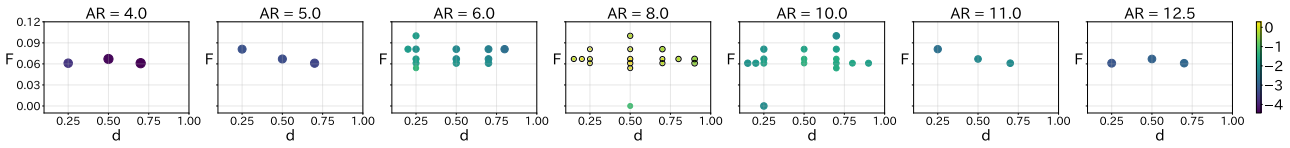

図 2　調音速度ごとの自然性スコア

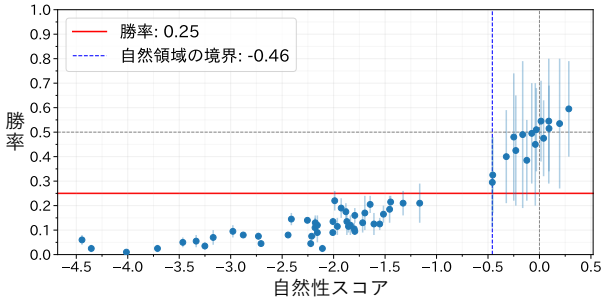

図 3　各制御点の自然性スコアと勝率

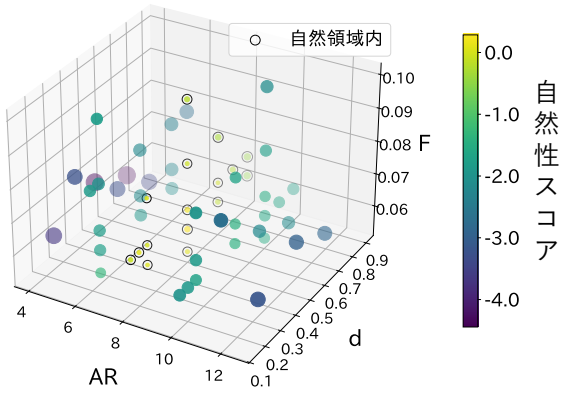

図 4　(AR，F，d) の三次元空間上における各制御点の潜在的自然性スコア分布

AR=8.0 の制御点が最も高い自然性スコアを示し，ポーズ頻度 F が 0.4 以上や 0.0 といった極端な制御点を除くと，AR=8.0 の全ての制御点が自然領域内と判定された。一方で，AR がこの値から離れるにつれて自然性スコアが低下する傾向が確認され，AR=8.0 以外の制御点は自然領域外と判定された。

このことから，本実験条件下においては，特定の調音速度条件（AR=8.0）が自然性に強く寄与していることが示唆される。この傾向は，3 節のコーパス分析において観測された調音速度分布と整合的である。J-MAC に含まれる朗読音声では，AR はおおよそ 7〜9 モーラ/秒の範囲に集中しており（付録 A，図 5），主観評価で用いた条件のうち AR=6.0 および AR=10.0 は，その最小値・最大値のやや外側に位置する。ただし，本研究で設定した AR の刻みは粗く，自然領域の境界を同定するために重要となる AR=7〜9 の範囲については詳細に評価できていない。

8.3　話速構成要素ごとの影響

話速構成要素別に見ると，調音速度 AR が自然性スコアに最も強く影響していることが明らかとなった。8.2 節で述べた AR ≈ 8 モーラ/秒の付近ではベースラインを上回るスコアの制御点も確認された。

一方，平均ポーズ長 d およびポーズ頻度 F については，外側の極端な値においてのみ自然性スコアが僅かに低下する傾向が見られ，比較的広い範囲にわたって自然性スコアの変化にほとんど影響を与えないことが分かった。このことから，朗読音声においてポーズ要素は，自然性を大きく左右する主要因ではなく，一定範囲内では自然性に対する感度が低い構成要素であることが示唆される。

F および d が一定範囲で自然性に与える影響が小さいという結果は，話者ごとの読みの個性や意味的強調を，自然性を保ったまま付与できる余地があることを示唆し，これは J-MAC における話者ごとの話速構成要素の分布（図 1）とも整合的である。

