情報処理学会論文誌 Vol.62 No.3 903–914 (Mar. 2021)

単言語話者のための日英コードスイッチング音声の認識と翻訳

中山 佐保子[1,2,a)] サクティ サクリアニ[1,2,b)] 中村 哲[1,2,c)]

受付日 2020年6月14日, 採録日 2020年12月1日

概要：会話の中で複数の言語が切り替わる現象は，コードスイッチングと呼ばれる．コードスイッチングは，言語が切り替わる場所や長さによってさまざまなものがある．従来の音声認識システムは，そのようなコードスイッチングを扱うのが難しく，解決すべき課題の 1 つであった．これまで研究されてきたコードスイッチング音声認識は，言語が混ざったコードスイッチング音声を，そのまま言語が混ざったコードスイッチングテキストに書き起こすことだけを目的とした．それは，結局，コードスイッチングを理解できる人だけが認識結果を理解することを想定する．一方，本研究は，コードスイッチング話者と単言語話者の会話を想定し，コードスイッチング話者の発言を単言語話者が理解できるように支援する．コードスイッチング音声認識の認識結果から Bidirectional Encoder Representations from Transformers（BERT）やニューラル機械翻訳を用いて単言語に音声翻訳するカスケードアプローチと，シングルタスクやマルチタスク学習でコードスイッチング音声から単言語テキストに直接音声翻訳する直接アプローチの，合わせて 4 手法を比較し，日英コードスイッチングの音声を単言語の日本語および英語に翻訳するシステムを開発する．

キーワード：コードスイッチング，音声認識，BERT，機械翻訳，マルチタスク学習

Recognition and Translation of Japanese-English Code-switching Speech for Monolingual Speakers

Sahoko Nakayama[1,2,a)] Sakriani Sakti[1,2,b)] Satoshi Nakamura[1,2,c)]

Received: June 14, 2020, Accepted: December 1, 2020

Abstract: Bilingual speakers often mix two or more languages in their conversation. Such a phenomenon is called code-switching (CS). The switching units and positions may be different variously, and the length of a unit can be from word unit to phrase length beyond the loanword unit. The CS phenomenon causes difficulties for automatic speech recognition (ASR) since the system has to be able to control multilingual input. The CS ASR for various language pairs has been investigated in the past. However, most of the goals for developing a CS ASR is to transcribe CS speech into CS text, which supposes that only those who understand the CS use it. In contrast, in this study, we focus on the conversations between CS speakers and monolingual speakers; and we aim to assist monolingual speakers to understand what CS speakers say. We develop a system that recognizes CS speech and translates to monolingual text. We investigated two cascade approaches from ASR by a neural machine translation (NMT) and Bidirectional Encoder Representations from Transformers (BERT), and two direct approaches by single-task learning and multi-task learning. In the end, we compare and review these four ways on a translation task from Japanese-English CS speech.

Keywords: code-switching, automatic speech recognition, BERT, machine translation, multi-task learning

[1] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology, Ikoma, Nara 630–0192, Japan
[2] 理化学研究所革新知能統合研究センター観光情報解析チーム
RIKEN, Center for Advanced Intelligence Project AIP, Tourism Information Analytics Team, Ikoma, Nara 630–0192, Japan
[a)] nakayama.sahoko@is.naist.jp
[b)] ssakti@is.naist.jp
[c)] s-nakamura@is.naist.jp

1. はじめに

教育，仕事，観光などの目的で，外国人居住者や旅行者が増えている．日本の国際結婚の数は，約 40 年前と比較して約 3.5 倍になり [1]，外国人旅行者の数も約 15 年間で約 5 倍に増加した [2]．このような国際化は，人々のコミュニケーションの仕方に影響を与える．バイリンガルの会話

でみられる，言語を混ぜて話す現象はコードスイッチング（CS）と呼ばれ，コミュニケーションの変化の 1 つである．実際に，バイリンガルの子供たちが 4 時間に 153 回の CS を使ったという報告があり [3]，日常生活での CS の使用が明らかになっている．CS は，以下のような，文の途中で言語を切り替える文中 CS と文の切れ目で言語を切り替える文間 CS がある [4]．

- 文中 CS：
  本物の企業家には the ability to see the needs of a market that doesn't yet exist がある．
- 文間 CS：
  I'm looking for a present for a friend in Japan. 何がいいと思いますか？

文中 CS は，単語の長さから外来語を超えるフレーズの長さまで，言語を切り替える単位に幅がある．また，上記の文中 CS の例のように日本語から始まって英語に切り替わるものもあれば，文間 CS の例のように日本語から始まって英語に切り替わるものもある（それらを一括りにして日英 CS と呼び，呼称には日英翻訳のような方向性を含む意味を持たない）．また，2 つの言語の能力がともに高く，簡単に言語を切り替えられることから生じる能力駆動型の CS と，1 つの言語の能力が低いため，その言語で話すときに別の言語を使いながら話さないとならない欠陥駆動型の CS がある [5]．このようにさまざまな CS があるため，単言語の認識を基本とする従来の音声認識（ASR）システムはうまく認識することができず，CS を認識させることは ASR にとって難しい課題の 1 つである．

これまで，多言語の音響モデルの構築 [6] や，言語識別と音素結合を組み合わせた手法 [7]，深層学習による手法 [8] など，さまざまな手法の CS 音声認識が提案されてきた．しかし，それらの CS 音声認識の主な目的は，言語が混ざった CS の音声を，そのまま言語が混ざった CS のテキスト（文字情報）に書き起こすことだけであり，CS を理解する人だけが使用することを想定していた．一方，本研究は，CS 話者と単言語話者の会話を想定し，CS 話者の発言を単言語話者が理解できるように支援する．CS は，CS 話者同士が会話をするときだけではなく，CS 話者と単言語話者が会話をするときにも用いられる．たとえば，アメリカに移民した子供たちのうち，半分以上が英語をうまく話せない両親を持つ [9]．子供たちは，学校では英語を話し，家では母語を話すので，流暢なバイリンガル話者になるという．このような場合，子供が CS を用いて話しても両親は子供たちが言うことを理解できない．そのため，単言語話者が CS 話者の発言を理解できるように，CS の音声を認識し，単言語テキストに翻訳するシステムを開発する必要がある．

テキストからテキストへの機械翻訳であれば，これまでにも，CS 機械翻訳の研究がいくつか存在する．Sinha ら [10] は，各言語を分離することにより，テキストからテキストへの CS 翻訳を実現した．しかし，言語ごとに異なる機械翻訳を用いるため，言語間を超えた文脈を考慮することが，この方法だと難しい．また，Johnson ら [11] によって提案された機械翻訳は，翻訳先の言語を入力で指定することにより多言語の翻訳を実現し，CS 翻訳の可能性を示唆した．しかし，音声からテキストへの翻訳の場合，翻訳先を入力で指定するためテキストと音声の結合が必要であり，正規化せずに用いると精度を悪化させる．また，翻訳先の言語を同じデコーダモデルで学習させると，お互いの言語がバイアスとなって精度を悪化させる．Menacer らはいくつかの方法でアラビア語と英語の CS を翻訳した [12]．最も BLEU スコアが高かったモデルは，翻訳する必要のない言語の部分はそのまま入力から出力にコピーし，それ以外は多言語を使って学習した機械翻訳で翻訳するという手法であった．しかし，この手法は，音声からテキストへの翻訳の場合，入力を出力にそのままコピーすることはできない．また，コピーはせずに CS の文全体を多言語を使って学習させたモデルで翻訳させた場合，比較モデルの中で BLEU スコアが最も低かった．

