睡眠を誘発する音声刺激の生成に向けた分析と評価

熊田 順一[1]　齋藤 佑樹[1,a)]　高道 慎之介[1]　渡邊 亞椰[1]　丹治 尚子[1]　長野 瑞生[2]　井島 勇祐[2]
猿渡 洋[1]

概要：睡眠障害は世界的に社会課題の一つと捉えられている．本研究では，睡眠を誘発する音声刺激の特徴の分析と生成について検討する．分析対象として女性プロ話者 1 名による約 3 時間の睡眠導入朗読動画と約 4 時間の雑談動画を YouTube から取得し，その音声特徴量の統計量を算出することにより，聴取者の睡眠誘発を意識した発話とそうでない発話に表出する発話スタイルの違いを調査する．主観評価の結果から，F0 とエネルギーの平均と標準偏差が小さい音声が睡眠を誘発する印象を与える傾向にあることを示す．

Junichi Kumada[1]　Yuki Saito[1,a)]　Shinnosuke Takamichi[1]　Aya Watanabe[1]　Naoko Tanji[1]
Mizuki Nagano[2]　Yusuke Ijima[2]　Hiroshi Saruwatari[1]

1. はじめに

睡眠障害は世界的に社会問題の一つとして捉えられており，近年の Covid-19 流行に伴う行動制限やリモート通学・通勤の推進に伴い，その深刻さも増す傾向にある [1]．睡眠障害の緩和のための一般的な臨床介入法は認知行動療法の導入と睡眠導入剤の使用であるが，これらは高価もしくは有害な副作用を引き起こすおそれがある [2]．故に，非侵襲的かつ低コストな睡眠障害への介入策が要求されている．

本研究では，利用者が手軽に聴取できる音声コンテンツ生成への応用を視野に入れ，睡眠導入を目的として発話された音声刺激の分析を実施する．分析対象として女性プロ話者 1 名による約 3 時間の睡眠導入朗読動画と約 4 時間の雑談動画を YouTube から取得して音声を抽出し，その音響的特徴の統計量を算出することにより，聴取者の睡眠誘発を意識した発話とそうでない発話に表出する発話スタイルの違いを調査する．本稿では，音声刺激の収集・前処理・分析の方法と，主観評価デザインおよびその結果について報告する．主観評価の結果から，F0 とエネルギーの平均と標準偏差が小さい音声が睡眠を誘発する印象を与える傾向にあることを示す．

[1] 東京大学
[2] NTT
[a)] yuuki_saito@ipc.i.u-tokyo.ac.jp

2. 関連研究

2.1 睡眠障害緩和のための音楽刺激の分析と生成

音楽療法は音楽の聴取や演奏に伴う生理的・心理的・社会的な効果により，心身の健康回復および向上を目的とする補完医療である [3]．Dickson らは音楽療法に適した音刺激を分析し，スペクトル重心が低く，音の変動が滑らかで，リズムに対する強調が少ない楽曲が適していることを示唆した [4]．Yang らは CycleGAN [5] に基づくスタイル転写学習により，睡眠障害緩和に適した楽曲が有する特徴をそうではない（ユーザが普段聴取するような）楽曲に反映させる機械学習モデル (SleepGAN) を提案した [6]．これらの研究に代表されるように，音楽刺激を対象とした睡眠誘発特徴の分析および生成に関する研究は近年進められているが，人間の声を対象としたものは発展途上といえる．

2.2 In-the-wild なコンテンツ由来の音声コーパス構築

多様な話者を含む高品質な大規模コーパスは整備されつつある [7], [8], [9] 一方で，これらが網羅可能な発話スタイルはいまだ限定的であり，より多様な音声表現を含む in-the-wild なコンテンツから音声コーパスを構築する試みが推進されている．Yang らは Bilibili 動画[*1]に投稿されたコンテンツを対象とし，音声に含まれるユーモア表現を自動で予測する機械学習モデルのためのコーパスである Bilibili コーパスを構築した [10]．Chen らはインター

[*1] https://www.bilibili.com/

ネット上のオーディオブックおよび，ポッドキャストやYouTube に投稿されたコンテンツから 10,000 時間の超大規模音声コーパスを構築する方法論を提案し，その成果物 (GigaSpeech) を用いた音声認識タスクでの性能を評価した [11]．Takamichi らは同様の方法論により，日本語における最大規模の多話者オーディオブック音声コーパスである J-MAC [12] と音声認識・話者照合タスクに向けた JTubeSpeech [13] を構築した．これらの in-the-wild データに基づく音声コーパスは収録環境や背景音と音響効果による修飾などを多く含み，必ずしもテキスト音声合成 [14] や声質変換 [15] などの生成系タスクに適しているという保証はないが，Seki らはモデル学習とデータ評価のループに基づくデータ選択法により，音声を高品質に合成可能な話者を増加できることを示した [16].