9　おわりに

本研究では，朗読音声合成におけるより自然な話速制御に向け，話速構成要素である調音速度，ポーズ頻度，平均ポーズ長の組合せと主観的自然性との関係を分析した。その結果，本研究で用いた合成音声条件下では，調音速度が自然性スコアに最も大きな影響を与える要因である一方，ポーズ頻度と平均ポーズ長の影響は限定的であることが示された。

これらの結果は，話速制御において話速 SR のみに基づき音声を等比的に操作するのではなく，話速を構成要素（AR，F，d）に分解した上で組合せとして制御することが，自然性を維持した話速調整につながることを示唆している。

今後の課題は，制御点を細分化し，自然領域の境界をより精緻に同定することである。

参考文献

[1] 杉藤美代子, 大山玄. 朗読におけるポーズと呼吸—息継ぎのあるポーズと息継ぎのないポーズ—. 日本人の声, pp. 87–103. 和泉書院, 1994.

[2] 杉藤美代子. ポーズの時間および発話時間と意味との関連—語り聞かせ「オオカミの大しくじり」の分析—. 日本人の声, pp. 27–42. 和泉書院, 1994.

[3] 杉藤美代子. 効果的な朗読へのアプローチ —「天気予報」と「大きなかぶ」—. 日本人の声, pp. 43–60. 和泉書院, 1994.

[4] 広実義人. 知覚上の発話速度に及ぼすポーズ数の影響. 日本音声学会, No. 205, 1994.

[5] 鈴木淳也, 佐川雄二, 田中敏光, 杉江昇, 下山博. 聞きやすい音声におけるポーズ長と話速の関係の分析. 名城大学総合研究所総合学術研究論文集, Vol. 4, , 2005.

[6] Jesuraja Bandekar, Sathvik Udupa, Abhayjeet Singh, Anjali Jayakumar, G Deekshitha, Sandhya Badiger, Saurabh Kumar, VH Pooja, and Prasanta Kumar Ghosh. Speaking rate attention-based duration prediction for speed control tts. 2023.

[7] Jae-Sung Bae, Hanbin Bae, Young-Sun Joo, Junmo Lee, Gyeong-Hoon Lee, and Hoon Young Cho. Speaking speed control of end-to-end speech synthesis using sentence-level conditioning. In **Proc. Interspeech**, 2020.

[8] M. Lenglet, O. Perrotin, and G. Bailly. Speaking rate control of end-to-end tts models by direct manipulation of the encoder's output embeddings. In **Proc. Interspeech**, 2022.

[9] Syed Abbas, Thomas Merritt, Alexis Moinet, Sri Karlapati, Ewa Muszynska, Simon Slangen, Elia Gatti, and Thomas Drugman. Expressive, variable, and controllable duration modelling in tts. In **Proc. Interspeech**, 2022.

[10] Shinnosuke Takamichi, Wataru Nakata, Naoko Tanji, and Hiroshi Saruwatari. J-MAC: japanese multi-speaker audiobook corpus for speech synthesis. In **Proc. Interspeech**, 2022.

[11] 竹下隼司, 松崎拓也. 朗読音声合成におけるポーズ長分布の多様性を吸収するための標準化の効果. 人工知能学会全国大会論文集, 2024.

[12] Alok Parlikar and Alan W. Black. Modeling pause duration for style-specific speech synthesis. In **Proc. Interspeech**, pp. 446–449, 2012.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In **North American Chapter of the Association for Computational Linguistics**, 2019.

[14] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. In **International Conference on Learning Representations**, 2021.

[15] Open jtalk. http://open-jtalk.sourceforge.net/.

[16] Ryuichi Yamamoto. pyopenjtalk, 2018. https://github.com/r9y9/pyopenjtalk.

[17] Ryosuke Sonobe, Shinnosuke Takamichi, and Hiroshi Saruwatari. Jsut corpus: free large-scale japanese speech corpus for end-to-end speech synthesis. 10 2017.

[18] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. **Biometrika**, Vol. 39, No. 3/4, pp. 324–345, 1952.