いずれにしても，本研究は，テキストからテキストへの翻訳ではなく，音声からテキストへの CS 音声翻訳を実現する．そのために，2 つのカスケード（Cascade）アプローチと 2 つの直接（Direct）アプローチを比較する．Cascade アプローチは，ニューラル機械翻訳（NMT）と Bidirectional Encoder Representations from Transformers（BERT）[13] を用いたものであり，Direct アプローチは，CS 音声から単言語テキストを出力するシングルタスク音声翻訳と，CS 音声から単言語テキストと CS テキストの 2 通りを出力するマルチタスク音声翻訳である．我々が以前行った CS 音声翻訳の研究は，CS 音声から英語への翻訳タスクのみを扱ったが [14]，本研究は CS 音声から日本語への翻訳タスクも扱う．

2. CS 音声から単言語翻訳への提案手法

CS 音声から単言語テキストへの音声翻訳を行うため，2 つの Cascade アプローチと 2 つの Direct アプローチを比較する．Cascade アプローチは，CS 音声認識によって CS 音声から書き起こした CS テキストに対して，NMT または BERT を使って機械翻訳を行う手法である．Direct アプローチは，CS 音声認識によって CS テキストに書き起こすプロセスを経ずに，直接シングルタスクまたはマルチタスクで CS 音声から単言語テキストを出力する手法である．これらの提案手法について，1 つずつ紹介する．

2.1 Cascade アプローチ

2.1.1 Cascade ASR+BERT モデル

Cascade アプローチの 1 つ目は，BERT を用いるアプ

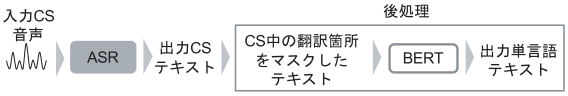

図 1 Cascade ASR+BERT モデル
Fig. 1 Cascade ASR+BERT.

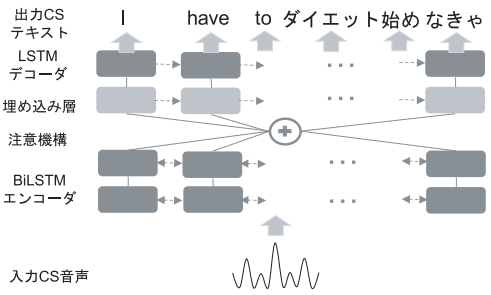

図 2 ASR のモデル構造
Fig. 2 Model architecture of ASR.

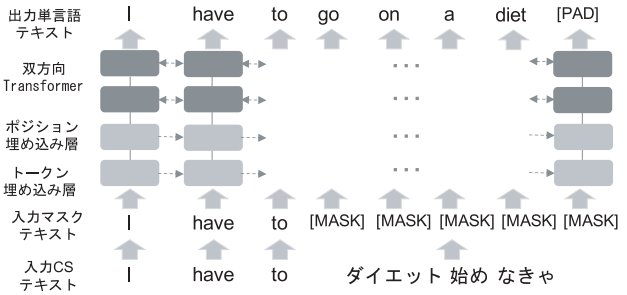

図 3 CS 翻訳のための BERT のモデル構造
Fig. 3 Model architecture of BERT for CS translation.

ローチである．処理の構造を**図 1** に示す．まず，ASR が CS 音声から CS テキストを生成する．次に，BERT 言語モデルが，CS テキストから単言語テキストを予測する．音声の認識に用いる ASR の構造を**図 2** に示す．ASR は，注意機構を持つエンコーダデコーダモデル [15] で，エンコーダは，1 方向あたり 256 の隠れユニットを持つ 3 層の双方向 LSTM（BiLSTM）を持ち，デコーダは 128 次元の埋め込み層と 512 の隠れユニットを持つ 1 層の LSTM を持つ．入力特徴量は対数メルスペクトログラムを用い，活性化関数は LeakyReLU（$l = 1e-2$）[16] を用いた．エンコーダとデコーダをマッピングする注意機構のアライメントスコアは多層パーセプトロンの計算を用いた [17]．

BERT 言語モデルの構造を**図 3** に示す．モデルのパラメータと初期値は，著者らが公開した学習済みの BERT_{Base} モデルに従った．これは，110M のパラメータを持ち，768 の隠れユニットを持つ 12 層の Transformer [18] から構成される．BERT 言語モデルは，従来の言語モデルに比べて強力な言語モデルである．従来の言語モデルは，入力系列に対して単一方向（左から右）のみを学習するが，単語間の文脈の学習に限界があった．一方，BERT は，Transformer を用いて，双方向（左から右および右から左）から単語間の文脈関係を学習する．

表 1 BERT 言語モデルによって予測する単言語テキストの例
Table 1 Example of the monolingual text recovered by BERT.

原文	i have to ダイエット 始め なきゃ before my belly explodes
マスクした文	i have to [MASK] [MASK] [MASK] [MASK] [MASK] before my belly explodes
ラベル	i have to go on a diet [PAD] before my belly explodes
参照訳	i have to go on a diet before my belly explodes

図 4 Cascade ASR+NMT のモデル
Fig. 4 Cascade of ASR+NMT.

BERT は，2 つの学習フェーズで構成される．(1) 言語表現の一般的なデータセットを使用した事前学習，(2) 固有表現抽出 [19]，質問応答 [20]，感情分析 [21] などの特定のタスクを解くためにドメイン固有のデータセットで学習するファインチューニング．事前学習フェーズでは，BERT はランダムに一部のトークンを [MASK] トークンに置き換え，その周辺のトークンを使用して表現を学習することにより，[MASK] トークンに置き換えられたトークンを予測する．Ghazvininejad らは，信頼度の低いトークンを [MASK] トークンに置き換えて予測することを繰り返し，単言語から別の単言語への変換タスクに BERT 言語モデルを活用した [22]．我々の CS 翻訳にも，事前学習フェーズで用いられる BERT 言語モデルを利用した．本研究は，ASR が CS 音声から CS テキストに書き起こした後，第 1 言語を維持しながら第 2 言語の単語を [MASK] トークンに置き換え，BERT 言語モデルが [MASK] トークンを予測して第 1 言語の単言語テキストに変換する．なお，置き換えられる単語の数が分からないため，第 2 言語の各位置に予測するトークンの最大数以上の [MASK] トークンを設定する．**表 1** は，BERT 言語モデルを使用して予測する単言語テキストの例を示す．予測するトークンの数が [MASK] トークンの数より少ない場合，予測箇所のラベルは [PAD] トークンで埋められる．

2.1.2 Cascade ASR+NMT モデル

もう 1 つの Cascade アプローチは，NMT を用いる．**図 4** に，処理の構造を示す．ASR が，CS 音声から CS テキストを予測した後，NMT は CS テキストから単言語テキストに翻訳する．ここで用いる ASR システムは，2.1.1 項で述べた図 2 のモデルと同様のモデルである．NMT モデルは，**図 5** に示すとおり，ASR モデルと同様に注意機構を持つエンコーダデコーダモデルで，エンコーダは，1 方向あたり 256 の隠れユニットを持つ 2 層の BiLSTM を持ち，

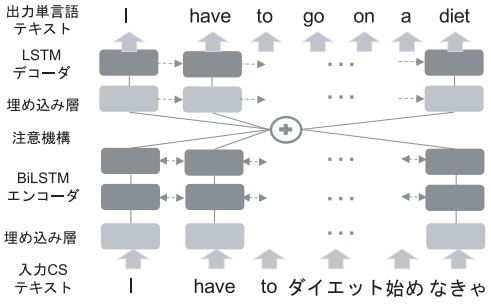

図 5 NMT のモデル構造
Fig. 5 Model architecture of NMT.

