解説

インターネット時代の音声コーパスの作成*

高道慎之介（東京大学）**

1. はじめに

音声音響工学はこの 30 年間で大きく進化してきたが，その背景には音声コーパスの存在が不可欠である．今日の研究においてコーパスは一般的に用いられているが，一方で 1998 年の論文には「音声コーパスの必要性やその意義については近年広く認められるようになってきた」 [1] という記述がある．これから，コーパスが当たり前のツールとして認識されるようになったのは，音声音響工学の歴史においてそれほど古くないと言える．

これまで，音声コーパスの作成には，専門家の監督の下で話者に発話させその内容を注釈する方法が採られてきた．しかしながら，音声音響処理技術の発展に伴い，コーパスの構築方法論も変化してきた．そのひとつが，既存の大量の音データから，研究目的に資するものを得る方法である．この方法の背景には，当然ながらインターネット及び音メディアがある．そこで本稿では「インターネット時代の音声コーパスの作成」と題し，コーパス構築の方法論を概説する．

2. なぜ我々はダークデータを使うべきなのか

本稿では，監督下で収集されるデータを**実験室データ**と称する．対して，(インターネット上に存在する）工学的な利用可能性が未知であるデータを**ダークデータ** [2, 3] と称する [1]．例えば，インターネットから無作為にダウンロードした音データがダークデータとなる．

実験室データに基づくコーパスは，所望の目的に資するよう専門家が設計，監督，収集することで得られる．一方でダークデータに基づくコーパスは，**図-1** のように収集済みデータから目的に資するものを選択，加工することで得られる．では，なぜ我々はダークデータを使うべきなのか [2]．その理由を整理する．

図–1 画像生成した「インターネット上の音データを収集して洗練する研究者」．Bing Image Creator を使用．

タスクの複合化：音声認識合成における基本タスク（例えば，朗読音声の認識合成）の性能が人間のそれと同程度に達し [4, 5]，より複合的なタスクを研究で扱えるようになった．これに適う実験室データを収集するには，基本タスクよりも多くの人的・金銭的コストがかかる．例えば，必要な話者や実験環境を準備できるとは限らず，準備できたとしてもコストが増大する．

データ規模のスケール則：学習データ量に対する機械学習モデル（特に Transformer [6]）の性能 [3] は，今のところスケーリング則に従っている [7]．故に，学習データ量は多いほうが好ましく，少量データを前提に機械学習を工夫する前に，大量データを観測する方法を検討するほうが良い．

データ中心の機械学習への移行：固定コーパスの下で機械学習モデルを改善する方法論を，**モデル中心**

---

* Speech Corpus Compilation in the Era of the Internet.

** Shinnosuke Takamichi (University of Tokyo)

[1] ダークデータに似た呼称に in-the-wild データがある．これは「所望の工学利用のために実環境で収集されたデータ」の意味合いが強い．両者は，含まれるデータの範囲において異なるが，本稿で紹介する方法の多くは，in-the-wild データにも利用できる．

[2] 実験室データが不要であるという意味ではない．

[3] 計算量や主観品質ではなく，目的関数値の良さを指す．

AI (model-centric artificial intelligence) と呼ぶ．現在，音声音響処理に関する国際会議論文の多くがこの方法論に基づく．他方，実応用において観測されるデータ集合は固定でなく更新される．また，低品質データがコーパスに含まれる場合，機械学習モデルの性能は顕著に劣化する．これらに基づいて展開される方法論が**データ中心 AI (data-centric AI)** である[4]．この方法論では，モデル（の集合）を固定した下で学習データを改善する．この方法論は，大量に存在するダークデータとの相性が良く今後の発展が見込まれる．

著作権法の改正：2019 年施行の著作権法において，機械学習における著作物利用の内容が記述された．著作権法 第 30 条の 4 「著作物は、次に掲げる場合その他の当該著作物に表現された思想又は感情を自ら享受し又は他人に享受させることを目的としない場合には、その必要と認められる限度において、いずれの方法によるかを問わず、利用することができる。（後略）」[5]と同 2 号により，情報解析の用に供する場合に，著作権者の許諾なしに著作物を機械学習に利用できることが明記された．これにより，日本国内において適法となるデータの範囲が拡大された．

