JVS：フリーの日本語多数話者音声コーパス

高道 慎之介[1,a)]　三井 健太郎[1]　齋藤 佑樹[1]　郡山 知樹[1]　丹治 尚子[1]　猿渡 洋[1]

概要：深層学習を含む機械学習技術の発展により，音声合成は機械学習タスクになっている．音声合成研究の加速のために我々は，学術機関だけでなく民間企業からも容易にアクセス可能な日本語音声コーパスを開発しており，その一環として 2017 年に JSUT コーパスを公開した．このコーパスは，end-to-end 型音声合成に向けて，単一話者による 10 時間の読み上げ音声データを含むよう設計された．音声変換や多数話者モデリングなどの一層多様な音声合成研究の加速を見据え，本研究では 3 スタイル（通常，ささやき，および裏声）・100 話者の音声データを含む JVS コーパスを構築した結果を報告する．このコーパスは，22 時間のパラレル読み上げ音声を含む 30 時間の音声データが含む．本コーパスは，プロジェクトページにて入手できる．

Shinnosuke Takamichi[1,a)]　Kentaro Mitsui[1]　Yuki Saito[1]　Tomoki Koriyama[1]　Naoko Tanji[1]
Hiroshi Saruwatari[1]

1. はじめに

深層学習に恩恵を受け，音声合成関連技術（テキスト音声合成，歌声合成，音声変換，音声符号化など）は，これまでにない急速な発展を見せており [1], [2], [3], [4]，機械学習タスクのひとつとなっている．容易にアクセス可能な音声コーパスは，そのような研究を加速させるのみならず研究の再現性を向上させるために有効である．2017 年に我々は，end-to-end テキスト音声合成のための大規模日本語音声コーパス JSUT を公開した [5]．このコーパスは，単一話者による 10 時間の読み上げ音声と常用漢字 [6] の全ての読みを含むよう設計された．プロジェクトページ [7] は，2017 年 10 月の公開から 60 カ国以上・6,000 回以上のアクセス（75%は国内，25%は国外）を受けており，JSUT コーパスは，今日の日本語音声合成研究における主要な音声コーパスとなっている [8], [9].

より多様なの音声研究を加速させるために，本稿では新たな音声コーパス JVS (Japanese versatile speech) を構築する．このコーパスは以下の利点を持つ．

高品質フォーマット: 音声ファイルは，24 kHz サンプリング，16 bit 量子化，RIFF WAV 形式で保存されている．

高音質録音: 音声ファイルは，プロ音響監督による監督下で収録室にて収録されたものである．

多数話者: 本コーパスは声優・俳優などのプロ話者 100 名の日本語話者を含む．

多数スタイル: 各話者は，読み上げ音声のみならず，ささやき声，裏声を発話している．

大規模: 本コーパスは合計で 30 時間の音声データからなる．

パラレル／ノンパラレル発話: 各話者は，話者間で共通する文と話者間で異なる文を発話している．

多数のタグ: 本コーパスは，音声データのみならず，読み上げテキスト・性別情報・F_0 レンジ・話者間類似度・音素アライメントを含む．

研究用途では無償: 本コーパスは，研究用途であれば学術機関だけでなく民間企業においても利用可能である．

容易にアクセス可能: 本コーパスはプロジェクトページ [10] にて入手できる．

以降では，本コーパスの設計について述べる．

2. コーパスデザイン

本コーパスは，4 つのサブコーパスからなる．サブコーパス名称は *[NAME][NUM_UTT]* のフォーマットで統一されている．*[NUM_UTT]* は各話者の発話数である．

parallel100: パラレル読み上げ音声 100 発話

nonpara30: ノンパラレル読み上げ音声 30 発話

whisper10: ささやき声 10 発話

[1] 東京大学 大学院情報理工学系システム情報学専攻, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.
[a)] shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp

