Vol.2020-NL-246 No.8
Vol.2020-SLP-134 No.8
2020/12/2

テレビ録画とその字幕を利用した大規模日本語音声コーパスの構築

安藤 慎太郎[1,†1,a)]　藤原 弘将[1,b)]

概要：本研究では大規模な日本語音声認識コーパスの開発について述べる．本コーパスは，日本のテレビ放送とその字幕を用いて開発され，約 2,000 時間もの音声が含まれている．本研究では，長時間音声と不正確なテキストのアラインメントを行う準教師つきデコーディング手法を繰り返し行うことで，音声とテキストの対応区間を従来より多く抽出し，コーパスを自動構築するシステムを開発した．本コーパスを用いて構築した音響モデルを，日本語話し言葉コーパス（CSJ）の評価セットおよび TEDx の講演音声で評価した結果を報告する．実験の結果，TEDx 講演音声に対しては，本コーパスで構築したモデルが CSJ 学習セットで構築したモデルを上回り，本コーパスの有用性が示された．本コーパスは，本稿での実験に用いた Kaldi のモデル学習スクリプトと共に，研究開発用途に公開した．

キーワード：音声認識，コーパス

Construction of a Large-scale Japanese ASR Corpus on TV Recordings and Their Subtitles

Shintaro Ando[1,†1,a)]　Hiromasa Fujihara[1,b)]

Abstract: This paper presents a new large-scale Japanese speech corpus for automatic speech recognition (ASR) systems. This corpus contains over 2,000 hours of speech with transcripts built on Japanese TV recordings and their subtitles. We develop herein an iterative workflow to extract matching audio and subtitle segments from TV recordings based on a conventional method for lightly-supervised audio-to-text alignment. We evaluate a model trained with our corpus using an evaluation dataset built on Japanese TEDx presentation videos and confirm that the performance is better than that trained with the Corpus of Spontaneous Japanese (CSJ). The experiment results show the usefulness of our corpus for training ASR systems. This corpus is made public for the research community along with Kaldi scripts for training the models reported in this paper.

Keywords: Automatic Speech Recognition, corpus

1. 序論

深層学習ベースの音声認識モデルの性能はその学習データの量に大きく依存している．英語では，商用目的の開発では 5,000 時間を超える音声データが用いられることも一般的であり [1–3]，研究目的でも，SwitchBoard-Fisher データセット（約 2,000 時間）や LibriSpeech [4]（約 960 時間）など，大規模な音声コーパスが公開されている．しかし，日本語の音声コーパスのデータ量は，英語と比べて十分とは言い難い．例えば研究用途で多用される日本語話し言葉コーパス（CSJ）[5] は約 600 時間，新聞記事読み上げ音声コーパス（JNAS）[6] は約 90 時間に留まっている．

高品質な音声認識システムを構築するためには大規模な

[1] 株式会社 Laboro.AI
Laboro.AI, Inc.
[†1] 現在，東京大学大学院工学系研究科
Presently with Graduate School of Engineering, The University of Tokyo
[a)] ando@laboro.ai
[b)] fujihara@laboro.ai

音声コーパスが必要だが，その構築にはコストのかかる書き起こしや録音作業が必要であり，容易ではない．その中，人手での作業を行わずに，自動的にデータ収集を行う手法が複数検討されている．VoxForge [7] や Common Voice [8] は，一般ユーザーに音を録音，アップロードしてもらう事で，データを収集している．その他，上述の LibriSpeech [4] コーパスは，一般公開されている大量のオーディオブックの読み上げ文とスクリプトを対応付けることで，既に存在するデータを整形し音声コーパスとしている．また，Web 上の動画とその字幕を用いて音声コーパスを構築する方法も検討されている．TED-LIUM [9] コーパスは，Web 上で公開されている TED プレゼンテーション動画から構築されたものである．一般公開されたデータセットでは無いが，YouTube の動画音声と字幕から得た 100,000 時間を超えるデータを用いて End-to-End 音声認識モデルを構築した先行研究も存在する [10].