話題が逸れるが，国際会議 INTERSPEECH 2023 において音声研究界隈への意識調査の結果が報告された [8]．本調査は 1997 年から 6 年毎に開催され，今回は 2021 年の結果が報告された．興味深いのは，今回の調査で初めて “no more need for speech research” (音声工学研究はこれ以上必要ない [時代は来るか]) に対し，“never” (来ない) と回答した割合が激減したことである．そして，いつその時代が来るかに対する回答の中央値は 2100 年であった[6]．音声音響処理を枯れた技術[7]にするために，処理に資するダークデータの採掘，およびダークデータに耐えうる音声音響処理を展開するのが現代のフェーズであると著者は考える．

[4] https://dcai.csail.mit.edu/. Andrew Ng 博士の解説が易しい．https://www.youtube.com/watch?v=06-AZXmwHjo を参照．

[5] https://elaws.e-gov.go.jp/document?lawid=345AC0000000048

[6] 著者もこれに概ね同意する．

[7] 広く使われることで信頼性の高い技術．

3. どう我々はダークデータを使うべきなのか

さて，2 章ではダークデータの使用を推奨したが，一方で実験室データの使用には無い問題が生まれる．本章ではこの問題と緩和方法を整理する．

管理下の音声収録であれば，収録環境を設計して収録音声の取捨選択を即座に行うことで，コーパス品質を担保できる．しかしながらダークデータの場合，環境と収録を監督できないため低品質データが含まれてしまう．例として，インターネットの音声コンテンツをダウンロードして，音声合成コーパスを構築する場合を考える．スタジオ品質の音声データのみを収集したいにも関わらず，背景音が多く含まれる，音声が歪んでいる，あるいは音声がそもそも含まれない場合もある．データが少量であれば手作業で評価してよいが，データ規模の大きさを享受するには，データ品質を自動で定量化するほうが好ましい．そこで本章では，ダークデータの「音声音響処理に使える度合い」を定量化する方法を紹介する．なお，ノイジーな（音響的な意味に限らない）データに対する音声音響処理 [9–11] もあるが，本稿では省略する．

3.1 音　　質

研究目的に応じて，音データに要請される品質条件は異なる．ダークデータには様々な品質の音データが含まれるため，その音質の定量化が必要となる．音質定量化手法として STOI (short-time objective intelligibility) [12] や PESQ (perceptual evaluation of speech quality)[8] 等のリファレンスあり手法 [13, 14] が従来使われてきたが，リファレンスの存在しないダークデータにこれらの手法を適用することはできない．

対して近年，リファレンスなし（reference-free）で音質を予測する機械学習が登場している．STOI 値をリファレンスなしに予測する手法 [15] に加え，主観評価値を主観評価なしに予測する方法もある [11, 16–19]．主観評価値は評価者や評価音セットによって異なるため，このバイアスを無視して予測することは適切でないが，これを緩和する方法が検討されている [20]．

当該音が所望する音質に至らない場合，音源分離 [21, 22] や復元 [23, 24] 等を用いて音質改善を試みることがある．一昔前の低性能な処理であれ

[8] https://www.itu.int/rec/T-REC-P.862.2

ばエラーの蓄積を無視できなかったが，性能が成熟した昨今は蓄積を無視できる場合も多い．

3.2 データ対としての質

音データをモダリティ変換に用いるには，音データと他モーダルデータ（例えば，言語，画像）の対が必要である．ダークデータからデータ対を選択するために，当該データ対がどの程度対応しているかを定量化する．

典型的には学習済みモデルを用いる．例えば，学習済みの音声認識モデルを用いて単語誤り率やアライメントスコアを計算することで，音声と書き起こし文の対応度合いを定量化できる [25–27]．昨今では，モダリティ間の対照学習 (contrastive learning) モデルを用いて類似の枠組みが展開されている．画像–説明文コーパス作成においては，CLIP (contrastive language-image pre-training) モデル [28][9] を用いる方法がある [29]．対の画像と説明文の埋め込みベクトルをそれぞれ求め，ベクトル間の距離が近いほど当該データ対が対応していると見做す [30]．音イベント–説明文版の CLIP である CLAP (contrastive language-audio pre-training) モデル [31] を用いることで，同様の枠組みで音イベント–説明文コーパスを作成できる [32]．