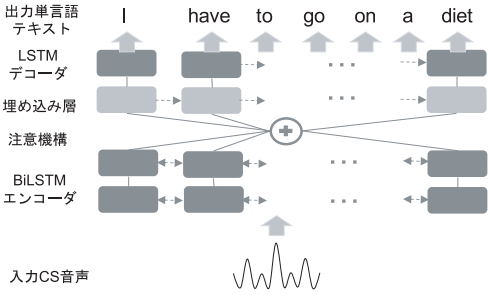

図 6 Direct シングルタスク音声翻訳のモデル構造
Fig. 6 Model architecture of direct single-task speech translation.

デコーダは 128 次元の埋め込み層と 512 の隠れユニットを持つ 2 層の LSTM を持つ．

2.2 Direct アプローチ

2.2.1 Direct シングルタスク音声翻訳

シングルタスク音声翻訳は，CS 音声から単言語テキストへ直接出力する音声翻訳システムである．先述した ASR モデルと同じ構造で，CS 音声から英語テキストを直接生成するようにモデルを学習させた．**図 6** に，このモデルの構造を示す．

2.2.2 Direct マルチタスク音声翻訳

マルチタスク音声翻訳は，マルチタスク学習を用いて CS 音声から単言語テキストと CS テキストの 2 通りの出力を学習し音声翻訳を実現するシステムである．マルチタスク学習はいくつかのバリエーションを持つが，エンコーダを共有し，2 つの並列デコーダを備えたマルチタスク学習 [23] を採用した．並列の 2 つのデコーダは，CS テキストと単言語テキストの両方を出力する．共有するエンコーダおよび並列の 2 つのデコーダは先述した ASR モデルと同じ構造を持つ．**図 7** に，このモデルの構造を示す．

3. コードスイッチングコーパス

合成音声の CS と自然音声の CS の 2 種類のコーパスを作成した．合成音声の CS は，機械翻訳と音声合成を用いて作成した CS であり，自然音声の CS は，日常的に CS を用いるバイリンガル話者から収集された CS である．

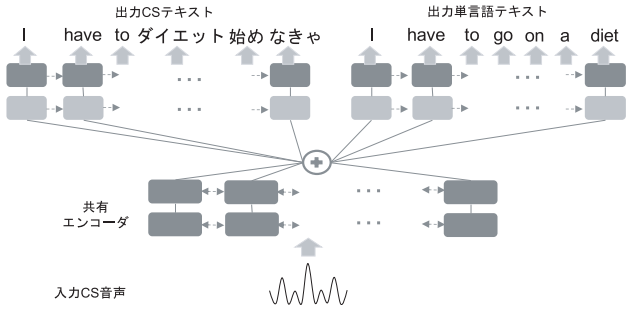

図 7 Direct マルチタスク音声翻訳のモデル構造
Fig. 7 Model architecture of direct multi-task speech translation.

CS は L1 と L2 を切り替えるため非母語音声を含む場合が多い．母語音声と非母語音声では同じ単語でも発音が異なるため，音響モデルと音声の不一致を生じさせ，認識精度の低下を招く原因となる [24]．そのため，L2 言語の習熟度に基づいて CS を分類し，L2 言語が母語話者に近い話者レベルのものを母語話者レベルの CS とし，L2 言語が母語話者の音声とは異なる非母語音声のものを非母語話者レベルの CS とする．本研究は，まだ非母語話者レベルの CS を入手できていないため，母語話者レベルの CS のみを扱うこととする．

また，外来語と引用は，理論的には CS ではないが，本研究は日英会話のすべての単語を翻訳することを目的とするため，外来語や引用も CS として処理することとする．

3.1 合成音声のコードスイッチング

モデルの学習に必要な大規模な CS 対訳コーパスは手に入らなかったため，日本語と英語の ATR 旅行会話コーパス（BTEC）[25] を使用して作成した．Menacer らが行った実験結果によると，機械翻訳を用いて作成した人工的な CS データによるモデルの学習は，CS 機械翻訳システムの開発にとって比較的有効な手法だと分かる [12]．また，Tjandra らが行った実験では，合成音声による実験の結果が自然音声による実験の結果と同じく提案手法の有効性を示すものとなっており，音声合成による音声で実験した結果と人の発話による自然音声で実験した結果の傾向は変わらないことが分かる [26]．そのため，本研究でも，機械翻訳と音声合成による CS データを作成した．

日本語文から作成した CS を「日本語ベース CS」とし，英語文から作成した CS を「英語ベース CS」として 2 種類の CS を作成した．切替え位置は，品詞タグ付けツールである TreeTagger [27] と Mecab [28] の結果に基づいて，選択した．日本語ベース CS は，助詞または名詞を選択し，英語ベース CS は，名詞または動詞，または前置詞を選択した．名詞を選択する場合は，名詞のみの 1 単語，助詞や動詞，前置詞を選択する場合は，それ以降の接続詞や句点の前までのフレーズを機械翻訳によって翻訳し，挿入した．音声の作成は，Google TTS [29] を用いた．

表 2 日英 BTEC の対訳文から作成された自然な CS 文
Table 2 A natural CS sentence created from a pair of Japanese-English BTEC.

日本語文	このごろ披露宴では花嫁さんが二度もお色直しをして，派手らしいですね．
英語文	I hear that nowadays the bride changes her clothes as often as twice during the reception and that the reception is luxurious.
作成文	このごろ披露宴では花嫁さん changes her clothes as often as twice during the reception and that the reception is 派手らしいですね．

表 3 作成された CS の切替え位置の統計
Table 3 The statistics of the created CS switching positions.