多くのテレビ放送には字幕情報が付与されており，それらを音声コーパスとして整形する方法に関しても検討されている．この際，テレビ字幕の時刻情報は不正確であるため，その扱いが課題となる．Multi-Genre Broadcast（MGB）チャレンジ [11,12] では，英語およびアラビア語のテレビ放送データを用いて，音声と字幕の対応をとるアライメントのタスクが提示され，複数の研究グループが強制アライメントおよび準教師つきデコーディング（lightly-supervised decoding）を用いて音声コーパスを作成する手法を提案した [12–15]．また Bang ら [16] は，元々の字幕に付与された時刻情報を有効活用しつつ自動的に微修正する手法を提案し，韓国のテレビ放送を用いて 336 時間の音声コーパスの構築を行った．

本研究では，テレビ放送のデータを活用する研究に倣い，日本のテレビ放送とその字幕を用いて，大規模日本語音声コーパス「LaboroTVSpeech」を構築した．本コーパスは，現時点では，8 ヶ月間のテレビ録画から抽出した約 2,000 時間の音声データから構成されている．コーパス構築手順は完全に自動化されているため，コーパスのデータ量は定期的に増加させることが可能である．

本研究の貢献は以下の通りである．(1) テレビ録画の音声と必ずしも正確ではない字幕データを用いて音声コーパスを自動的に構築するシステムを開発した．(2) 構築された 2,000 時間の音声コーパス LaboroTVSpeech を学術研究目的で活用可能な音声コーパスとして公開した．[*1]本コーパスは，著者らの知る限りにおいて，公開されている日本語音声コーパスとして最大である．(3) 異なるコーパスから学習された音声認識モデルの性能を比較するためのデータセット「TEDxJP-10K」を構築した．このデータセットは LaboroTVSpeech や他のコーパスから独立しているため，学習に用いたコーパスの性能を正当に評価可能である．

本稿の構成は以下の通りである．まず，2 節ではテレビ録画と字幕から音声コーパス LaboroTVSpeech を構築した手順の詳細を述べる．続く 3 節では本コーパスの統計を含めたその詳細を述べる．4 節では音声認識実験を通して本コーパスの比較評価を通してその有用性を報告し，5 節で結論とする．

2. コーパス構築手順

2.1 データ収集と前処理

コーパス構築のための音声・字幕データ及び，番組名やジャンルなどの基本情報は，地上波デジタル放送のワンセグ放送から取得した．ワンセグ放送は，通常のフルセグ放送とは異なり，B-CAS カードを用いたアクセスコントロールがが適用されていないためである．2020 年 2 月から 9 月にかけて，合計 9,142 番組のテレビ録画を録画した．なお録画の際，既に録画済みの番組の再放送や，元々字幕が付与されていない番組は除外した．録画番組に関する統計は，3 節で報告する．

コーパス構築のための前処理として，全音声データは 16 kHz にダウンサンプリングを行った．字幕データは，各字幕セグメントに対して MeCab [17] を用いて形態素単位に分かち書きを行った．固有名詞などに対応するため，辞書として mecab-ipadic-NEologd [18] を用いた．MeCab は 2 桁以上のアラビア数字の発音を推定しないため，アラビア数字は漢字に変換したうえでその発音を推定した．また，英単語のカタカナ読みに対応するため，Bilingual Emacspeak Project [19] で配布されている辞書を追加で利用した．

2.2 音声と字幕のアライメント

テレビ放送には下記のような特徴があり，テレビ放送とその字幕のアライメントを求める際に留意する必要がある．

- 字幕の時刻情報は必ずしも正確とは限らない．例えばニュースなどの生放送番組で字幕が付与されている場合，10 秒以上の遅れを伴うことも多い．また，生放送に限らず，実際の発話と字幕のタイミングは数秒程度ずれることがある．
- 番組自体には字幕が付与されていても，コマーシャルなど，字幕が存在しない音声区間が存在する．
- 字幕テキストが音声の忠実な書き起こしとは限らない．可読性向上のため，口語的な表現は整形されることも多く，特に日本語の場合，助詞や細かい言い回しの変更が頻繁に発生する．
- 特にバラエティー番組などにおいて，発話の一部が字幕テキスト情報ではなく，番組の映像にいわゆるテロップとして付与される場合がある．この場合，字幕テキスト上では複数の単語や文が不規則に削除されて