学習済みモデルを利用しない方法も有効である．例えば，音声認識や翻訳においては，入力音声の長さに対する出力文の長さの見当を事前につけられる．そのため，長さに関する制限を設けることで明らかに不対応であるデータ対を排除できる [33, 34]．

3.3 多様性，画一性

研究目的に応じて様々な多様性と画一性がコーパスに要請される．例えば，古典的な音声合成コーパス（例えば ATR503 [35]，JSUT [36]）では，音素エントロピーあるいは文字カバレッジに基づいてテキスト多様性を担保しつつ，画一的なスタイルでの発話を収集している．では，所望の多様性と画一性を担保するために，どのようにダークデータを選択すれば良いだろうか．

多様性について，データに付随するメタ情報（例えば，動画のタイトル文，カテゴリ，ユーザコメント文）で評価する簡便な方法がある．例えば，動画カテゴリのカバレッジにより発話の多様性を評価できる [25, 37, 38]．また，本内容はコアセット選択と見做すことができ，多様性指標に基づく選択が可能である．多様性指標として，従来のエントロピーやカバレッジ [39, 40] の他に，埋め込みベクトル（例えば，BERT (bidirectional encoder representations from Transformers) の出力ベクトル [41]）の共分散行列の行列式 [42] や類似度総和 [43] がある．多様性担保では，この指標値を最大にするようにデータを選択する．

画一性についても種々の方法がある．多様性と画一性は真逆の性質であるため，画一性担保では，多様性担保に類似した指標を逆向きに改善すれば良い．話者のベクトル表現である x-vector [44]，収録機器特性の device-quality x-vector [45]，対話文脈の dialogue embedding [46] 等があり，これらの共分散行列の行列式 [26] や類似度総和 [46, 47] を最小化する．

3.4 合成音声の検出

音合成によって学習データをかさ増しすること（data augmentation）は，機械学習の常套手段である [48–51]．一方で，自然音と合成音はその性質が異なるため，これらを区別しなければならない場合がある．

ダークデータを対象とした研究ではないが，ディープフェイクとして合成音を検出する方法がある [52–54]．この方法では，自然音と合成音を識別するモデルを学習する．一定の性能が認められるものの，ダークデータに適用する場合，当該音が対象自然音（あるいは合成音）のみから構成されるとは限らず，背景音の混入等を考慮しなければならない．既存の識別モデルは，耐雑音性能が顕著に悪いことが報告されており [55]，本件は未解決問題である．背景音の混入した合成音を検出する方法が現在も研究されている [56, 57]．

3.5 データの類似性

インターネットからダークデータを収集すると，ほぼ同一のデータを重複して収集してしまうことがある．例えば，ある動画を引用（あるいは転載）した別動画が存在する場合，同一の音データを複数回収集することになる．これを使用する場合，音声音響処理において予期しないデータ加重[10]やデータリーク[11]を生じさせる恐れがある．

動画の転載検出を目的として，自己教師あり学

[9] 対応する画像と説明文を同じベクトルに埋め込むよう学習されたモデル

[10] 各学習データに学習時の重要度を付与すること．

[11] 学習データと同一のものが評価データに含まれること．

習モデルを利用する方法がある [58]．このモデルの特徴量は発話内容に加え話者情報を内包している [59, 60] ため，発話間の特徴量系列の距離を図ることで転載された音データを検出する．また関連研究として，言語モデルの学習において頻出単語列のデータ重みを抑える方法がある [61]．これは転載検出を目的とした手法ではないが，重複データの重みを減ずる方法と見做すことができる．

3.6 不適切表現，有害コンテンツ，固有名詞

データソースによっては，NSFW (not safe for work) や，ヘイトスピーチ [12]，有害コンテンツなどを含む．これらのデータがコーパスに含まれないように排除しなければならない．