日本語から始まる CS	
品詞	割合（%）
助詞	77
名詞	12
助動詞	5
動詞	2
副詞	2
接続詞	1
句点	1

英語から始まる CS	
品詞	割合（%）
名詞	33
動詞	15
接続詞	15
句点	12
前置詞	6
副詞	6
限定詞	6
形容詞	4
代名詞	3

3.2 自然音声のコードスイッチング

自然音声の CS は，日常生活の中で CS を頻繁に用いる日英バイリンガル話者によって作成された．作成者は，英語圏に住んでいるが，1 年だけ日本に留学したことがあり，家庭内では日本人の家族と日本語を話すため，CS を頻繁に用いる．1,000 組の日英 BTEC 文から自然な CS 文を作成してもらった．**表 2** に，作成された CS 文を示す．結果の CS は，日本語の単語を 3,251 単語（24%）と英語の単語を 10,214 単語（76%）含んだ．また，全発話数のうち，日本語から始まるものが 57%で英語から始まるものが 43%であった．さらに，**表 3** に示すとおり，それぞれの言語の切替え位置を調査し，日本語から始まる CS は，助詞が切替え位置の大部分を占め，英語から始まる CS は，名詞や動詞などで切り替わることが多かった．最後に，作成した CS テキストを，テキストの作成者とは異なるバイリンガル話者（幼少期の 5 年間を英語圏で過ごした日本人で，ふだんの会話にも CS を用いる）に読んでもらい，静かな部屋で音声を録音した．作成された 1,000 発話のうち，900 発話は学習セット用に 100 発話はテストセット用に使用した．

3.3 前処理

表 4 に，学習セットとテストセットの統計を示す．英語ベース CS と日本語ベース CS について，それぞれ「人工 CS」と「自然 CS」のコーパスを用意した．「人工 CS」コーパスは，合成音声の CS のみを使用するコーパスであり，「自然 CS」コーパスは「人工 CS」コーパスに自然音声の CS を追加したコーパスである．自然音声の CS は，学習セット用に 900 発話しかなかったため，音声の速度を変化させることでデータを拡張させる speed perturbation を用いた data augmentation（データ拡張）を適用した [30], [31]．90%，100%，110%の速度で speed perturbation を行い，3 倍の 2,700 発話に拡張した．「自然 CS」コーパスは，その自然音声 2,700 発話と合成音声の 50K を合わせて 52.7K のコーパスになっている．また，英語へ翻訳するタスクは英語ベース CS を用い，日本語へ翻訳するタスクは日本語ベース CS を用いた．

表 4 学習セットとテストセットの統計
Table 4 Statistics of the training and evaluation corpora.

			英語ベース		日本語ベース	
			人工 CS	自然 CS	人工 CS	自然 CS
学習セット	発話数	全体	50K	52.7K	100K	102.7K
		合成音声	50K	50K	100K	100K
		自然音声	-	900x3	-	900x3
	単語数	英語	521K	546K	264K	289K
		日本語	60K	69K	146K	155K
	時間		76	81	95	100
テストセット	発話数	全体	500	500	500	500
		合成音声	500	400	500	400
		自然音声	-	100	-	100
	単語数	英語	5.4K	5.4K	1.4K	2.7K
		日本語	0.6K	0.7K	2.5K	2.8K

音声特徴量については，Librosa library [32] を使用して，16 kHz のサンプリングレートの音声波形から 80 次元の対数メルスペクトログラムを抽出した．フレームの窓幅は 50 msec とし，シフト幅は 12.5 msec とした．作成したすべての文はトークン化され，日本語の文は形態素解析器である Mecab [28] を適用し，英語の文は WordPiece [33] を適用した．WordPiece は，未知語を効率良く減らすためのサブワード単位である．

4. 実験

まずは予備実験として CS テキストから単言語テキストへのテキスト翻訳の実験を行い，その後，本実験として CS 音声から単言語テキストへの音声翻訳の実験を行う．予備実験については，英語への翻訳のみを行い，本実験で英語への翻訳に加え，日本語への翻訳を実施する．評価指標は，単語誤り率（WER），文字誤り率（CER），BLEU スコア [34] を用いる．WER と CER は値が低いほどモデルの精度が良いことを示し，BLEU は値が高いほどモデルの精度が良いことを示す．

表 5 BERT 言語モデルと NMT の CS から英語へのテキスト翻訳精度
Table 5 Translation performance of BERT and NMT from CS text to English text.

テスト			CS の多言語テキストを単言語の参照訳で評価した場合	BERT 言語モデル 人工 CS	BERT 言語モデル 自然 CS
人工 CS	WER%↓	全体	27.43	11.14	12.14
		CS 部分	179.52	60.11	66.40
		CS 以外	0.11	5.17	5.20
	CER%↓	全体	19.79	12.01	12.62
		CS 部分	106.66	61.98	65.29
		CS 以外	0.07	1.03	1.09
	BLEU↑	全体	66.36	78.46	79.42
自然 CS	WER%↓	全体	35.23	20.61	19.53
		CS 部分	164.96	83.06	72.02
		CS 以外	3.44	8.67	8.77
	CER%↓	全体	24.80	19.31	18.56
		CS 部分	104.94	74.72	68.44
		CS 以外	2.54	3.69	3.92
	BLEU↑	全体	61.85	72.03	73.11

テスト			CS の多言語テキストを単言語の参照訳で評価した場合	NMT モデル 人工 CS	NMT モデル 自然 CS
人工 CS	WER%↓	全体	27.43	7.47	6.56
		CS 部分	179.52	34.31	35.11
		CS 以外	0.11	2.64	1.43
	CER%↓	全体	19.79	8.87	7.94
		CS 部分	106.66	32.92	33.35
		CS 以外	0.07	3.39	2.08
	BLEU↑	全体	66.36	86.54	88.11
自然 CS	WER%↓	全体	35.23	37.56	17.94
		CS 部分	164.96	60.69	57.88
		CS 以外	3.44	28.18	6.64
	CER%↓	全体	24.80	28.87	17.72
		CS 部分	104.94	50.59	49.96
		CS 以外	2.54	22.37	7.16
	BLEU↑	全体	61.85	57.46	77.86

4.1 コードスイッチングテキストから単言語テキスト

最終目標は音声からテキストへの翻訳の評価であるが，ASR の認識誤りの影響を調べるためにテキストからテキストへの翻訳の評価も行った．**表 5** に，BERT 言語モデルと NMT モデルの翻訳精度を示す．「CS の多言語テキストを単言語の参照訳で評価した場合」の列は，参照訳の単言語テキストに対する原文の CS テキストの WER，CER，および BLEU であり，参照訳の単言語テキストに対して原文の CS テキストがどれくらい遠いのか，近いのかを表す．この値を翻訳によって小さくできた（BLEU の場合は大きくできた）場合，参照訳に近づけたと考えられるため，参考値として翻訳結果の評価値と比較する．ただし，これらの比較は学習するデータサイズによっても変わってくるため，システムどうしの比較を主眼とする．また，原文の CS テキストにおける翻訳すべき他言語部分を「CS 部分」とし，それ以外を「CS 以外」として分けて評価した値も算出し，「全体」だけではなく「CS 部分」に対しても，どの程度正しく翻訳できたかを確認する．「CS の多言語テキストを単言語の参照訳で評価した場合」の列中の「CS 部分」では，WER および CER の値がいずれも 100%を超えたが，これは，参照訳の「CS 部分」を正解としたときに，原文の「CS 部分」に間違って挿入，置換，削除された文字の数が参照訳の「CS 部分」の長さよりも多かったため 100%を超えている．

BERT 言語モデルは，「CS の多言語テキストを単言語の参照訳で評価した場合」よりも WER と CER が低く，BLEU は高いので，翻訳したことで参照訳に近づいたと分かる．NMT モデルについても，学習していないコーパスによるテストを行った人工 CS モデルの自然 CS テスト以外では，「CS の多言語テキストを単言語の参照訳で評価した場合」の値より WER と CER が低くなっており，BLEU は高くなったため，参照訳に近づいたと分かる．また，BERT 言語モデルと NMT モデルの間で比較すると，「全体」と「CS 部分」の両方で NMT の方が優れる傾向だった．人工 CS モデルの自然 CS テストの「全体」における評価だけ，BERT 言語モデルの翻訳精度の方が優れていた．ただ，そこでも「CS 部分」で比較すると，NMT モデルが BERT 言語モデルよりも良い精度だった．そこから，「CS 以外」の性能によって，BERT 言語モデルの「全体」の性能が良かったと分かる．BERT 言語モデルは「CS 以外」のところが入力テキストのコピーとなるため，入力テキストにエラーがなければ精度が良くなると考えられる．次の章で，ASR による認識誤りが含まれた場合はどうなるかを検証する．