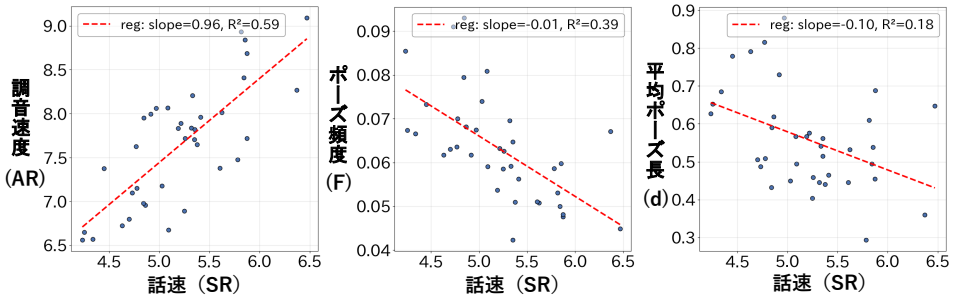

図 5　話速と調音速度，ポーズ頻度，平均ポーズ長の関係

A　話速と話速構成要素との関係

3 節で述べた相関傾向の確認のため，図 5 に話者ごとの話速と話速構成要素との散布図を示す。

B　文章作品ごとの話速構成要素の分布

図 6 に文章作品ごとの平均ポーズ長 d とポーズ頻度 F の分布を示す。

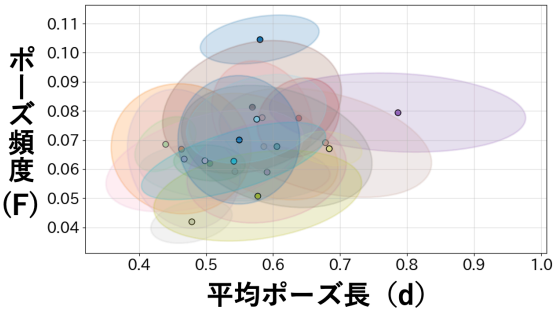

図 6　文章作品ごとの話速構成要素の分布

C　制御値と実測値との差異

本研究では，話速構成要素の制御値をもとに可視化・分析を行い，各制御点の潜在的な自然性スコアを推定するとともに，調音速度・ポーズ頻度・平均ポーズ長の組合せと潜在的な自然性スコアとの関係を分析した。本節では，その判断の背景として，制御値と実測値との間に生じる差異とその影響について整理する。

聴取者が実際に聴取する合成音声上での話速構成要素は，合成音声に対して強制アライメントを適用することで実測値として計測することができる。しかし，持続時間推定以降の音声合成処理の影響により，指定した制御値と合成音声における話速構成要素の実測値との間には，制御点および文章ごとに若干の差異が生じる。

この差異により，実測値の組を評価対象と見なす場合，評価対象となる点がばらけてしまい，Bradley-Terry モデルのパラメータ数当たりの標本数が減少することで，推定結果の信頼性が低下する可能性がある。そのため，本研究では制御の際の指定値に基づく評価結果を用いて潜在的自然性スコアの推定を行った。

ただし，制御値と実測値との乖離の程度の把握は，推定結果の解釈や話速制御モデルへの統合を考える上で重要である。図 7 に，各制御点における話速構成要素の制御値と実測値との差異を示す。

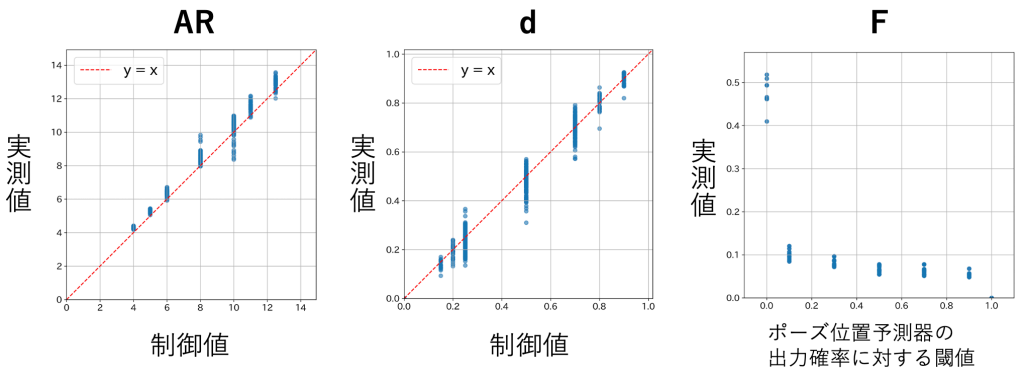

図 7　各制御点における話速構成要素の制御値と実測値の差異