画像に対する有害コンテンツ検出 [62] や，テキストに対するヘイトスピーチ検出 [63] が存在する．後者については日本語を対象とした研究もあり，データセットや辞書の整備が進められている [64][13]．また近年では，前述した画像–説明文間の対照学習モデルを用いた検出も提案されている [65, 66]．著者の知る限り音データを直接的に扱う研究は今のところ無いが，発話内容や動画像に対する検出を利用できる．

上記に関連して，固有名詞を削除したい場合がある．例えば，audio captioning[14] において特定企業名を用いて入力音の内容を表現しない [33]，あるいは，prompt text-to-speech[15] において特定個人名の入力から当該個人の声を再現しない [38] ために，コーパスから固有名詞を削除する．これに対しては，文に対する固有表現抽出を用いる [33, 38]．

3.7 ドメインらしさ

何らかのドメイン（例えば，音声認識における発話ドメイン）に特化した処理を行う場合，データを無作為に集めるよりも所望ドメインのデータを収集したい．では，ダークデータから所望ドメインの音データを選択するにはどうすべきか．

自然言語処理において，モデル尤度差に基づいてドメイン特化の文セットを得る方法がある [67, 68]．具体的には，ドメイン非依存の汎用言語モデルとドメイン特化言語モデルの尤度差を図ることで，対象文のドメインらしさを定量化している．この音声版と見做される方法として，自己教師あり学習モデルによる音声シンボル系列（例えば，HuBERT [69] の出力ベクトルを k-means クラスタリングしたもの [70]）を用いる方法が提案されている [71]．文字列も音声シンボル系列も離散シンボル系列であるため，音声シンボル系列を前述の方法と同様に尤度差で扱う．

音データに付随するメタ情報を利用する方法もある．例えば，当該動画が楽音に関するかを識別する方法 [72]，あるいは魅力的な音声を含むかを識別する方法 [38] がある．これらは，必要なラベルを少量のメタデータに予め付与し，言語モデル（例えば BERT [41]）に基づく識別器を学習したのち，所望ドメインの音データを選択している．

4. どう我々はダークデータに使われるべきなのか

3 章に述べたように，ダークデータは（その名称に反して）有望なデータである．これを活かすために我々がこれからやるべきことを，著者個人の考えに基づいて整理する．

4.1 インターネット時代からシミュレーション時代へ

ダークデータは無尽蔵ではない．「低品質データを含めても，2030 年から 2050 年の間にインターネット上のテキストデータを使い果たす」と予測する論文もある [73]．当該論文において音データに関する言及はないが，もう少し遅い時期に同様の結末を迎えることは想像に固くない．また，データの権利がこれから整理されていくだろう（本章で後述する）．すなわち，本稿で述べた「インターネット時代の音声コーパス作成論」の寿命は，10〜20 年前後と予想される．では，インターネット時代の次に何が来るのか．著者はシミュレーション時代，すなわち人工データ（生成データ）のみで行われる音声音響処理の時代だと考える．

3 章に述べたように，現在もデータ拡張の目的で人工データが使用されている．今後はこの枠組みがより複雑なコンテキストと複数のモダリティに拡張されていくだろう．また，これらを音声音響信号の情報シミュレーションと捉えるならば，信号の物理を扱う物理シミュレーションとも融合さ

[12] 人の内的属性（人種、宗教、ジェンダーなど）に基づいて、ある集団や個人を標的とし、社会の平和をも脅かす可能性のある攻撃的言説.

[13] https://github.com/MosasoM/inappropriate-words-ja

[14] 環境音の内容をテキストで記述する技術.

[15] テキストに基づいて音声の声色を制御する技術.