[*1] 本コーパスは，https://laboro.ai/column/eg-laboro-tv-corpus-jp/ から入手の手続きができる．

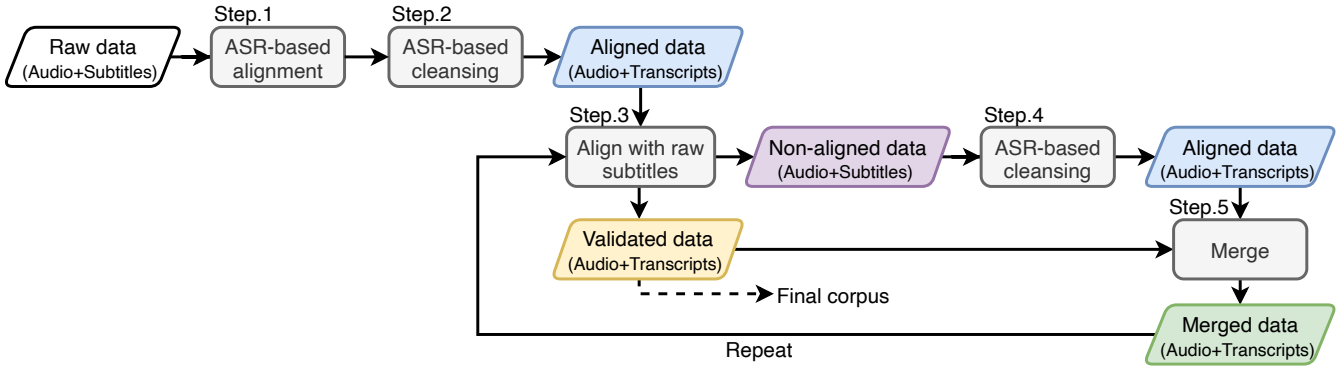

図 1 コーパス構築の概要.
Fig. 1 Diagram of our framework.

いるような状態となる.

MGB チャレンジにおける Alignment タスクの参加者ら [13–15] は, 字幕の時刻情報を利用せず, lightly-supervised decoding を用いたアライメントを検討した. これは, 字幕テキストを用いて再学習した言語モデルを用いて音声認識を行い, 出力テキストと元の字幕テキストのアライメントをとることで音声・字幕対応セグメントを抽出する手法である.

しかしながら先行研究 [9,15] で提案されている手法を今回の日本語テレビ音声・字幕対応タスクに用いたところ, 我々が望むほどの結果は得られなかった. 全番組累計の単語抽出率は 68 % に留まり, また 1 秒未満の短すぎる音声セグメントを除去したところ 63 % にまで低下した. 特にバラエティー番組に関しては, 半分に満たない 45 % という結果であった. これは先行研究で扱われたテレビ放送やオーディオブック等とは異なる, 上述の日本のテレビ番組特有の問題がタスクの難易度を上げたことに一因があると推測される.

[16] において, 著者らは元々の字幕の時刻情報を有効活用する手法を提案している. これは隣接する字幕を結合した上で, 字幕の時刻情報を前後に数秒間延長したうえで, 該当区間の強制アライメントを行うものであるが, 数秒に留まらない遅れを伴う生放送などの字幕には対応できない.

本研究は先行研究における lightly-supervised decoding のアプローチに基づいている. 具体的には, [15] の提案手法が実装された音声認識ツールキット Kaldi [20] のスクリプト[*2]を利用した. 本研究では, 上記の低抽出率の問題を解決するため, lightly-supervised decoding によるアライメントを複数回繰り返すことで, 一度対応が取れなかった音声・字幕区間からも可能な限り抽出を行う手法を提案する.

