報道関係各位

Press Release

すべての産業の新たな姿をつくる

Laboro

オーダーメイド AI 開発
『カスタム AI』

株式会社 Laboro.AI

2024 年 4 月 30 日

データ量を約３倍に増量しアップデート
日本語音声コーパス「LaboroTVSpeech2」を提供開始

株式会社 Laboro.AI
代表取締役 CEO　椎橋徹夫・代表取締役 COO 兼 CTO　藤原弘将

オーダーメイドの AI・人工知能ソリューション開発および AI 導入コンサルティング『カスタム AI』を展開する株式会社 Laboro.AI（ラボロエーアイ、東京都中央区、代表取締役 CEO 椎橋徹夫・代表取締役 COO 兼 CTO 藤原弘将。以下、当社）は、2020 年に TV 録画から長時間音声と字幕テキストを抽出して音声コーパスを自動構築する独自システムを用いた音声データから構築した日本語音声コーパス「LaboroTVSpeech（ラボロティービースピーチ）」を開発し、学術研究用に無償公開しておりましたが、この度、データ量を約３倍に増量し、より高品質な音声データとしてアップデートした「LaboroTVSpeech2」を開発し、提供を開始いたしました。

LaboroTVSpeech2 開発背景

昨今、AI と機械学習の分野では、大規模なデータセットの存在が重要になってきています。例えば、生成 AI で注目を集める言語モデル GPT では、GPT-1 から GPT-4 への進化においてトレーニングデータサイズの劇的な増加が行われており、現代に求められる AI モデルを開発するためには、大量かつ高品質なデータが AI の精度に大きなプラス効果をもたらすことを示しました。

当社でも、2020 年に提供を開始した旧版 LaboroTVSpeech について、より高品質な音声データを提供したいという思いから、リリース後もテレビ番組データの収集を継続し、今般の開発にいたりました。

LaboroTVSpeech2 について

LaboroTVSpeech2 は、旧版 LaboroTVSpeech と同様に B-CAS カードによるアクセス制限がないワンセグ放送を利用し、2022 年 12 月～2023 年 11 月放送、12 ジャンルの 39,248 の TV 番組、計 6,620 時間のデータから構成されております。

旧版 LaboroTVSpeech が 12 ジャンルの 9,142 の TV 番組、計 2,049 時間のデータで構成されていることと比べると、そのデータ量は約 3 倍と大幅に増加しております。

	音声の長さ（時間）	
ジャンル	旧版LaboroTVSpeech	LaboroTVSpeech2
ニュース／報道	767	2,126
バラエティ	316	924
情報／ワイドショー	323	1,197
ドラマ	206	529
ドキュメンタリー／教養	175	619
趣味／教育	101	639
スポーツ	66	232
アニメ／特撮	39	153
音楽	17	46
福祉	20	80
映画	6	23
劇場／公演	10	52
計	**2,049** 約3倍	**6,620**

LaboroTVSpeech2 を構成する番組ジャンルと音声の長さ

なお、LaboroTVSpeech2 は、旧版と同様に当社が独自開発したシステムにより構築しています。具体的には、テレビ番組の長時間の音声データと、その不完全な書き起こしである字幕データの時間的な対応関係を抽出する手法である準教師付きデコーディング（lightly-supervised decoding）と呼ばれる手法をベースとしています。これにより、本来であればテレビ番組のデータから音声と字幕がセットになって抽出されるべきところ、先のような何らかの問題で対応した情報として取得できなかった場合に、準教師付デコーディングによる音声と字幕の対応関係の抽出を繰り返し行うことで、一度対応が取れなかった区間からも可能な限りデータ抽出を行う仕組みを採用しています。

LaboroTVSpeech2 比較実験について

LaboroTVSpeech2 を用いたモデルの音声認識の性能を確認するため、日本語の TEDx を用いて構築した独自の音声認識システム評価用データセット（※1）を用意した上で、旧版 LaboroTVSpeech との比較実験を行いました。音声認識のツールキットとしては End-to-End 方式を採用する ESPnet を用いました。

その結果、文字誤認識率（CER）が旧版の 13.0％に対して 11.4%となり、1.6%の改善が見られたことを確認いたしました（※２）。

（※1）Youtube 上のプレイリスト「TEDx talks in Jpanaese」に含まれる動画から音声と字幕データを取得したもの。
（※2）上記の結果は、実環境での音声認識システムの性能とは異なる場合があります。

LaboroTVSpeech2 の利用について

LaboroTVSpeech2 に含まれる音声及びテキストデータの権利は、元のテレビ放送の著作権者に帰属していますが、著作権法 30 条の４に基づき、情報解析等の用途のために、大学等の学術研究機関に対して無償で公開いたします。ただし、元のテレビ番組の音声を再構成し鑑賞する事を防ぐために、発話単位でランダムに並び替えられており、かつ番組名や放送局等の付加情報は含まれておりません。

ご利用にあたっては、LaboroTVSpeech2 の利用相談の旨を明記の上、当社 HP 内お問い合わせフォーム（https://laboro.ai/contact/other/）よりお問い合わせください。また、営利企業における研究開発用途や商用目的での利用をご希望の場合も、同じく当社 HP 内お問い合わせフォームからご相談ください。

なお、お問い合わせをいただいてから配布まで最短 3 週間前後のお時間を頂戴しておりますことを何卒ご了承ください。

ご参考情報

■株式会社 Laboro.AI 会社概要
会 社 名：株式会社 Laboro.AI（ラボロ エーアイ）
所 在 地：〒104-0061 東京都中央区銀座八丁目 11-1
代 表 者：代表取締役 CEO　椎橋徹夫
　　代表取締役 COO 兼 CTO　藤原弘将
設　　立：2016 年４月１日
事業内容：機械学習を活用したオーダーメイド型 AI『カスタム AI』の開発
　　カスタム AI 導入のためのコンサルティング
Ｕ Ｒ Ｌ：https://laboro.ai/

株式会社 Laboro.AI は、オーダーメイドの AI ソリューション『カスタム AI』の開発・提供を事業とし、アカデミア（学術分野）で研究される先端の AI・機械学習技術をビジネスへとつなぎ届け、すべての産業の新たな姿をつくることをミッションに掲げています。業界に隔たりなく様々な企業のコアビジネスの改革を支援しており、その専門性から支持を得る国内有数の AI スペシャリスト集団です。

以　上

＜本リリースに関するお問い合わせ＞
株式会社 Laboro.AI マーケティング部 広報担当 中村麗奈
Mail：press@laboro.ai　Tel：03-6280-6564（代表）