れていくだろう．これらが実現され将来の研究活動を人工データで賄えるよう，現在の我々がダークデータを用いて基盤を構築しなければならない．

4.2 学習データと生成データの権利を追う

2 章 において，著作物を適法的に機械学習に利用できることを述べた．ただし執筆時点（2023 年 10 月）において，著作物を機械学習に使用することと，機械学習モデルから生成したデータが著作物に類似することは，法的に分けて扱われることに注意したい．音声合成における事例については著者の発表資料を参照されたい [74]．当該資料では，合成音声のセリフ，演技，話者性が，実在人物の著作権，パブリシティ権を侵害するか等を扱っている．ダークデータの利用は，工学に強い恩恵をもたらすが，社会に危機をもたらす兵器にも成り得る．そのため，学習データ及び生成データの法的扱いがこの数年で大きく変わると予想される．ダークデータを音声音響処理に用いる場合には，思考を工学のみに閉じずに適切な技術を構築しなければならない．

4.3 覚える機械学習から忘れる機械学習へ

ダークデータを機械学習に用いる際の懸念は，ダークデータによる学習効果を排除しなければならない場合があることである．これは，前述した法整備やデータ権利者から学習データの部分的な削除を依頼される場合である．欧州議会 [16]，アメリカ合衆国著作権局 [17]，日本の AI 戦略会議 [18]の方針を鑑みても，学習データの透明性が直近の課題となることは明白である．また，メンバーシップ推論攻撃 (membership inference attack) [75–77][19] により学習データが漏洩する懸念もある．

これに対し，指定されたデータの学習効果を学習済みモデルから除く手法 [20]の研究が進められている．この手法は**機械アンラーニング (machine unlearning)** と呼ばれる [78]．2023 年には，NeurIPS 2023 Machine Unlearning Challenge が開催され [21]，顔画像からの年齢推定タスクにおける検討が進んでいる．音声については未着手だが，同様の検討が必要である．

5. ま と め

本稿では，インターネット時代の音声コーパスの作成と題して，インターネットから得られるダークデータの扱いについて述べた．ダークデータの使用は懸念も多く，ダークデータであらゆる音声処理をカバーできるわけでも無い．しかしながら，ダークデータを正しく利用することで研究を次の時代に繋げられるだろう．我々は工学を超えた検討を進めなければならない．

謝辞:本研究は科研費 21H04900, 22H03639, 23H03418, 23K18474，JST 創発的研究支援事業 JP23KJ0828，ムーンショット JPMJPS2011 の助成を受けた．また，本稿の執筆にあたり東京大学 大学院情報理工学系研究科 修士課程 関 健太郎氏からの助言を受けた．

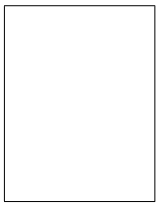
文 献

[1] 板橋秀一, “音声コーパスと音声処理システムの評価,” *Journal of Signal Processing*, vol. 2, no. 6, Nov. 1998.
[2] D. Trajanov et al., “Dark data in Internet of things (IoT): challenges and opportunities,” in *7th Small Systems Simulation Symposium*, 2018, pp. 1–8.
[3] B. Schembera, J. M. Durán, “Dark data as the new challenge for big data science and the introduction of the scientific data officer,” *Philosophy & Technology*, vol. 33, no. 1, pp. 93–115, 2020.
[4] G. Saon et al., “English Conversational Telephone Speech Recognition by Humans and Machines,” in *Proc. INTERSPEECH*, 2017, pp. 132–136.
[5] J. Shen et al., “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” in *Proc. ICASSP*. IEEE, 2018, pp. 4779–4783.
[6] A. Vaswani et al., “Attention is all you need,” in *Advances in Neural Information Processing Systems*, I. Guyon et al., Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
[7] T. Henighan et al., “Scaling laws for autoregressive generative modeling,” *CoRR*, vol. abs/2010.14701, 2020. [Online]. Available: https://arxiv.org/abs/2010.14701
[8] R. K. Moore, R. Marxer, “Progress and Prospects for Spoken Language Technology: Results from Five Sexennial Surveys,” in *Proc. INTERSPEECH 2023*, 2023, pp. 401–405.
[9] K. Saijo, T. Ogawa, “Self-remixing: Unsupervised speech separation via separation and remixing,” in *Proc. ICASSP*, 2023, pp. 1–5.
[10] T. Fujimura et al., “Noisy-target training: A training strategy for dnn-based speech enhancement