コーパス構築のための提案手法は以下の 5 ステップから構成されている.

Step 1 各番組の音声に対して, 番組別の字幕データを用いて適応した言語モデル音声認識した上で, Smith–Waterman アライメントを用いて認識結果と元の字幕とのアラインメントをとり, 音声と対応付いた字幕区間のセグメントを抽出する. 本ステップでは `segment_long_uttreances_nnet3.sh` スクリプトを用いた.

Step 2 Step 1 で得た各セグメントに対して, その区間のみに対応するテキストで適応した言語モデルを用いて再度音声認識を行う. そしてその認識結果に基づいて各セグメント内の信頼度の低い区間を除去する. 本ステップでは `clean_and_segment_data_nnet3.sh` スクリプトを用いた. このとき 10 単語に満たない音声セグメントは信頼性が低いため除去した. なおここで得られるテキストは音声認識結果であり, 必ずしも元の字幕テキストと一致しているとは限らない.

Step 3 本ステップまでで抽出されたデータと, 元々の字幕テキストとのアライメントを取ることで, 字幕テキストとテキストが完全に一致した音声セグメントを得る. また, このステップでは, Step 1 および 2 でアラインメントが取れなかった音声・テキスト区間を特定することが可能である. アライメントにあたっては python の `difflib` パッケージにおける `SequenceMatcher` モジュールを利用した. この際アライメントを前向き・後ろ向きでそれぞれ行い, 両方向で全く同一の単語とアライメントが取れたもののみを有効とした.

Step 4 Step 3 で元の字幕とのアライメントが取れなかった各部分に対して, Step 2 と同一のスクリプトを用いて再度アラインメントを行う. ただし, 本ステップでは単語数に基づくセグメント除去は実施していない.

Step 5 Step 3 で一度元の字幕とのアライメントが取れたセグメントと, Step 4 で抽出できたセグメントを合成する. この合成後データを再度 Step 3 を適用し, 3–5 の手順を繰り返す.

[*2] `https://github.com/kaldi-asr/kaldi/blob/master/egs/wsj/s5/steps/cleanup/{segment_long_utterances_nnet3.sh,clean_and_segment_data_nnet3.sh}`

上記の通り Step 3–5 を繰り返すことで，Step 1, 2 だけでは抽出ができなかった区間を抽出することができるため，従来法より多くのデータを得ることができる．これにより，本コーパスの音響的・言語的なバリエーションを増加させることを期待した．現在の実装では最大繰り返し回数は 2 回としている．この繰り返しの後，Step 5 から得られる合成後データを再度 Step 3 に渡すことで，元々の字幕データとの対応付いたデータセットを得た．なお，あまりに短すぎる音声区間は学習時に悪影響を与える可能性があるため，最終的なコーパスからは 1 秒未満のセグメントは除去した．

上記の手順におけるすべての音声認識に関しては，CSJ の講演音声（約 520 時間）で学習した Kaldi のニューラルネット nnet3 形式・TDNN-chain 構造のモデルを使用した．このモデルの構築には Kaldi の公式 CSJ レシピ[*3]を用いた．

3. コーパスの詳細

3.1 コーパスの統計

録画されたテレビ番組とそこから抽出されたコーパスのジャンル別統計を**表 1** に示す．なお各番組は複数のジャンルを持っている場合もあるが，簡単のため各番組で第一ジャンルのみを採用している．最終列の単語の抽出率は（抽出されたコーパスの総単語数）÷（抽出されたコーパスの総単語数）で計算した．音声長の抽出率は，各番組内のコマーシャルの長さ等へも依存するため，ここには示していない．