falsetto10: 裏声 10 発話

コーパスのディレクトリ構造を以下に示す．話者名称は jvs*[SPKR_ID]* のフォーマットで統一されている．*[SPKR_ID]* は，1 から 100 の値を取る話者インデックスである．

```
jvs001
  parallel100
    wav24kHz16bit
    lab
    transcripts_utf8.txt
  nonpara30
    wav24kHz16bit
    lab
    transcripts_utf8.txt
  whisper10
    wav24kHz16bit
    transcripts_utf8.txt
  falsetto10
    wav24kHz16bit
    transcripts_utf8.txt
jvs002
...
jvs100
speaker_similarity_male.csv
speaker_similarity_female.csv
duration.txt
gender_f0range.txt
```

2.1 サブコーパス

本節では，4 つのサブコーパスについて概説する．

2.1.1 parallel100

話者間で対応する発話（パラレル発話）は，声質変換 [11], [12] や話者モデリング [8], [13] などに使用される． 我々は，JSUT コーパスのサブコーパス “voiceactress100” から音素バランス 100 文を使用し，話者に発話させた．本サブコーパスは，その音声データのみならず読み上げテキスト (“parallel100/transcript_utf8.txt”) と音素アライメント (“parallel100/lab”) を含む．

2.1.2 nonpara30

話者間で異なる発話（ノンパラレル発話）の使用は，パラレル発話の使用に比べてより現実的な設定である．我々は，JSUT コーパスの “voiceactress100” 以外のサブコーパスからランダムに文を選択し，話者毎に異なる 30 文を発話させた．本サブコーパスは，“parallel100” と同様，読み上げテキストと音素アライメントを含む．ただし，本コーパスの文は，“parallel100” と異なり，音素バランス文でないことに注意する．

2.1.3 whisper10

ささやき声は，第三者への情報漏えいを回避するサイレントコミュニケーションの手段であり，これまでにもその分析 [14]・合成 [15]・認識 [16]・変換 [17] 技術が研究されている．本サブコーパスの最初の 5 文は，“parallel100” の最初の 5 文と同じである．残る 5 文は，“nonpara30” の最初の 5 文と同じである．すなわち，各話者の 10 発話は，ささやき声と読み上げ声間でパラレルである．

2.1.4 Falsetto10

裏声は，読み上げ音声の F_0 レンジを超えた発話であり，その発声原理は，読み上げ音声の原理と異なることが知られている [18]．本サブコーパスの最初の 5 文は，“parallel100” の最初の 5 文と同じである．残る 5 文は，“nonpara30” の 5 文と同じだが “whisper10” の 5 文と異なる．すなわち，各話者の 5 発話は，裏声と読み上げ音声でパラレルであり，5 発話は，ささやき声と裏声でパラレルである．

2.2 タグ

本節では，コーパスに含まれるアノテーションを述べる．

- **F_0 レンジ (gender_f0range.txt)**: 典型的な F_0 抽出器 [19], [20], [21] では F_0 探索のレンジを設定できる．この設定は，分析後の結果にしばしば強い影響をもたらす．本コーパスは，各話者の読み上げ音声に対して手動で付与された F_0 レンジを含む．
- **話者間類似度 (speaker_similarity_*.csv)**: 話者間の知覚的類似度は，話者（もしくは話者依存モデル）選択 [22] や話者空間構築 [23] に使用される．本コーパスは，各性別の全話者組み合わせに対する主観的類似度スコアを含む．
- **継続長 (duration.txt)**: 本コーパスは，継続長（データサイズ）と話速を含む．音素レベルの継続長は，音素アライメントの結果から計算できる．