[16] https://www.europarl.europa.eu/news/en/press-room/20230505IPR84904/ai-act-a-step-closer-to-the-first-rules-on-artificial-intelligence
[17] https://www.copyright.gov/ai/
[18] https://www8.cao.go.jp/cstp/ai/ai_senryaku/5kai/gijiyoushi5kai.pdf
[19] 機械学習モデルの学習データを推論する攻撃．モデルが公開されずアクセスのみ可能な場合にも発生しうる．
[20] Approximate unlearning と呼ばれる [78].
[21] https://unlearning-challenge.github.io/

without clean speech," in *Proc. EUSIPCO*, 2021, pp. 436–440.
[11] T. Saeki et al., "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022," in *Proc. INTERSPEECH*, 2022, pp. 4521–4525.
[12] C. H. Taal et al., "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. ICASSP*, 2010, pp. 4214–4217.
[13] J. L. Roux et al., "SDR – half-baked or well done?" in *Proc. ICASSP*, 2019, pp. 626–630.
[14] W. A. Jassim et al., "Warp-Q: Quality prediction for generative neural speech codecs," in *Proc. ICASSP*, 2021, pp. 401–405.
[15] A. Kumar et al., "Torchaudio-squim: Reference-less speech quality and intelligibility measures in torchaudio," in *Proc. ICASSP*, 2023, pp. 1–5.
[16] T. Sellam et al., "SQuId: Measuring speech naturalness in many languages," in *Proc. ICASSP*, 2023, pp. 1–5.
[17] P. Manocha, A. Kumar, "Speech Quality Assessment through MOS using Non-Matching References," in *Proc. INTERSPEECH*, 2022, pp. 654–658.
[18] G. Mittag et al., "NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets," in *Proc. INTERSPEECH*, 2021, pp. 2127–2131.
[19] C. K. A. Reddy et al., "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. ICASSP*, 2021, pp. 6493–6497.
[20] E. Cooper et al., "The VoiceMOS Challenge 2023: Zero-shot subjective speech quality prediction for multiple domains," *arXiv:2310.02640*, 2023.
[21] S. Rouard et al., "Hybrid transformers for music source separation," in *ICASSP 23*, 2023.
[22] A. Défossez, "Hybrid spectrogram and waveform source separation," in *Proc. ISMIR*, 2021.
[23] H. Liu et al., "VoiceFixer: A Unified Framework for High-Fidelity Speech Restoration," in *Proc. INTERSPEECH*, 2022, pp. 4232–4236.
[24] T. Saeki et al., "SelfRemaster: Self-Supervised Speech Restoration with Analysis-by-Synthesis Approach Using Channel Modeling," in *Proc. INTERSPEECH*, 2022, pp. 4406–4410.
[25] G. Chen et al., "GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio," in *Proc. INTERSPEECH*, 2021, pp. 3670–3674.
[26] S. Takamichi et al., "JTubeSpeech: corpus of Japanese speech collected from youtube for speech recognition and speaker verification," *arXiv:2112.09323*, 2021.
[27] Y. Yin et al., "Reazonspeech: A free and massive corpus for japanese asr," in **言語処理学会 第 *29* 回年次大会 発表論文集**, Mar. 2023.
[28] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. ICML*, ser. Proceedings of Machine Learning Research, M. Meila, T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763.