全体の単語抽出率が 73.8 % であり，全ジャンルにおいて単語抽出率が 50 % 以上であったことから，抽出プロセスが正常に機能していることが分かる．2.2 で述べたように，先行研究の手法をそのまま用いた場合の単語抽出率が 63 % であったことを踏まえると，提案法により，より多くの音声区間を抽出できていることがわかる．より詳細に見ると，単語抽出率は最も低いバラエティーでは 52.3 %，最も高いニュース番組では 89.9 % とジャンル間の差が大きい．これは背景ノイズの量や発話そのものの自発性の程度などの差に起因するものと考えられる．また，元々の全録画におけるジャンルの偏りがあるため，最終的なコーパスでもジャンルごとのデータ量は不均一になっている．とはいえ本コーパスには多種多様な話者・音響環境の発話が収録されており，音声認識モデルの構築に有益であるといえる．

3.2 公開データ

各音声データは字幕と同じ位置で区切られており，サンプリング周波数は 16 kHz である．本コーパスをテレビ番組として鑑賞できなくするため，全発話は録画日や番組，ジャンルに関係無くランダムに並べ替えられている．各音声セグメントに対して話者ラベルは付与されていない．書き起こしテキストは，分かち書き後の単語列の形式であり，各形態素には字幕の前処理の際に MeCab で取得した「名詞」「動詞」等の単純な品詞情報が付与されている．

コーパスは**表 2** に示す通り，学習データ（train）と開発データ（dev）の 2 つのサブセットに分割されている．コーパスの全セグメントから表 1 に示した各ジャンルで 1,000 セグメントずつランダムにサンプリングし，結果得られた 13.7 時間のデータを開発データ，残りの 2036.2 時間を学習データとした．なお本コーパスではテストデータを用意していない．これは，多くの芸能人が様々なの番組に登場し，かつ全番組の正確なキャストの情報を得ることはできないことにより，テストデータの話者を学習・開発データに含めないようデータセットを区切ることが不可能なためである．そこで本コーパスのテストデータの代わりとして，本研究では新たにデータセット「TEDxJP-10K」を構築した．これは日本語の TEDx プレゼンテーションの音声・字幕データから作成したものであり，その詳細は 4.1.1 で述べる．

4. 実験

4.1 実験条件

本研究では，LaboroTVSpeech の音声コーパスとしての性能を，TEDxJP-10K と CSJ の評価セット（eval 1–3）の計 2 つのデータセットを用いて評価した．TEDxJP-10K は音声コーパス評価のため本研究で独自に構築したものである．評価には文字認識誤り率（Character Error Rate; CER）を用いて，このときフィラーや言いよどみ等の全ての文字を評価対象とした．

4.1.1 TEDxJP-10K

TEDxJP-10K は，自発的な発話や背景ノイズ等を含む環境における日本語音声認識システムの性能を評価するために構築したデータセットである．機械学習に用いられる多くのデータセットでは，テストデータはデータセット全体のサブセットとして用意される．そのため学習データとテストデータは類似した特徴を持つことが多い．これはデータを固定してモデルの性能を比較する場合には問題にならないが，データ同士を比較する場合には，テストデータと類似したデータで学習したモデルの性能が自然と高くなってしまい，正当な比較を行うことができないという問題が発生する．本研究においては，CSJ で学習されたモデルと本コーパスで学習されたモデルを比較する際に，CSJ のテストセットや，本コーパスの開発データを用いてしまうと，結果が偏ってしまう．この問題を解決するため，独立したデータセットとして TEDxJP-10K を構築した．

データセットを構築するにあたり，まず日本語の TEDx

[*3] https://github.com/kaldi-asr/kaldi/tree/master/egs/csj/s5

表 1 元々の録画（Raw）と，そこから抽出された最終的なコーパス（Extracted）の番組ジャンル別統計．抽出後の音声長の列における括弧内の数字は，各ジャンルの音声が総合計時間に占める割合を示す．

Table 1 Statistics of the raw recordings and the extracted corpus. The numbers shown in parentheses in the audio lengths denote the percentages of each genre accounted for in the final extracted data.