3. 音声収集の結果

3.1 スペック

コーパス構築にあたり，我々は日本語プロ話者 100 名（男性 49 名・女性 51 名）を雇用した．全話者の収録は，プロ音響監督の監督下で収録室にて実施した．各話者の収録は 1 日以内に実施した．音声は，48 kHz サンプリングで収録された後，SPTK [24] を用いて 24 kHz にダウンサンプリングした．音声ファイルのフォーマットには，16 bit RIFF WAV 形式を使用した．読み上げ文は UTF-8 形式で保存した．各文の読点を，読み上げ音声の呼気段落間に対応する位置に付与した． 文のフルコンテキストと音素ラベルは，Open JTalk [25] を用いて自動生成した．音素アライメントは，Julius [26] を用いて自動生成した．F_0 レンジは，戸田による資料 [27] を参考に手動で付与した．F_0 抽出には WORLD [20], [28] を使用した．話者間類似度は，

表 1　話者毎の継続長の統計量．この計算には非音声区間も含まれることに注意する．

	Mininum [min.]	Average [min.]	Maximum [min.]	Total (100 speaker) [hour]
parallel100 (100 utterances)	10.11 (jvs020)	13.11	18.24 (jvs084)	22
nonpara30 (30 utterances)	2.12 (jvs099)	2.62	3.86 (jvs036)	4.4
whisper10 (10 utterances)	0.95 (jvs045)	1.24	1.69 (jvs018)	2.0
falsetto10 (10 utterances)	0.90 (jvs045)	1.18	1.61 (jvs035)	2.0
Total	-	-	-	30.4

齋藤らの研究 [23] を参考に，クラウドソーシングサービス Lancers [29] を使用して評価した．各評価者には，各話者対の類似度スコアを −3 (全く異なる) から +3 (非常に似ている) の 7 段階で付与させた．最終的なスコアは，話者対毎に全評価者のスコアを平均したものとした．各話者対のスコアは，10 名の異なる評価者により付与した．最終的な評価者数は 1,000 人だった．

3.2　分析

3.2.1　継続長

話者毎の継続長の統計量を Table 1 に示す．本コーパスは，合計で 26 時間の読み上げ音声，4 時間のささやき声／裏声を含み，各話者について，平均 15.7 分の読み上げ音声，1.24 分のささやき声，1.18 分を含む．サブコーパス “parallel100” では話者間の読み上げ文は共通するが，その継続長は話者間で非常に異なることがわかる．例えば，話者 “jvs084” は話者 “jvs020” に比べ 1.8 倍遅く発話している．

3.2.2　主観的話者間類似度

主観的話者間類似度の行列を Fig. 1 に示す．例えば，最も主観的に類似した話者対は，話者 “jvs019” と “jvs096” である．また，他話者から知覚的に最も異なる話者は，話者 “jvs010” である．

4.　まとめ

本稿では，多様な音声研究に向けて，JVS コーパスの設計手順と仕様を述べた．本コーパスのテキストデータのライセンスは，JSUT コーパスの LICENCE ファイルに記載されている．本コーパスのタグは，CC BY-SA 4.0 によってライセンスされている．音声データは，以下の場合に限り使用可能である．

- アカデミック機関での研究
- 非商用目的の研究（営利団体での研究も含む）
- 個人での利用（ブログなどを含む）

なお，プロジェクトページ [10] では商用利用に関する情報を提供している．

5.　謝辞

本研究の一部は，セコム科学技術振興財団の支援を受けて実施した．また，本研究開発は総務省 SCOPE(受付番号 182103104) の委託を受けた．

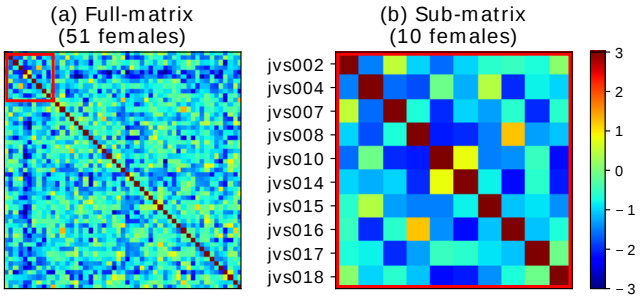

図 1　Speaker similarity matrix of 51 Japanese females and (b) its sub-matrix obtained by large-scale subjective scoring.