[29] P. Sharma et al., "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2556–2565. [Online]. Available: https://aclanthology.org/P18-1238
[30] C. Schuhmann et al., "LAION-5B: An open large-scale dataset for training next generation image-text models," *arXiv:2210.08402*, 2022.
[31] B. Elizalde et al., "CLAP: Learning audio concepts from natural language supervision," *arXiv:2206.04769*, 2022.
[32] Y. Wu et al., "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," *arXiv:2211.06687*, 2022.
[33] X. Mei et al., "WavCaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," *arXiv:2303.17395*, 2023.
[34] R. Ye et al., "GigaST: A 10,000-hour Pseudo Speech Translation Corpus," in *Proc. INTERSPEECH 2023*, 2023, pp. 2168–2172.
[35] A. Kurematsu et al., "Atr japanese speech database as a tool of speech recognition and synthesis," *Speech communication*, vol. 9, no. 4, pp. 357–363, 1990.
[36] R. Sonobe et al., "JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis," *arXiv:1711.00354*, 2017.
[37] S. Ando, H. Fujihara, "Construction of a large-scale japanese asr corpus on tv recordings," in *Proc. ICASSP*, 2021, pp. 6948–6952.
[38] A. Watanabe et al., "Coco-Nut: Corpus of Japanese utterance and voice characteristics description for prompt-based control," *2309.13509*, 2023.
[39] K. Stasaski et al., "More diverse dialogue datasets via diversity-informed data collection," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 4958–4968. [Online]. Available: https://aclanthology.org/2020.acl-main.446
[40] D. Wells et al., "A Low-Resource Pipeline for Text-to-Speech from Found Data With Application to Scottish Gaelic," in *Proc. INTERSPEECH 2023*, 2023, pp. 4324–4328.
[41] J. Devlin et al., "BERT: Pre-training of deep bidirectional Transformers for language understanding," vol. 1, 2019, pp. 4171–4186.
[42] Y.-A. Lai et al., "Diversity, density, and homogeneity: Quantitative characteristic metrics for text collections," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 1739–1746. [Online]. Available: https://aclanthology.org/2020.lrec-1.215
[43] K. Seki et al., "Diversity-based core-set selection for text-to-speech with linguistic and acoustic features," *arXiv:2309.08127*, 2023.
[44] D. Snyder et al., "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*. IEEE, 2018, pp. 5329–5333.
[45] P. O. Gallegos et al., "An Unsupervised Method to Select a Speaker Subset from Large Multi-Speaker Speech Synthesis Datasets," in *Proc. INTERSPEECH*, 2020, pp. 1758–1762.
[46] M. Cekic et al., "Self-supervised speaker recognition training using human-machine dialogues," in *Proc. ICASSP*, 2022, pp. 6132–6136.
[47] I. Yakovlev et al., "VoxTube: a multilingual speaker recognition dataset," in *Proc. INTERSPEECH 2023*, 2023, pp. 2238–2242.