3. 睡眠誘発を目的として発話された音声刺激の収集と分析

2 節で概説した関連研究の進展を踏まえ，本研究では YouTube に投稿された睡眠誘発音声コンテンツから音声刺激を収集し，その特徴を分析する．なお，音声収集手順は JTubeSpeech のスクリプト[*2]の実行に基づいている．

3.1 動画検索フレーズの定義と分析対象話者の選定

JTubeSpeech では Wikipedia の記事と Google Trends に基づき動画検索フレーズを自動取得するが，本研究では対象ドメインを睡眠誘発音声に限定し，かつ，多くの動画製作者からのコンテンツの投稿が見込まれる朗読，絵本読み聞かせ，個人ラジオを候補とし，動画検索フレーズを以下のように定義した．

- 【睡眠導入】,【眠くなる声】
- 【睡眠朗読】,【睡眠用朗読】,【睡眠導入朗読】,【睡眠前の朗読】,【睡れる朗読】,【寝かしつけ朗読】,【おやすみ前の朗読】,【おやすみ前に朗読】
- 【眠れる絵本読み聞かせ】,【睡眠用読み聞かせ】,【寝かしつけ絵本】
- 【寝落ちラジオ】

これらのフレーズを用いた動画検索結果を精査し，(1) 自動字幕付き，(2) 背景音が最小限，(3) 睡眠導入以外の発話スタイルのデータを含む，日本語を母語とする 40 代の女性プロ話者 1 名[*3]を分析対象話者として選定した．

3.2 音声セグメンテーション

本研究では，3.1 で選定した話者による「あのときの王子くん」の睡眠誘発朗読音声 (3.08h) を分析対象音声とした．音声のセグメンテーションは (1) Whisper[*4]により付与された自動字幕に基づく分割と (2) inaSpeechSegmenter [17] に基づく分割の 2 段階で実施した．前者が朗読音声全体を粗いセグメントに分割する作業であり，後者が音響特徴量由来の音声区間ラベリング（音声/無音/ノイズ/音楽）に基づく精緻な分割を行う作業に対応する．

睡眠誘発音声の対照データとして，当該話者による雑談配信 YouTube Live のアーカイブ動画 3 件 (合計 4.01h) から音声データを取得し，同様の音声セグメンテーションを実施した．当該データは，リアルタイムに寄せられる視聴者からのコメントや，事前に寄せられたお便りの読み上げ・応答などを含み，朗読音声とはドメインが大きく異なる．

表 1 本研究で収集した睡眠誘発音声と雑談配信音声

	セグメント数	セグメント長平均 [s]	総時間 [h]
睡眠誘発	2,890	2.27	1.83
雑談配信	4,020	2.96	3.33

表 2 inaSpeechSegmenter による音声区間ラベリング結果

	音声	無音	ノイズ	音楽
睡眠誘発	0.602	0.199	0.185	0.014
雑談配信	0.859	0.115	0.019	0.007

3.3 音声収集結果

表 1 に本研究で収集した音声刺激のセグメント数と総時間を示す．音声セグメントの割合は，睡眠誘発，雑談配信でそれぞれ 60.2%，85.9%であった．この結果を詳細に分析するため，inaSpeechSegmenter による音声セグメントの分類結果を表 2 に示す．どちらの動画にも冒頭 100 秒程度の区間に背景音があり，睡眠誘発動画にはさらに終了 130 秒程度の区間にも同様の背景音があったため，これらが「音楽」として分類されていることを確認した．また，「無音」と「ノイズ」として分類されているセグメントの割合は睡眠誘発のほうが高いが，これらが朗読の段落切り替わり時の間の取り方や，実際の無音区間中で瞬間的に発話が生じるセグメントにより構成されていることも確認した．

3.4 韻律特徴量の分析

睡眠を誘発する音声とそうではない音声の違いを明らかにするために，音声の韻律特徴量として基本周波数 (F0)，エネルギーをセグメント毎に抽出し，その平均と標準偏差 (std) を計算した．F0 の推定には WORLD ボコーダ [18], [19] を用いた．図 1 に計算結果のバイオリンプロットを示す．いずれの統計量についても，睡眠導入音声が中央値の観点で小さく，分布の広がりの観点で狭くなる傾向にあることが確認できる．

