マルチリンガルマシーンスピーチチェーンを用いたゼロショットコードスイッチングの音声認識と音声合成*

☆中山佐保子[1,2], チャンドラ アンドロス[1,†], サクティ サクリアニ[1,2], 中村哲[1,2]
([1]NAIST, [2]RIKEN AIP)

1 はじめに

会話の中で言語が切り替わる現象はコードスイッチング（CS）と呼ばれる。これまでの CS 音声認識や CS 音声合成は CS データによる教師あり学習が多かった [1, 2]。しかし，CS の音声とテキストのペアデータを大規模に集めるのは難しい。そのため，われわれはマシーンスピーチチェーンを利用して半教師あり学習の CS 音声認識と CS 音声合成の開発を進めている。マシーンスピーチチェーンは音声認識（ASR）と音声合成（TTS）をループ結合して互いに学習させる仕組みであり，入力に対する正解データがないラベルなしデータに対しても，互いのモデルを用いて推測させることで学習可能になる。本研究では，マシーンスピーチチェーンに言語 ID 分散表現と言語識別（LID）の学習を組み込み，言語 ID 情報を用いることで複数の CS および学習に含まれない未知の CS（ゼロショット CS と呼ぶ）に対してもロバストな ASR と TTS を実現する。

2 マルチリンガルスピーチチェーン

マシーンスピーチチェーンとは，人間のコミュニケーションメカニズムのスピーチチェーン [3] から着想した手法で，ASR と TTS をループ結合して相互に学習させる [4]。教師あり学習と教師なし学習の 2 つの学習フェーズを持ち，教師あり学習はラベルありデータで ASR と TTS をそれぞれ学習させ，教師なし学習は，ASR と TTS の互いのモデルを用いてラベルなしデータを推測させることで学習を行う。CS スピーチチェーンはこの仕組みを利用し，教師あり学習にモノリンガルデータを使い，教師なし学習に CS データを使うことで，ラベルなしの CS データを用いた学習を実現した [5]。本研究のマルチリンガルスピーチチェーンは，CS スピーチチェーンに言語 ID 情報を組み込む。ASR は，2 つのソフトマックス層を用いたマルチタスクで，入力音声を認識文字列と言語 ID に変換し，TTS は，ASR から受け取った言語 ID 情報を分散表現に変換して文字分散表現と結合して学習する。学習プロセスは，Fig. 1 に示す。まずは ASR+LID，TTS，そして多数話者情報を処理する話者認識モデル（SPKREC）を，それぞれモノリンガルデータを用いて教師あり学習を行う。次に，それらのモデルを用いてモノリンガルの教師あり学習を続けながら CS データの教師なし学習を実行する。TTS が文字と言語 ID の 2 つのテキストデータ $[y^{CSChr}, y^{CSLng}]$ を入力として受け取り，CS 音声 $\hat{x}^{CS}$ を生成する。生成した音声を，ASR が認識し，文字と言語 ID の 2 つのテキストデータ $[\hat{y}^{CSChr}, \hat{y}^{CSLng}]$ を生成する。そして，$L_{ASR}^{CSChr}(\hat{y}^{CSChr}, y^{CSChr})$ と $L_{ASR}^{CSLng}(\hat{y}^{CSLng}, y^{CSLng})$ の損失関数の合計を計算する。次に，ASR が CS 音声 x^{CS} を認識して文字系列 $\hat{y}^{CSChr}$ と言語 ID 情報 $\hat{y}^{CSLng}$ を生成する。生成したテキストから，TTS が CS 音声 $\hat{x}^{CS}$ を生成し，損失関数 $L_{TTS}^{CS}(\hat{x}^{CS}, x^{CS})$ を計算する。最後に，教師ありのモノリンガル損失と教師なしの CS 損失を，以下のように α と β のハイパーパラメータでバランスを調整しながら一つの損失関数に合計し，ASR と TTS のパラメータを更新する。

$$L = \alpha * ((L_{ASR}^{MonoChr} + L_{ASR}^{MonoLng}) + L_{TTS}^{Mono}) + \beta * ((L_{ASR}^{CSChr} + L_{ASR}^{CSLng}) + L_{TTS}^{CS}) \quad (1)$$

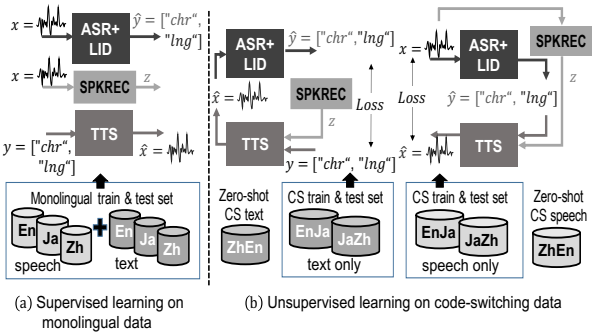

Fig. 1 Multilingual machine speech chain.