	Audio length (hours)			#Words in transcripts		
Genre	Raw	Extracted		Raw	Extracted	Extraction rate (%)
News, report	1800.0	767.0	(37.4 %)	9,380 K	8,436 K	89.9
Variety show	1382.9	316.4	(15.4 %)	6,330 K	3,368 K	53.2
Information/tabloid show	820.8	323.3	(15.8 %)	4,812 K	3,604 K	74.9
Drama	664.0	206.0	(10.1 %)	3,175 K	2,304 K	72.6
Documentary/culture	462.4	175.6	(8.6 %)	2,158 K	1,780 K	82.5
Hobby/education	304.1	101.0	(4.9 %)	1,613 K	1,092 K	67.7
Sports	223.9	66.6	(3.2 %)	1,140 K	726 K	63.7
Animation/special effect movies	175.2	39.7	(1.9 %)	801 K	430 K	53.7
Music	99.2	17.6	(0.9 %)	281 K	184 K	65.6
Welfare	58.7	20.1	(1.0 %)	286 K	212 K	74.0
Movies	28.4	6.2	(0.3 %)	121 K	72 K	59.4
Theatre/public performance	23.5	10.4	(0.5 %)	150 K	118 K	78.5
Total	6042.9	2049.9	(100.0 %)	30,253 K	22,331 K	73.8

表 2 LaboroTVSpeech におけるサブセット．

Table 2 Data subsets in LaboroTVSpeech.

	train	dev
Audio length (hours)	2036.2	13.7
# Audio segments	1.6 M	12 K
# Words	22 M	147 K

プレゼンテーションがまとめられた YouTube 上のプレイリスト “TEDx talks in Japanese”[*4]に含まれる動画から音声と字幕データを取得した．そして人手による字幕書き起こしが存在する（すなわち自動生成でない）動画について，音声を字幕の時刻情報に基づいて区切り，それらの全音声セグメントのうち 10,000 セグメントをランダムにサンプリングした．この 10,000 セグメントに対して，アノテータが人手で時間情報の修正と，字幕テキストの修正を行ったものを最終的なデータセットとした．字幕テキストの修正では，字幕が忠実な書き起こしとなるよう，必要に応じてフィラーや言いよどみの挿入や書き換えが行われた．TEDxJP-10K の音声セグメント数は 10,000，合計音声長は 8.8 時間であり，性別・年齢・出自等の異なる 273 名[*5]の話者の音声が含まれている．なお TEDxJP-10K を再構築可能なデータセットとするため，元動画の URL およびアノテータによる編集の差分を公開する[*6]．

4.1.2 音響モデル及び言語モデル

音響モデルは，CSJ，TV，CSJ+TV の 3 種類を比較した．CSJ モデルは 2.2 で使用されたものと同一である．TV モデルは LaboroTVSpeech の学習データを用いて構築し，CSJ+TV モデルは CSJ と LaboroTVSpeech の学習データを混合したコーパスから構築した．全ての音響モデルは Kaldi の公式 CSJ レシピと同一の手順により学習した．使用するコーパスの違いのみを比較するため，パラメータのチューニングは行っていない．また LaboroTVSpeech のデータは各発話に対して話者ラベルが付与されていないため，ケプストラム平均正規化や話者適応学習および *i*-vector 抽出器の計算においては，各セグメントが独立した話者による発話として扱った．

言語モデルに関しては，CSJ レシピに倣い，modified Kneser-Ney discounting を適用した 3-gram 言語モデルを SRILM [21] ツールキットにより学習した．CSJ の評価セットの実験では，全音響モデルに対して CSJ 言語モデルを使用した．TEDxJP-10K の実験では，CSJ 言語モデルに加えて，LaboroTVSpeech の字幕データから学習したモデル（TV 言語モデル）と，TV 言語モデルに Web 上の日本語テキストから構築したモデルを補完したモデル（TV+OSCAR 言語モデル）の 3 つを比較した．TV 言語モデルは LaboroTVSpeech の学習データの全テキストを使用して構築した．その語彙数は約 170 K である．TV+OSCAR 言語モデルは，Web 上の日本語テキストから構築された約 100 GB のテキストデータである OSCAR データセット [22] を用いて学習したモデルと，TV 言語モデルを補完することで作成した．OSCAR データセットの処理においては，絵文字等の特殊文字を取り除いたうえで 2.1 節と同様の手順で単語分かち書きを行い，単語の出現