4. 音声印象に関する主観評価

4.1 評価条件

3.4 節で示された睡眠誘発音声の統計的差異が音声の主

[*2] https://github.com/sarulab-speech/jtubespeech
[*3] YouTube チャンネル登録者は 12 万人超（本稿執筆時点）
[*4] https://github.com/openai/whisper

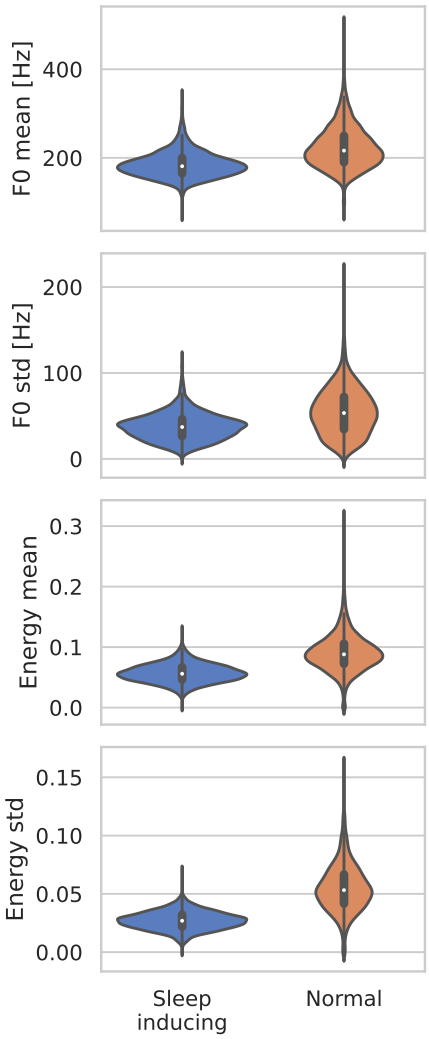

図 1 韻律特徴量の統計量

観的印象に影響を及ぼすかどうかを調査するための主観評価を実施した．

音声刺激の準備： 図 1 に示す 4 つの統計量の四分位数 (Q1–Q4) を計算し，その値に基づいて (1) 睡眠誘発音声，(2) 通常発話音声をそれぞれ 4 つのグループに分割した．四分位数を計算する際，音声刺激長が 3.5 [s] 以下のセグメントは除外した．音声刺激のサンプリング周波数は 48,000 [Hz] であり，信号処理による音声加工を想定し，WORLD ボコーダによる分析再合成処理を施した．最終的に，分割された各グループから無作為に抽出された 50 サンプルずつを主観評価で提示する音声刺激として用いた．

評価方法： クラウドソーシングにより被験者を募集し，提示された音声刺激に対する印象を valence（不快–快）と arousal（睡眠–覚醒）の 2 軸 [20] で回答させる主観評価を実施した．評価軸のスケールは 1 から 7 の 7 段階とし，「不快」および「睡眠」を 1，「快」および「覚醒」を 7 として設定した．評価時のインストラクションとして，森らの先行研究 [21] を参考に，valence と arousal を以下の基準で回答するよう指示した．