3 実験条件

本実験では，BTEC[6] の対訳コーパスを Google TTS で合成して使用する。教師あり学習には日本語，英語，中国語の発話を 25K ずつ用いたデータ（Ja+En+Zh）を使い，教師なし学習には，読点以降のフレーズを翻訳して作成した CS 発話（EnJaCS, JaZhCS，ZhEnCS）を 10K ずつ用いた。音響特徴量は ASR に 80 次元の対数メルスペクトログラム，TTS に 80 次元の対数メルスペクトログラムと 1024 次元

*Multilingual machine speech chain for zero-shot code-switching ASR and TTS. by Sahoko Nakayama[1,2], Andros Tjandra[1,†], Sakriani Sakti[1,2], Satoshi Nakamura[1,2], ([1]NAIST, [2]RIKEN AIP)

†The work was done when he was at NAIST, he is currently at Facebook AI, USA.

Table 1 Comparison between ASR baselines with/without LID in CER%.

Train:Ja+En+Zh	Ja	En	Zh
ASR without LID [chr]	8.8	9.1	5.8
ASR with LID [chr,lng]	8.9	8.5	5.1

Table 2 CER% of proposed ASR model.

	EnJaCS	JaZhCS	ZhEnCS
[Baseline] Supervised: labeled mono			
Ja+En+Zh	**14.1**	**16.9**	**16.0**
[Proposed] Semisupervised: unlabeled CS			
+ EnJaCS+JaZhCS	11.6	8.3	**10.5**
+ EnJaCS+ZhEnCS	11.2	**9.2**	9.7
+ ZhEnCS+JaZhCS	**11.9**	10.4	11.3
[Topline] Supervised: labeled CS			
+ EnJaCS+JaZhCS	8.9	6.7	**8.1**
+ EnJaCS+ZhEnCS	10.8	**7.3**	8.1
+ ZhEnCS+JaZhCS	**10.3**	7.7	8.0

の対数振幅スペクトログラムを使用する。フレームの窓幅は 50msec とし，シフト幅は 12.5msec とした。また，テキストは全ての文字を小文字のアルファベットに変換して使用した。ASR には注意機構を用いたエンコーダデコーダモデル [7] を用い，3 層の双方向 LSTM エンコーダと 1 層の LSTM デコーダ，多層パーセプトロンを用いた注意機構から構成され，活性化関数には LeakyReLU（$l = 1e - 2$）を用いた。TTS は Tacotron[8] をベースとし，活性化関数を LeakyReLU（$l = 1e - 2$），CBHG の畳み込みフィルタを 8 セット，デコーダの GRU を 2 層の LSTM に変更して用いた。

4 実験結果

4.1 音声認識結果

まず，Ja+En+Zh で教師あり学習したベースラインシステムを用いて LID の性能を確認する。Table 1 の結果は，LID が ASR の性能改善に役立つ可能性があることを示す。

次に，ゼロショット CS を含む複数の CS での提案手法の性能を調査する。提案手法の CS スピーチチェーンは Ja+En+Zh で教師あり学習をした後に EnJaCS+JaZhCS，EnJaCS+ZhEnCS，ZhEnCS+JaZhCS で教師なし学習をしたモデルである。結果は，Table 2 に示す通り，提案手法はラベルなしの CS データを用いたにも関わらず，ベースラインと比較して全ての CS テストケースで性能を改善した。太字で示したゼロショット CS も性能を改善したことが分かる。ラベルありの CS データを用いて教師あり学習を行なったトップラインのモデルと比較しても，近い性能を発揮した。

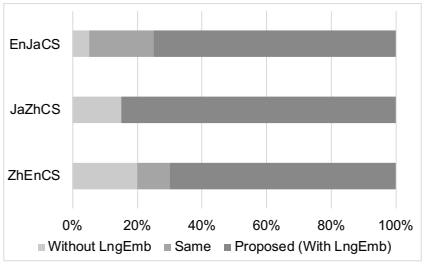

Fig. 2 AB preference subjective evaluation.

4.2 音声合成結果

言語 ID 分散表現を用いた TTS の性能を確認するため，プリファレンス AB テストを行なった。各 CS 言語ペアの合成音声に対して 10 人のバイリンガル話者に参加してもらった。結果は，Fig. 2 に示す通り，言語 ID 分散表現（LngEmb）が性能改善に役立つことを示した。

5 おわりに

マシーンスピーチチェーンに言語 ID 情報を組み込んだマルチリンガルスピーチチェーンを紹介した。提案手法は，ゼロショット CS を含め複数の CS の認識性能の改善を確認できた。

謝辞 本研究は科研費 [JP17H06101] の助成を受けております。

参考文献

[1] Li *et al.*, “Towards code-switching ASR for end-to-end CTC models,” Proc. of ICASSP, 2019.
[2] Sitaram *et al.*, “Speech synthesis of code-mixed text,” Proc. of LREC, 2016.
[3] Denes *et al.*, “The Speech Chain: The Physics And Biology Of Spoken Language,” Proc. of ICASSP, 1993.
[4] Tjandra *et al.*, “Machine Speech Chain,” IEEE/ACM TASLP, 2020.
[5] Nakayama *et al.*, “Speech Chain for Semi-Supervised Learning of Japanese-English Code-Switching ASR and TTS,” Proc. of SLT, 2018.
[6] Takezawa *et al.*, “Multilingual Spoken Language Corpus Development for Communication Research,” IJCLCLP, 2007.
[7] Chan *et al.*, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” Proc. of ICASSP, 2016.
[8] Wang *et al.*, “Tacotron: Towards end-to-end speech synthesis,” Proc. of INTERSPEECH, 2017.