[*4] https://www.youtube.com/playlist?list=PLsRNoUx8w3rOHjXIU5EE4KOiIagv9yQaG
[*5] 動画によっては 2 話者以上が発話を行う場合もあるが，各動画につき 1 話者として扱った．
[*6] https://laboro.ai/column/eg-laboro-tv-corpus-jp/

表 3　TEDxJP-10K の音声認識結果（CER%）.
Table 3　ASR decoding results (CER%) on the TEDxJP-10K dataset.

	Acoustic model		
LM	CSJ	TV	CSJ+TV
CSJ	23.41	21.83	20.61
TV	22.14	18.98	18.22
OSCAR	23.52	19.28	18.71
TV+OSCAR	21.89	18.29	17.92

表 4　CSJ 評価セットの音声認識結果（CER%）. 全音響モデルで CSJ 言語モデルを使用した.
Table 4　ASR decoding results (CER%) on the CSJ eval dataset using CSJ-LM.

	Acoustic model		
Dataset	CSJ	TV	CSJ+TV
eval1	8.00	14.46	**7.83**
eval2	**6.44**	11.30	6.62
eval3	**5.94**	10.94	6.22

頻度に基づき 200 K 語彙を選択した．2-gram と 3-gram 確率の枝刈りは 10^{-8} を閾値とした．補間の際の各言語モデルの重みは LaboroTVSpeech の開発データのテキストを用いて EM アルゴリズムで推定した.

4.2　実験結果

TEDxJP-10K データセットの音声認識結果を**表 4** に示す．CSJ 音響モデルと TV 音響モデルを比較すると，いずれの言語モデルを使用した場合にも TV 音響モデルが常に良い結果を示した．TV+OSCAR 言語モデルを使用した場合，TV 音響モデルでは CSJ 音響モデルから相対 16.4 % ポイントの CER 低下が観測された．CSJ と TV の混合データによるモデルでは更に性能が向上し，TV+OSCAR 言語モデルを使用した場合，CSJ 音響モデルから CER は相対 18.1 % ポイント低下した.

CSJ 評価セットの音声認識結果を**表 3** に示す．4.1.1 で述べた通り，ドメインの一致により CSJ 音響モデルが TV 音響モデルより良い結果を示した．しかしながら，CSJ+TV を混合したデータセットで学習したモデルの性能は大幅に CSJ 音響モデルとの差を縮め，特に eval 1 では CSJ+TV 音響モデルは相対 2.1 % の CER 低下を示した.

表 3 と表 4 に示すとおり，CSJ+TV 混合モデルは両評価データセットで良い性能を示した．これらの結果は LaboroTVSpeech が音声認識器の学習において有用性のあるコーパスであることを示している.

5.　結論

本研究では，テレビ録画の音声とその字幕から，音声長 2,000 時間を超える日本語音声コーパス LaboroTVSpeech の構築を行った．本コーパスは学術研究用途に公開されている．本稿では本コーパスの詳細を述べた上で，2 つの評価データセットを用いて音声認識システム構築用コーパスとしての性能を評価し，実験の結果より本コーパスの有用性を確認した.

本コーパスの大きな利点は，テレビ放送をデータソースとしているため，データ量を絶えず増加させることが可能である点である．そのため，今後は本コーパスの定期的なアップデートを予定している．また，各テレビ番組はジャンル分けがなされているため，特定ジャンルの番組から抽出したデータのみから構築されたサブセットを作成することができる．そのようなサブセットは特定ドメインにおける音声認識システムの構築に有用であると考えられる.