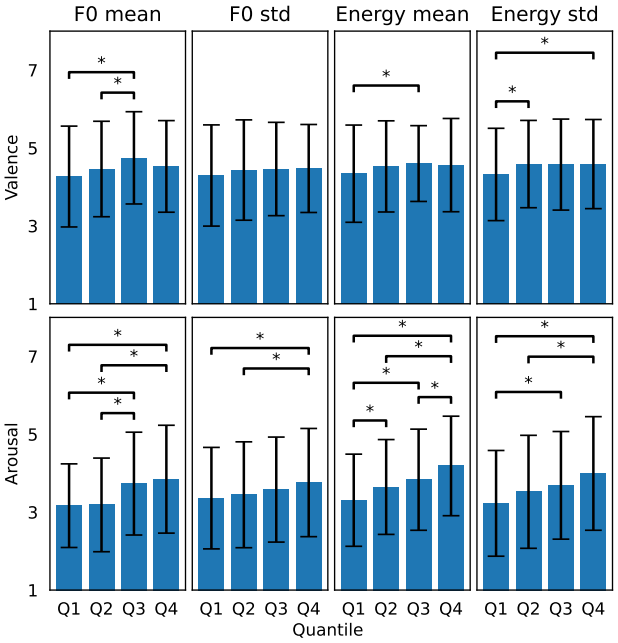

図 2 主観評価結果（睡眠誘発音声）

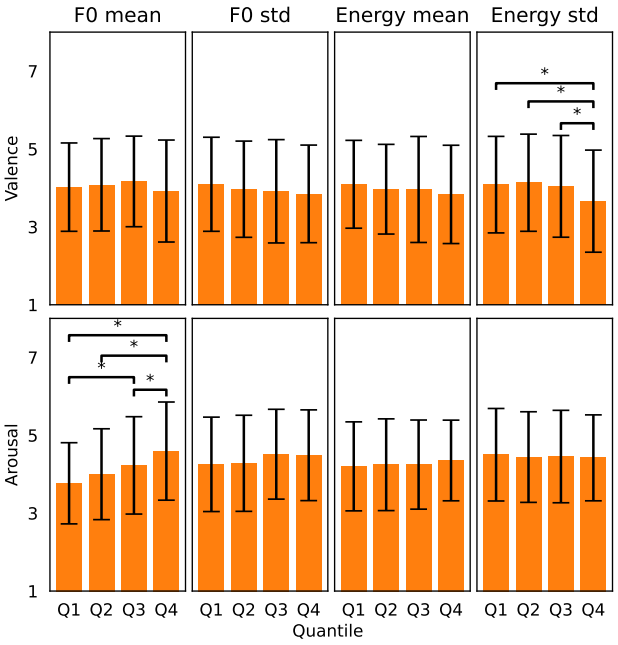

図 3 主観評価結果（通常発話音声）

- **Valence**：この音声を聴いたあとのあなたの気分の不快さ・快さ
- **Arousal**：この音声を聴いたあとのあなたの心理状態の活発さ

主観評価では，四分位数基準で分割された 4 グループ（50 サンプル × 4）から無作為に抽出した 20 サンプル (5 × 4) の音声刺激を被験者に提示した．評価の数は 4（統計量）× 2（睡眠誘発 or 通常発話）= 8 であり，各評価で 30 名ずつ（計 240 名）被験者を募集した．

図 2 と図 3 にそれぞれ睡眠誘発音声と通常発話音声を提示した主観評価の結果を示す．ここで，図中の “*” はグ

ループ間に $p < 0.05$ で有意差があることを意味する．

まず，睡眠誘発音声に対する主観評価結果（図 2）について，valence（不快–快）の評価軸で有意差が観測された四分位グループ対は以下のみであった．

- **F0 mean:** Q1 < Q3, Q2 < Q3
- **Energy mean:** Q1 < Q3
- **Energy std:** Q1 < Q2, Q1 < Q4

これらの結果は，韻律特徴量の平均が小さく，エネルギーのばらつきが小さい音声は不快と評価されやすい傾向を示唆するが，その要因としてはこのような音声に対するボコーダ分析再合成の品質劣化が顕著であったことが考えられる．一方，arousal（睡眠–覚醒）の評価軸では，概して韻律特徴量の平均・標準偏差が小さい音声が低いスコアを獲得しており，特に，エネルギーの平均で四分位グループ分割を実施した場合の評価結果は全ての比較対で有意差が観測された．これらの結果は，睡眠誘発を意識して発話された音声の中でも受聴を通じて知覚される arousal には有意な差があるが，抑揚が小さく，音高・音量ともに小さい音声セグメントが特に高い睡眠誘発効果を与える可能性を示唆した．

次に，通常発話音声に対する主観評価結果（図 3）について，valence（不快–快）の評価軸で有意差が観測された四分位グループ対は以下のみであった．

- **Energy std:** Q1 > Q4, Q2 > Q4, Q3 > Q4

この結果は，通常発話においてエネルギーの標準偏差が極端に大きい音声は不快と評価される傾向にあることを示唆する．一方，arousal（睡眠–覚醒）の評価軸では，睡眠誘発音声に対する評価結果とは異なり，有意差は F0 の平均で四分位グループ分割を実施した場合のみに観測され，Q1 < Q3, Q1 < Q4, Q2 < Q4, Q3 < Q4 であった．この結果は，睡眠誘発を意識せずに通常どおり発話された音声であっても，F0 の平均を下げるような加工によって受聴者に睡眠誘発効果を擬似的に与えられるという可能性を示唆した．

5. おわりに

本研究では，利用者が手軽に聴取できる睡眠導入音声コンテンツ生成への応用を視野に入れ，睡眠導入を目的として発話された音声刺激の分析を実施し，主観評価の結果から，F0 とエネルギーの平均と標準偏差が小さい音声が睡眠を誘発する印象を与える傾向にあることを示した．今後は，本研究で立てられた仮説を検証するために，音声加工を介入させた場合の主観的印象の変化を調査する．また，韻律特徴量を操作可能なテキスト音声合成モデル [22] により，人工的に合成した音声に対しても韻律操作により睡眠誘発効果を与えられるかどうかも検討する．

参考文献

[1] E. Altena, C. Baglioni, C. A. Espie, J. Ellis, D. Gavriloff, B. Holzinger, A. Schlarb, L. Frase, S. Jernelöv, and D. Riemann, “Dealing with sleep problems during home confinement due to the COVID-19 outbreak: Practical recommendations from a task force of the european CBT-I Academy,” *Journal of Sleep Research*, vol. 29, no. 4, May 2020.