4.2 コードスイッチング音声から単言語テキスト

表 6 に，Cascade ASR+BERT および Cascade ASR+NMT を用いて，コードスイッチング音声から音声翻訳した結果を示す．まず，「CS の多言語テキストを単言語の参照訳で評価した場合」の値と比較すると，人工 CS モデルの自然 CS テストは学習しなかったテストのため難しかったが，それ以外の「全体」と「CS 部分」の WER，CER，BLEU では改善されている．また，Cascade ASR+BERT と Cascade ASR+NMT の間で比較をすると，テキストの機械翻訳タスクでは，BERT 言語モデルが，NMT の精度を上回るケースがあったが，音声翻訳では，Cascade

表 6 Cascade ASR+BERT と Cascade ASR+NMT の CS から英語への音声翻訳精度の比較
Table 6 Comparison between cascade ASR+BERT and cascade ASR+NMT from CS speech to English text.

テスト			CS の多言語テキストを単言語の参照訳で評価した場合	モデル			
				人工 CS		自然 CS	
				Cascade		Cascade	
				ASR+BERT	ASR+NMT	ASR+BERT	ASR+NMT
人工 CS	WER%↓	全体	27.43	16.16	10.76	15.80	9.20
		CS 部分	179.52	66.76	36.08	69.15	34.13
		CS 以外	0.11	8.74	6.16	8.52	5.37
	CER%↓	全体	19.79	15.83	9.73	14.97	8.92
		CS 部分	106.66	65.74	33.39	67.11	31.86
		CS 以外	0.07	4.76	4.22	3.65	3.31
	BLEU↑	全体	66.36	71.48	80.79	73.17	82.08
自然 CS	WER%↓	全体	35.23	46.42	41.94	25.61	22.34
		CS 部分	164.96	89.67	56.50	74.23	57.15
		CS 以外	3.44	32.61	32.73	14.76	13.17
	CER%↓	全体	24.80	36.95	31.56	22.46	20.23
		CS 部分	104.94	77.12	49.97	69.91	49.26
		CS 以外	2.54	24.36	24.35	8.85	10.28
	BLEU↑	全体	61.85	47.89	54.43	64.24	69.82

表 7 Cascade と Direct アプローチの CS から英語への音声翻訳精度の比較
Table 7 Comparison of translation performance from CS speech to English text between cascade and direct approaches.

テスト			CS の多言語テキストを単言語の参照訳で評価した場合	モデル					
				人工 CS			自然 CS		
				Cascade ASR+NMT	Direct single-task ST	Direct multi-task ST	Cascade ASR+NMT	Direct single-task ST	Direct multi-task ST
人工 CS	WER%↓	全体	27.43	10.76	11.13	10.15	9.20	13.04	8.71
		CS 部分	179.52	36.08	40.96	34.40	34.13	37.59	33.16
		CS 以外	0.11	6.16	5.73	5.73	5.37	8.60	4.95
	CER%↓	全体	19.79	9.73	9.69	8.85	8.92	10.83	8.50
		CS 部分	106.66	33.39	34.52	31.14	31.86	34.43	30.60
		CS 以外	0.07	4.22	3.88	3.71	3.31	5.16	2.99
	BLEU↑	全体	66.36	80.79	80.82	81.99	82.08	78.67	82.87
自然 CS	WER%↓	全体	35.23	41.94	38.87	34.56	22.34	29.63	23.21
		CS 部分	164.96	56.50	60.69	55.85	57.15	61.20	56.79
		CS 以外	3.44	32.73	29.77	26.16	13.17	19.79	14.78
	CER%↓	全体	24.80	31.56	29.42	28.26	20.23	24.98	21.55
		CS 部分	104.94	49.97	50.59	48.91	49.26	53.43	49.05
		CS 以外	2.54	24.35	23.08	21.66	10.28	16.00	12.54
	BLEU↑	全体	61.85	54.43	56.72	58.28	69.82	64.38	68.46

ASR+NMT がすべてのケースで Cascade ASR+BERT を上回った．「CS 部分」の精度をみても，すべてのケースで，Cascade ASR+NMT モデルが Cascade ASR+BERT よりも精度が良かった．Cascade ASR+BERT の精度が悪化したのは，ASR の認識誤りが [MASK] トークンの数を増やしたため，BERT にとって，タスクを解決するのが困難になったことが原因と考えられる．したがって，以降は Cascade ASR+NMT モデルを Direct アプローチのシングルタスク音声翻訳（Direct single-task ST）および Direct アプローチのマルチタスク音声翻訳（Direct multi-task ST）と比較し，CS 音声から英語と日本語への音声翻訳タスクを行う．

表 7 に，日英 CS から英語への音声翻訳における，ASR+NMT の Cascade アプローチと，シングルタスクおよびマルチタスク学習を使用した Direct アプローチの

表 8 自然 CS テストにおける CS から英語への音声翻訳の出力例
Table 8 Output examples of speech translation from CS speech to English text on natural CS.

	原文	oh , no . i don ' t have any change so , どう し たら いいん でしょ う ?
	参照訳	oh , no . i don ' t have any change so , what should i do ?
出力結果	Cascade ASR+NMT	oh , no . i don ' t have any change so , what should i do ?
	Direct single-task ST	oh , no . i don ' t have any change so , what should i do ?
	Direct multi-task ST	oh , no . i don ' t have any change so , what should i do ?
	原文	if you want to watch tv , you should 宿題 は 終え た ほう が いい です よ .
	参照訳	if you want to watch tv , you should finish your homework first .
出力結果	Cascade ASR+NMT	if you want to watch tv , you should go to you homework .
	Direct single-task ST	if you want to watch tv , you should be able to talk .
	Direct multi-task ST	if you want to watch tv , you should have any homework .
	原文	when students from four year universities and junior colleges are put together there are probably about 二百 五十 万 人 ぐらい の 学生 が い る はず で す .
	参照訳	when students from four year universities and junior colleges are put together , there are probably about two point five million students .
出力結果	Cascade ASR+NMT	when i ' m sorry that the students is no good , i ' m going to have a student .
	Direct single-task ST	when i ' m students for your your university and junior college are put together , there are probably about you .
	Direct multi-task ST	when you ' re students from four year your university and junior college are put together , there are probably about two hundred million yen .
	原文	how would you like to pay 現金 です か , カード です か ?
	参照訳	how would you like to pay , cash or charge ?
出力結果	Cascade ASR+NMT	how would you like to pay , cash or charge ?
	Direct single-task ST	how would you like to pay , cash or charge ?
	Direct multi-task ST	how would you like to pay , cash or charge ?
	原文	i ' ll do my best to find your baggage but first i ' d like you to fill in this 手 荷物 事故 報告 書 .
	参照訳	i ' ll do my best to find your baggage , but first i ' d like you to fill in this property irregularity report .
出力結果	Cascade ASR+NMT	i ' ll do my best to find your baggage but first i ' d like you to fill in this tennis .
	Direct single-task ST	i ' ll do my best to fly your baggage , but first i ' d like you to fill in this morning , so i ' ll do you ' ll be in the kanto person .
	Direct multi-task ST	i ' ll do my best to find your baggage , but first i like you to fill in this form to yourself in japan .