謝辞　本研究は，株式会社 NTTPC コミュニケーションズの InnovationLAB から GPU リソースの支援を受けました.

参考文献

[1] Jaitly, N., Nguyen, P., Senior, A. and Vanhoucke, V.: Application of Pretrained Deep Neural Networks to Large Vocabulary Speech Recognition, *INTERSPEECH 2012*, pp. 2578–2581 (2012).
[2] Amodei, D. et al.: Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin, *International Conference on Machine Learning*, pp. 173–182 (2016).
[3] Parthasarathi, S. H. K. and Strom, N.: Lessons from Building Acoustic Models with a Million Hours of Speech, *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6670–6674 (2019).
[4] Panayotov, V., Chen, G., Povey, D. and Khudanpur, S.: Librispeech: An ASR Corpus Based on Public Domain Audio Books, *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210 (2015).
[5] Maekawa, K.: Corpus of Spontaneous Japanese: Its Design and Evaluation, *ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 7–12 (2003).
[6] Itou, K. et al.: JNAS: Japanese Speech Corpus for Large Vocabulary Continuous Speech Recognition Research., *Journal of the Acoustical Society of Japan (E)*, Vol. 20, No. 3, pp. 199–206 (1999).
[7] VoxForge: Free Speech... Recognition (Linux, Windows and Mac) - voxforge.org, http://www.voxforge.org/.
[8] Ardila, R. et al.: Common Voice: A Massively-Multilingual Speech Corpus, *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 4218–4222 (2020).
[9] Hernandez, F. et al.: TED-LIUM 3: Twice as Much Data and Corpus Repartition for Experiments on Speaker Adaptation, *SPECOM 2018. Lecture Notes in Computer Science*, Vol. 11096, pp. 198–208 (2018).
[10] Soltau, H., Liao, H. and Sak, H.: Neural Speech Recognizer: Acoustic-to-Word LSTM Model for Large Vocabulary Speech Recognition, *INTERSPEECH 2017*, pp. 3707–3711 (2017).
[11] University of Edinburgh: The MGB Challenge, http://www.mgb-challenge.org/.
[12] Bell, P. et al.: The MGB Challenge: Evaluating Multi-

Genre Broadcast Media Recognition, *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

[13] Bell, P. and Renals, S.: A system for automatic alignment of broadcast media captions using weighted finite-state transducers, *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 675–680 (2015).

[14] Lanchantin, P. et al.: Selection of Multi-Genre Broadcast Data for the Training of Automatic Speech Recognition Systems, *INTERSPEECH 2016*, pp. 3057–3061 (2016).

[15] Manohar, V., Povey, D. and Khudanpur, S.: JHU Kaldi System for Arabic MGB-3 ASR Challenge Using Diarization, Audio-Transcript Alignment and Transfer Learning, *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 346–352 (2017).

[16] Bang, J.-U. et al.: Automatic Construction of a Large-Scale Speech Recognition Database Using Multi-Genre Broadcast Data with Inaccurate Subtitle Timestamps.

[17] Kudo, T.: MeCab: Yet Another Part-of-Speech and Morphological Analyzer, https://taku910.github.io/mecab.

[18] Sato, T.: Neologism Dictionary Based on the Language Resources on the Web for Mecab, https://github.com/neologd/mecab-ipadic-neologd (2015).

[19] Bilingual Emacspeak Project: Bilingual Emacspeak Project, http://www.argv.org/bep/.

[20] Povey, D. et al.: The Kaldi Speech Recognition Toolkit, *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (ASRU)* (2011).

[21] Stolcke, A.: SRILM — AN EXTENSIBLE LANGUAGE MODELING TOOLKIT, *Intl. Conf. on Spoken Language Processing*, Vol. 2, pp. 901–904 (2002).

[22] Suárez, P. J. O., Sagot, B. and Romary, L.: Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures, *Proceedings of the Workshop on Challenges in the Management of Large Corpora*, pp. 9–16 (2019).