[2] L. Degenhardt, S. Darke, and P. Dillon, “GHB use among Australians: characteristics, use patterns and associated harm,” *Drug and Alcohol Dependence*, vol. 67, no. 1, pp. 89–94, Jun. 2002.

[3] D. Aldridge, “An overview of music therapy research,” *Complementary Therapies in Medicine*, vol. 2, no. 4, pp. 204–216, Oct. 1994.

[4] G. T. Dickson and E. Schubert, “Musical features that aid sleep,” *Musicae Scientiae*, vol. 26, no. 3, pp. 497–515, Dec. 2020.

[5] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proc. ICCV*, Venice, Italy, Oct. 2017, pp. 2223–2232.

[6] J. Yang, C. Min, A. Mathur, and F. Kawsar, “SleepGAN: towards personalized sleep therapy music,” in *Proc. ICASSP*, Singapore, May 2022, pp. 966–970.

[7] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, “LibriTTS: a corpus derived from LibriSpeech for text-to-speech,” in *Proc. INTERSPEECH*, Graz, Austria, Sep. 2019, pp. 1526–1530.

[8] S. Takamichi, R. Sonobe, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, “JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research,” *Acoustical Science and Technology*, vol. 41, no. 5, pp. 761–768, Sep. 2020.

[9] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, “AISHELL-3: A multi-speaker Mandarin TTS corpus,” in *Proc. INTERSPEECH*, Brno, Czech Republic, Sep. 2021, pp. 2756–2760.

[10] Z. Yang, B. Hu, and J. Hirschberg, “Predicting humor by learning from time-aligned comments,” in *Proc. INTERSPEECH*, Graz, Austria, Sep. 2019, pp. 496–500.

[11] G. Chen, S. Chai, G.-B. Wang, J. Du, W.-Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang, M. Jin, S. Khudanpur, S. Watanabe, S. Zhao, W. Zou, X. Li, X. Yao, Y. Wang, Z. You, and Z. Yan, “GigaSpeech: an evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio,” in *Proc. INTERSPEECH*, Brno, Czech Republic, Sep. 2021, pp. 3670–3674.

[12] S. Takamichi, W. Nakata, N. Tanji, and H. Saruwatari, “J-MAC: Japanese multi-speaker audiobook corpus for speech synthesis,” in *Proc. INTERSPEECH*, Incheon, South Korea, Sep. 2022, pp. 2358–2362.

[13] S. Takamichi, L. Kürzinger, T. Saeki, S. Shiota, and S. Watanabe, “JTubeSpeech: corpus of Japanese speech collected from YouTube for speech recognition and speaker verification,” *arXiv*, vol. abs/2112.09323, 2021.

[14] Y. Sagisaka, “Speech synthesis by rule using an optimal selection of non-uniform synthesis units,” in *Proc. ICASSP*, New York, U.S.A., Apr. 1988, pp. 679–682.

[15] 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, Nov. 2007.

[16] K. Seki, S. Takamichi, T. Saeki, and H. Saruwatari,

"Text-to-speech synthesis from dark data with evaluation-in-the-loop data selection," in *Proc. ICASSP*, Rhodes, Greece, Jun. 2023, (ACCEPTED).

[17] D. Doukhan, J. Carrive, F. Vallet, A. Larcher, and S. Meignier, "An open-source speaker gender detection framework for monitoring gender equality," in *Proc. ICASSP*, Calgary, Canada, Apr. 2018, pp. 5214–5218.

[18] 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.

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

[20] J. A. Russel, M. Lewicka, and T. Niit, "A cross-cultural study of a circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 57, no. 5, pp. 848–856, Apr. 1989.

[21] 森大毅, 相澤宏, and 粕谷英樹, "対話音声のパラ言語情報ラベリングの安定性," **日本音響学会誌**, vol. 61, no. 12, pp. 690–697, 2005.

[22] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *Proc. ICLR*, Vienna, Austria, May 2021.