精度を比較した結果を示す．まず，「CS の多言語テキストを単言語の参照訳で評価した場合」の値と比較すると，学習していないコーパスのため難しかった人工 CS モデルの自然 CS テスト以外では，ほとんどすべてのモデルで「全体」と「CS 部分」の WER および CER，そして BLEU の値が改善された．Direct シングルタスク音声翻訳の自然 CS モデルの自然 CS テストだけ，CER が「CS の多言語テキストを単言語の参照訳で評価した場合」の値より増えているが，WER では改善できているのが分かる．Cascade ASR+NMT，Direct シングルタスク音声翻訳，Direct マルチタスク音声翻訳の間で比較をすると，Direct マルチタスク音声翻訳が最も良い精度を示す傾向にあった．自然 CS モデルの自然 CS テストだけ，「全体」の WER と CER で Direct マルチタスク音声翻訳よりも Cascade ASR+NMT の方が精度は良かったが，そこでも「CS 部分」になると Direct マルチタスク音声翻訳の方が精度は良かった．

また，**表 8** に示した，自然 CS テストでの CS から英語への音声翻訳の出力例を見ると，「どうしたらいいんでしょう？」というような翻訳箇所が 1 つの文として完結するものについてはすべてのモデルがうまく翻訳できたが，「you should 宿題 は 終え た ほう が いい です よ」や「there are probably about 二百 五十 万 人 ぐらい の 学生 が い る はず で す」のような翻訳箇所が他言語のフレーズ内に挿入されたものについては難しい傾向があった．また，短い単語や「現金 です か , カード です か ?」のような単語の組合せで成り立つようなフレーズは比較的簡単に翻訳できるが，単語でも「手 荷物 事故 報告 書」のような複数のトークンにまたがるような複合語については，難しいと分かった．難しいケースをみると，「you should 宿題 は 終え た ほう が いい です よ」では，Cascade ASR+NMT や Direct マ

表 9 Cascade と Direct アプローチの CS から日本語への音声翻訳精度の比較
Table 9 Comparison of translation performance from CS speech to Japanese text between cascade and direct approaches.

テスト			CS の多言語テキストを単言語の参照訳で評価した場合	モデル					
				人工 CS			自然 CS		
				Cascade ASR+NMT	Direct single-task ST	Direct multi-task ST	Cascade ASR+NMT	Direct single-task ST	Direct multi-task ST
人工 CS	WER%↓	全体	17.52	13.11	19.96	13.88	12.94	16.25	12.05
		CS 部分	106.42	32.59	42.25	36.40	33.48	36.28	31.96
		CS 以外	0.61	9.19	15.51	9.44	8.90	12.29	8.18
	CER%↓	全体	35.23	12.10	18.93	13.10	12.23	15.38	11.68
		CS 部分	187.66	27.32	36.05	30.85	28.59	31.10	26.92
		CS 以外	0.53	9.05	15.55	9.40	8.86	12.34	8.41
	BLEU↑	全体	74.14	77.42	68.14	77.35	77.77	72.31	79.36
自然 CS	WER%↓	全体	28.67	41.80	44.81	27.51	25.77	30.51	25.52
		CS 部分	104.56	61.46	68.67	61.18	60.40	63.44	61.67
		CS 以外	3.66	25.78	33.85	16.31	14.41	19.04	13.46
	CER%↓	全体	54.29	33.97	46.70	25.69	23.85	28.04	22.98
		CS 部分	202.69	56.87	66.90	54.98	53.88	57.72	53.43
		CS 以外	1.90	20.85	35.65	15.95	14.29	18.91	13.18
	BLEU↑	全体	64.64	59.02	52.81	64.47	65.75	60.32	67.24

ルチタスク音声翻訳はキーワードとなる「homework」を予測できたが，Direct シングルタスク音声翻訳は予測できておらず「talk」という異なる意味の動詞の予測を行った．他の例でも，Direct シングルタスク音声翻訳は音声翻訳が一番難しいという印象を受けた．これは，Direct シングルタスク音声翻訳が CS 中の英語音声と日本語音声を同じ英語テキストへ書き起こすので，英語音声と日本語音声の区別ができずに難しくなったと考えられる．Cascade ASR+NMT は「手 荷物 事故 報告 書」のところを「tennis」と訳した．ここは ASR が誤って「テニス」と認識しており，エラーが伝播したと考えられる．このように，Cascade ASR+NMT は ASR からのエラー伝播によるダメージを受けるので，それが Direct マルチタスク音声翻訳との差につながったと考えられる．Direct マルチタスク音声翻訳については，Direct シングルタスク音声翻訳と同じく CS 中の英語音声と日本語音声から同じ英語テキストへ書き起こしを行うが，マルチタスクとして CS テキストへの書き起こしも行ったため，CS 中の英語音声と日本語音声の間でうまく区別がつけられ，ASR によるエラー伝播も少ないので他システムより優れた傾向にあると考えられる．

また，**表 9** に，日英 CS から日本語への音声翻訳精度を示す．まず，「全体」と「CS 部分」の WER および CER については，ほとんどが「CS の多言語テキストを単言語の参照訳で評価した場合」の値よりも改善された．しかし，Cascade ASR+NMT と Direct シングルタスク音声翻訳は，人工 CS モデルの自然 CS テストの「全体」で，「CS の多言語テキストを単言語の参照訳で評価した場合」の値よりも WER や CER が高くなり，Direct シングルタスク音声翻訳は自然 CS モデルの自然 CS テストでも「CS の多言語テキストを単言語の参照訳で評価した場合」の値よりも WER が高くなった．人工 CS モデルの自然 CS テストについては，学習しなかった自然 CS によるテストが難しかったと考えられた．自然 CS モデルの自然 CS テストでも精度が下がっている Direct シングルタスク音声翻訳については，CS 中の英語音声と日本語音声を区別せずに，同じ日本語へ書き起こすのでモデルの精度を低下させ，エラーが増えたと考えられる．BLEU についても，Direct シングルタスク音声翻訳は，人工 CS モデルの自然 CS テストや自然 CS モデルの自然 CS テストで，「CS の多言語テキストを単言語の参照訳で評価した場合」の値よりも下回っており，CS 中の英語音声と日本語音声を区別せずに，同じ日本語へ書き起こすので，モデルの精度を低下させ，BLEU の値も低くなったと考えられる．Cascade ASR+NMT，Direct シングルタスク音声翻訳，Direct マルチタスク音声翻訳の間で比較をすると，英語への翻訳と同様に，Direct シングルタスク音声翻訳の精度が最も悪く，ASR のエラー伝播が少ない場合は Cascade ASR+NMT が Direct マルチタスク音声翻訳より優れる場合もあるが，Direct マルチタスク音声翻訳が最も良い精度を示す傾向にある．ただ，人工 CS モデルの人工 CS テストでは，「全体」と「CS 部分」において，Cascade ASR+NMT の方が良い傾向にあった．日本語ベース CS は名詞のみの 1 単語の CS が含まれており，ASR にとって簡単なタスクになり，ASR のエラー伝播が少なかったと考えられる．また，自然 CS モデルにおける

表 10 自然 CS テストにおける CS から日本語への音声翻訳の出力例
Table 10 Output examples of speech translation from CS speech to Japanese text on natural CS.