参考文献

[1] G. Hinton, L. Deng, D. Yu, G. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *Signal Processing Magazine of IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[2] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, “WaveNet: A generative model for raw audio,” vol. 1609.03499, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499

[3] S. Takamichi, K. Tomoki, and H. Saruwatari, “Sampling-based speech parameter generation using moment-matching network,” in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017.

[4] Y. Saito, S. Takamichi, and H. Saruwatari, “Statistical parametric speech synthesis incorporating generative adversarial networks,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 84–96, Jan. 2018.

[5] R. Sonobe, S. Takamichi, and H. Saruwatari, “JSUT corpus: free large-scale japanese speech corpus for end-to-end speech synthesis,” vol. arXiv 1711.00354, 2017.

[6] G. o. J. Agency for Cultural Affairs, “List of daily-use kanjis http://www.bunka.go.jp/kokugo_nihongo/sisaku/joho/joho/kijun/naikaku/kanji/index.html,” 2010.

[7] “JSUT: Japanese speech corpus of Saruwatari Lab, the University of Tokyo corpus,” https://sites.google.com/site/shinnosuketakamichi/publication/jsut.

[8] S. Ueno, M. Mimura, S. Sakai, and T. Kawahara, “Multi-speaker sequence-to-sequence speech synthesis for data augmentation in acoustic-to-word speech recognition,” in *Proc. ICASSP*, Brighton, United Kingdom, May 2019, pp. 6161–6165.

[9] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki, “Emotional voice conversion using dual supervised adversarial net-

works with continuous wavelet transform F0 features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1535–1548, Oct. 2019.

[10] "JVS: Japanese Versatile Speech corpus," https://sites.google.com/site/shinnosuketakamichi/research-topics/jvs_corpus.

[11] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[12] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[13] H. Lu and S. King, "Factorized context modeling for Text-to-Speech synthesis," in *Proc. ICASSP*, Vancouver, Canada, May 2013.

[14] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Communication*, vol. 45, no. 2, pp. 139–152, 2005.

[15] V. A. Petrushin, L. I. Tsirulnik, and V. Makarova, "Whispered speech prosody modeling for TTS synthesis," Chicago, U.S.A., May 2010.

[16] S.-C. Jou, T. Schultz, and A. Waibel, "Whispery speech recognition using adapted articulatory features," in *Proc. ICASSP*, vol. 1, Philadelphia, U.S.A., Mar. 2005, pp. 1009–1012.

[17] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2505–2517, Nov. 2012.

[18] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *The Journal of the Acoustical Society of America*, vol. 90, no. 5, pp. 2394–2410, Nov. 1991.

[19] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveigne, "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. 3–4, pp. 187–207, 1999.

[20] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE transactions on information and systems*, vol. E99-D, no. 7, pp. 1877–1884, 2016.

[21] D. Talkin, "REAPER: Robust Epoch And Pitch EstimatoR," https://github.com/google/REAPER.

[22] P. Lanchantin, M. J. Gales, S. King, and J. Yamagishi, "Multiple-average-voice-based speech synthesis," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 285–289.

[23] Y. Saito, S. Takamichi, and H. Saruwatari, "DNN-based speaker embedding using subjective inter-speaker similarity for multi-speaker modeling in speech synthesis," in *Proc. SSW10*, Vienna, Austria, Sep. 2019.

[24] "Speech signal processing toolkit (SPTK)," http://sp-tk.sourceforge.net/.

[25] "Open jtalk," http://open-jtalk.sourceforge.net/.

[26] A. Lee, T. Kawahara, and K. Shikano, "Julius — an open source real-time large vocabulary recognition engine," in *Proc. EUROSPEECH*, Aalborg, Denmark, Sep. 2001, pp. 1691–1694.

[27] T. Toda, "Hands on voice conversion," https://www.slideshare.net/NU_I_TODALAB/hands-on-voice-conversion, 2018.

[28] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, 2016.

[29] "Lancers," http://www.lancers.jp.