[48] S. E. Eskimez et al., “GAN-Based Data Generation for Speech Emotion Recognition,” in *Proc. INTERSPEECH*, 2020, pp. 3446–3450.
[49] E. Song et al., “TTS-by-TTS 2: Data-Selective Augmentation for Neural Speech Synthesis Using Ranking Support Vector Machine with Variational Autoencoder,” in *Proc. INTERSPEECH*, 2022, pp. 1941–1945.
[50] S. Ueno et al., “Data augmentation for ASR using TTS via a discrete representation,” in *Proc. ASRU*. IEEE, 2021, pp. 68–75.
[51] T. Sugiura et al., “Audio synthesis-based data augmentation considering audio event class,” in *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)*, 2021, pp. 60–64.
[52] H. Khalid et al., “FakeAVCeleb: A novel audio-video multimodal deepfake dataset,” *arXiv:2108.05080*, 2021.
[53] C. Wang et al., “Fully automated end-to-end fake audio detection,” in *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, ser. DDAM ’22. New York, NY, USA: Association for Computing Machinery, 2022, p. 27–33. [Online]. Available: https://doi.org/10.1145/3552466.3556530
[54] J. Yamagishi et al., “ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection,” in *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 2021, pp. 47–54.
[55] 和田賢造 et al., “合成音検出を用いた話者照合のためのデータクレンジングの検討,” in **信学技報**, Feb. 2023, pp. 259–263.
[56] J. Yi et al., “ADD 2022: the first audio deep synthesis detection challenge,” in *Proc. ICASSP*. IEEE, 2022.
[57] L. Cuccovillo et al., “Open challenges in synthetic speech detection,” in *2022 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2022, pp. 1–6.
[58] V. T. Pham et al., “Vietnam-Celeb: a large-scale dataset for Vietnamese speaker recognition,” in *Proc. INTERSPEECH 2023*, 2023, pp. 1918–1922.
[59] A. Pasad et al., “Layer-wise analysis of a self-supervised speech representation model,” in *Proc. ASRU*, 2021, pp. 914–921.
[60] G.-T. Lin et al., “On the utility of self-supervised models for prosody-related tasks,” in *Proc. SLT*, 2023, pp. 1104–1111.
[61] Y. Levine et al., “PMI-masking: Principled masking of correlated spans,” in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=3Aoft6NWFej
[62] G. Laborde, “Deep NN for NSFW Detection.” [Online]. Available: https://github.com/GantMan/nsfw_model
[63] M. Gada et al., “Cyberbullying detection using lstm-cnn architecture and its applications,” in *2021 International Conference on Computer Communication and Informatics (ICCCI)*, 2021, pp. 1–6.
[64] 荒井ひろみ et al., “ソーシャルメディアにおけるヘイトスピーチ検出に向けた日本語データセット構築の試案,” in **言語処理学会 第 *27* 回年次大会 発表論文集**, Mar. 2021, pp. 466–470.
[65] P. Schramowski et al., “Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content?” in *Proc. ACM FAccT*, ser. FAccT ’22. New York, NY, USA: Association for Computing Machinery, 2022, p. 1350–1361. [Online]. Available: https://doi.org/10.1145/3531146.3533192
[66] Z. J. Wang et al., “DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 893–911.
[67] R. C. Moore, W. Lewis, “Intelligent selection of language model training data,” in *Proceedings of the ACL 2010 Conference Short Papers*. Uppsala, Sweden: Association for Computational Linguistics, Jul. 2010, pp. 220–224. [Online]. Available: https://aclanthology.org/P10-2041
[68] 鈴木潤 et al., “ニューラル言語モデルの効率的な学習に向けた代表データ集合の獲得,” in **言語処理学会 第 *28* 回年次大会 発表論文集**, Mar. 2022, pp. 344–348.
[69] W. Hsu et al., “HuBERT: Self-supervised speech representation learning by masked prediction of hidden units,” *CoRR*, vol. abs/2106.07447, 2021. [Online]. Available: https://arxiv.org/abs/2106.07447
[70] K. Lakhotia et al., “On generative spoken language modeling from raw audio,” *Transactions of the ACL*, vol. 9, pp. 1336–1354, 2021.
[71] Z. Zheng et al., “Unsupervised Active Learning: Optimizing Labeling Cost-Effectiveness for Automatic Speech Recognition,” in *Proc. INTERSPEECH 2023*, 2023, pp. 3307–3311.
[72] Q. Huang et al., “MuLan: A joint embedding of music audio and natural language,” in *Proc. ISMIR*, 2022.
[73] P. Villalobos et al., “Will we run out of data? an analysis of the limits of scaling datasets in machine learning,” 2022.
[74] 高道慎之介, “Ai 音声合成の技術動向,” in *CEDEC* - **一般社団法人コンピュータエンターテインメント協会**, Aug. 2023.
[75] R. Shokri et al., “Membership inference attacks against machine learning models,” in *2017 IEEE Symposium on Security and Privacy (SP)*. Los Alamitos, CA, USA: IEEE Computer Society, may 2017, pp. 3–18.
[76] S. Yeom et al., “Privacy risk in machine learning: Analyzing the connection to overfitting,” in *Proc. IEEE CSF Symposium*, 2018, pp. 268–282.
[77] A. Sablayrolles et al., “White-box vs black-box: Bayes optimal strategies for membership inference,” in *Proc. ICML*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 5558–5567. [Online]. Available: https://proceedings.mlr.press/v97/sablayrolles19a.html
[78] H. Zhang et al., “A review on machine unlearning,” *SN Computer Science*, vol. 4, no. 4, p. 337, April 2023.



高道慎之介

2011 年に長岡技術科学大学を卒業．2013 年・2016 年それぞれに奈良先端科学技術大学院大学 博士前期・後期課程を修了．2023 年より東京大学 講師（現職）．博士（工学）．音声合成変換，音声信号処理の研究に従事．