原文		七月 二 十 日 二 回 目 公演 の エー 指定 席 券 を please give me two pieces .
参照訳		七月 二 十 日 二 回 目 公演 の エー 指定 席 券 を 二 枚 ください .
出力結果	Cascade ASR+NMT	七月 二 十 日 二 回 目 の エー 指定 席 券 を 二 枚 ください .
	Direct single-task ST	七月 二 十 日 二 回 公園 の エー 席 券 を 二 枚 ください .
	Direct multi-task ST	七月 二 十 日 二 回 目 公演 の エー 指定 席 券 を 二 枚 ください .
原文		ゆで 卵 一 個 と orange juice and bread please .
参照訳		ゆで 卵 一 個 と オレンジ ジュース と パン を お願い し ます .
出力結果	Cascade ASR+NMT	ゆで 卵 一 個 と オレンジ ジュース と パン を ください .
	Direct single-task ST	ゆで 卵 一 個 と オレンジ ジュース 二つ お願い し ます .
	Direct multi-task ST	ゆで 卵 一 個 と オレンジ ジュース と パン を お願い し ます .
原文		サイズ の 番号 は わかり ませ ん が the size of the foot is 24 points five centimeters .
参照訳		サイズ の 番号 は わかり ませ ん が 足 の 大き さ は 二 十 四 点 五 センチ です .
出力結果	Cascade ASR+NMT	サイズ の 番号 は わかり ませ ん が 二 十 五 点 六 六 円 です .
	Direct single-task ST	サイズ の 番号 は わかり ませ ん が 五 十 四 センチ の 名前 は .
	Direct multi-task ST	サイズ の 番号 は わかり ませ ん が 残念 ながら 二 十 四 点 五 十 センチ です .

自然 CS テストでの WER は，「全体」は Direct マルチタスク音声翻訳の方が良いものの，「CS 部分」では Cascade ASR+NMT の方が精度は良かった．しかし，CER では「全体」も「CS 部分」もどちらも Direct マルチタスク音声翻訳の方が良く，トークナイズの違いによるわずかな差だと分かる．そして，自然 CS モデルの人工 CS テスト，および人工 CS モデルの自然 CS テストでは「全体」と「CS 部分」の両方で，Direct マルチタスク音声翻訳が最も優れた性能だった．

表 10 の，日本語への音声翻訳の出力例を見ると，英語への出力例と同じく，「please give me two pieces」のような翻訳箇所が 1 つの文として完結するものや「orange juice and bread please」のような単語の組合せで成り立つようなフレーズはうまく翻訳できた．「the size of the foot is 24 points five centimeters」の「24 points five centimeters」は複数のトークンにまたがる名詞であったが，Direct マルチタスク音声翻訳だけが「二 十 四 点 五 センチ」のところを「二 十 四 点 五 十 センチ」と翻訳し，比較的近いものを当てることができた．

以上の考察をまとめると，翻訳箇所が 1 つの文として完結するものや，単語や単語の組合せで成り立つようなフレーズの翻訳は難しくないが，単語でも複数のトークンにまたがるような複合語や，翻訳箇所が他言語のフレーズ内に挿入されたものは難しいと分かった．Cascade ASR+NMT，Direct シングルタスク音声翻訳，Direct マルチタスク音声翻訳を比較すると，Direct シングルタスク音声翻訳の精度が最も悪く，Cascade ASR+NMT が Direct マルチタスク音声翻訳より優れることがあるものの，Direct マルチタスク音声翻訳が最も良い精度を示す傾向にある．Direct シングルタスク音声翻訳は，CS 中の英語音声と日本語音声から同じ単言語テキストへ書き起こすので日本語音声と英語音声の区別ができず，モデルの精度の低下を招いた．Cascade ASR+NMT は ASR のエラー伝播がある．一方，Direct マルチタスク音声翻訳は，直接音声から予測するため ASR のエラー伝播が少なく，また Direct シングルタスク音声翻訳とは違って，CS 音声から CS テキストへの出力も一緒に学習することで，CS 中の英語音声と日本語音声の区別ができるようになるため，最も良い精度を示す傾向にある．

5. おわりに

日英 CS の音声翻訳を開発するため，NMT と BERT による 2 つの Cascade アプローチ，およびシングルタスク学習とマルチタスク学習による 2 つの Direct アプローチを調査した．実験の結果，Cascade ASR+BERT は，ASR の認識誤りによって，予測するものが増えると同時に予測のために必要な情報が減るため，CS の音声翻訳は難しくなることが分かった．また，Direct シングルタスク音声翻訳は，CS 中の英語音声と日本語音声から同じ単言語テキストへ書き起こすので，英語音声と日本語音声の区別ができず，モデルの精度の低下を招いた．また，Cascade ASR+NMT は ASR のエラー伝播の影響を受けることが分かった．一方，Direct マルチタスク学習は，直接予測することで ASR のエラー伝播が少なく，また Direct シングルタスク学習と同じく CS 中の英語音声と日本語音声から単言語テキストへの書き起こしを予測するにもかかわらず，同時に CS 音声から CS テキストを予測することで CS 中の英語音声と日本語音声を区別できるため，Direct マルチタスク学習が最も良い精度を示す傾向にあった．今後は，翻訳結果の主観評価や，非母語話者レベルの CS を用いた実験を行いたい．

謝辞 本研究は科研費 JP17H06101，JP17K00237 の助成を受けております．

参考文献

[1] 厚生労働省：平成 29 年度人口動態統計（確定数），入手先 ⟨https://www.mhlw.go.jp/toukei/saikin/hw/jinkou/kakutei17/xls/29toukei.xls⟩ (2017).

[2] 日本政府観光局（JNTO）：2018 年 訪日外客数（総数），入手先 ⟨https://www.jnto.go.jp/jpn/statistics/since2003_visitor_arrivals.pdf⟩ (2019).

[3] Fotos, S.S.: Japanese-English Code Switching in Bilingual Children, *JALT Journal*, Vol.12, No.1, pp.75–98 (1990).

[4] Poplack, S.: Sometimes I'll start a sentence in Spanish y termino en español: Toward a typology of code-switching, *The Bilingualism Reader*, Vol.18, No.2, pp.221–256 (2000).

[5] Bautista, M.L.S.: Tagalog-english code switching as a mode of discourse, *Asia Pacific Education Review*, Vol.5, No.2, pp.226–233 (online), DOI: 10.1007/BF03024960 (2004).

[6] White, C.M., Khudanpur, S. and Baker, J.K.: An investigation of acoustic models for multilingual code switching, *Proc. INTERSPEECH*, pp.2691–2694 (2008).

[7] Vu, N.T., Lyu, D.-C., Weiner, J., Telaar, D., Schlippe, T., Blaicher, F., Chng, E.-S., Schultz, T. and Li, H.: A first speech recognition system for Mandarin-English code-switch conversational speech, *Proc. ICASSP*, pp.4889–4892 (2012).

[8] Yilmaz, E., den Heuvel, H. and van Leeuwen, D.: Investigating Bilingual Deep Neural Networks for Automatic Recognition of Code-switching Frisian Speech, *Proc. SLTU*, Vol.81, pp.159–166 (2016).

[9] Hernandez, D.J., Denton, N.A. and Macartney, S.E.: Children in Immigrant Families: Looking to America's Future, *Social Policy Report*, Vol.22, No.3, Society for Research in Child Development (2008).

[10] Sinha, R.M.K. and Thakur, A.: Machine translation of bi-lingual hindi-english (hinglish) text, *Proc. MT Summit X*, pp.149–156 (2005).

[11] Johnson, M., Schuster, M., Le, Q.V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M. and Dean, J.: Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation, *Proc. TACL*, Vol.5, pp.339–351 (2017).

[12] Menacer, M.A., Langlois, D., Jouvet, D., Fohr, D., Mella, O. and Smaïli, K.: Machine Translation on a parallel Code-Switched Corpus, *Proc. Canadian AI*, pp.426–432 (2019).

[13] Devlin, J., Chang, M., Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, *Proc. NAACL-HLT*, pp.4171–4186 (2019).

[14] Nakayama, S., Kano, T., Tjandra, A., Sakti, S. and Nakamura, S.: Recognition and Translation of Code-switching Speech Utterances, *Proc. O-COCOSDA* (2019).

[15] Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P. and Bengio, Y.: End-to-end attention-based large vocabulary speech recognition, *Proc. ICASSP*, pp.4945–4949 (2016).

[16] Xu, B., Wang, N., Chen, T. and Li, M.: Empirical evaluation of rectified activations in convolutional network, *Deep Learning Workshop*, *ICML*, pp.1–5 (2015).

[17] Luong, T., Pham, H. and Manning, C.D.: Effective Approaches to Attention-based Neural Machine Translation, *Proc. EMNLP*, pp.1412–1421 (2015).

[18] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I.: Attention is all you need, *Proc. NIPS*, pp.5998–6008 (2017).

[19] Tjong Kim Sang, E.F. and De Meulder, F.: Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition, *Proc. NAACL-HLT*, pp.142–147 (2003).

[20] Rajpurkar, P., Zhang, J., Lopyrev, K. and Liang, P.: SQuAD: 100,000+ Questions for Machine Comprehension of Text, *Proc. EMNLP*, pp.2383–2392 (2016).

[21] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A. and Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank, *Proc. EMNLP*, pp.1631–1642 (2013).

[22] Ghazvininejad, M., Levy, O., Liu, Y. and Zettlemoyer, L.: Constant-time machine translation with conditional masked language models, arXiv preprint arXiv: 1904.09324 (2019).

[23] Weiss, R.J., Chorowski, J., Jaitly, N., Wu, Y. and Chen, Z.: Sequence-to-Sequence Models Can Directly Translate Foreign Speech, *Proc. INTERSPEECH* (2017).

[24] Tan, Z., Fan, X., Zhu, H. and Lin, E.: Addressing Accent Mismatch In Mandarin-English Code-Switching Speech Recognition, *Proc. ICASSP*, pp.8259–8263 (2020).

[25] Takezawa, T., Kikui, G., Mizushima, M. and Sumita, E.: Multilingual Spoken Language Corpus Development for Communication Research, *Proc. ACLCLP*, Vol.12, No.3, pp.303–324 (2007).

[26] Tjandra, A., Sakti, S. and Nakamura, S.: Listening while Speaking: Speech Chain by Deep Learning, *Proc. ASRU*, pp.301–308 (2017).

[27] Schmid, H.: TreeTagger – A language independent part-of-speech tagger, available from ⟨http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/⟩ (1994).

[28] Kudo, T.: Mecab: Yet another part-of-speech and morphological analyzer, available from ⟨http://taku910.github.io/mecab⟩.

[29] Durette, P.N.: gTTS – Google Text-to-Speech, available from ⟨https://pypi.org/project/gTTS/⟩.

[30] Ko, T., Peddinti, V., Povey, D. and Khudanpur, S.: Audio augmentation for speech recognition, *Proc. INTERSPEECH* (2015).

[31] Ko, T., Peddinti, V., Povey, D., Seltzer, M.L. and Khudanpur, S.: A study on data augmentation of reverberant speech for robust speech recognition, *Proc. ICASSP*, pp.5220–5224, IEEE (2017).

[32] McFee, B., McVicar, M., Nieto, O., Balke, S., Thome, C., Liang, D., Battenberg, E., Moore, J., Bittner, R., Yamamoto, R., et al.: librosa 0.5.0, available from ⟨https://librosa.github.io/librosa/0.5.0/index.html⟩ (2017).

[33] Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al.: Google's neural machine translation system: Bridging the gap between human and machine translation, arXiv preprint arXiv:1609.08144 (2016).

[34] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J.: BLEU: A method for automatic evaluation of machine translation, *Proc. ACL*, pp.311–318 (2002).

中山 佐保子 （学生会員）

2012 年早稲田大学国際教養学部国際教養学科卒業．2019 年奈良先端科学技術大学院大学博士課程前期修了．同年より，同大学院博士課程後期在学，理化学研究所革新知能統合研究センター目的指向基盤技術研究グループ観光情報解析チーム研究アシスタント．コードスイッチングの音声言語処理に関する研究に従事．

サクティ サクリアニ

1999 年インドネシアバンドン工科大学情報学部卒業．2002 年ドイツウルム大学修士課程修了．2003 年音声言語コミュニケーション研究所研究員．2006 年（独）情報通信研究機構専門研究員．2008 年ドイツウルム大学博士課程修了．2009 年インドネシア大学コンピューターサイエンス学部客員教授．2011 年奈良先端科学技術大学院大学助教．2015 年フランス INRIA Paris-Rocquencourt の客員研究員．現在，奈良先端科学技術大学院大学准教授，理化学研究所革新知能統合研究センター目的指向基盤技術研究グループ観光情報解析チーム研究員．JNS，SFN，ASJ，ISCA，IEICE，IEEE 各会員．

中村 哲 （正会員）

1981 年京都工芸繊維大学工芸学部電子工学科卒業．京都大学博士（工学）．シャープ株式会社．1994 年奈良先端科学技術大学院大学助教授，1996 年米国 Rutgers 大学客員教授（文科省在外研究員），2000 年 ATR 音声言語コミュニケーション研究所室長，2005 年所長，2006 年（独）情報通信研究機構音声コミュニケーション研究グループリーダ，2010 年研究センター長，けいはんな研究所長等を経て，2011 年奈良先端科学技術大学院大学情報科学研究科教授，2017 年データ駆動型サイエンス創造センター長，2017 年理化学研究所革新知能統合研究センター観光情報解析チームリーダー．2003 年からカールスルーエ大学客員教授．音声翻訳，音声対話等の音声言語情報処理，自然言語処理の研究に従事．電子情報通信学会論文賞，AAMT 長尾賞，日本音響学会技術開発賞，人工知能学会研究会優秀賞，情報処理学会喜安記念業績賞，総務臣表彰，文部科学大臣表彰，Antonio Zampoli 賞，ドコモモバイルサイエンス賞，IBM Faculty Award，Google AI Focused Research Award 等受賞．ISCA 理事，IEEE SLTC 委員を歴任．ATR フェロー，IEEE フェロー，ISCA フェロー，本会フェロー．