End-to-End 日本語方言音声認識と 日本語方言識別のための結合モデリング手法の比較*

☆今泉遼 (都立大), 増村亮 (NTT), 塩田さやか, △貴家仁志 (都立大)

1 はじめに

近年, 音声認識 (Automatic speech recognition, ASR) や方言識別など多くのタスクにおいて深層学習を用いた end-to-end (E2E) システムが広く用いられるようになってきている. 高性能な E2E システムの構築には大量のデータが必要であることが知られているため, 録音環境・話し方・言語など条件が異なる複数の音声データベースを組み合わせて使うことも多い. 異なる音声データを組み合わせて使うタスクの例として, 方言と標準語を組み合わせたデータベースを用いる方言識別 (Dialect identification, DID) や多方言音声認識 (Multi-dialect ASR, MD-ASR) が挙げられる. DID とは, 入力テキストや入力音声からその方言を識別するシステムである. DID の実現手法は 2 つのグループに分類できる. 1 つはアクセントや話者固有の情報など, 音響的特徴を入力特徴量として利用する手法であり, もう 1 つはテキスト情報を言語的特徴として入力特徴量に利用する手法である. 日本語の各方言は音響的にも言語的にも多様性があるため, 日本語 DID では音響的特徴量と言語的特徴量の両方を同時に考慮して識別することが望ましい. しかしながら, 2 つの特徴量を同時に考慮している研究は報告されていない. 一方, ASR とは入力された音声をテキストに変換するシステムである. その中でも入力する音声データに方言や標準語などドメインの異なるものを扱い, テキストを生成するシステムが MD-ASR である. 日本語の方言と標準語は音響的・言語的特徴が大きく異なるため, 日本語多方言を認識する MD-ASR は学習が難しいタスクとされている. 従来の研究では, DID を先に行うことで方言特有の特徴を用いながら日本語 MD-ASR を行うことで認識性能が上がると報告されている [1]. しかし, テスト時に方言情報が既知であるという条件付けがあり, 方言情報が不明な音声には適用できないという問題がある.

上述から, 方言には音響的に固有な特徴と言語的に固有な特徴の両方が含まれており, DID と MD-ASR の学習においても音響的特徴と言語的特徴を相互に利用することで両方の性能が向上できると考えられる. そこで, 本報告では, 日本語 DID と日本語 MD-ASR の順序が異なる 3 種類の結合モデリングの提案と性能比較を行う. 提案する結合モデリング手法は DID, MD-ASR の順番で行う “DID2ASR,” MD-ASR, DID の順番で行う “ASR2DID,” 同時に行う “DID+ASR” の 3 種類である. 3 つの提案手法は DID と MD-ASR の順序を変えることで, 結合モデリングごとに方言識別情報・音声認識情報のどちらを重視するかを変更できるため, 方言識別や音声認識の性能が変化すると考えられる. 実験では, 6 つの方言からなる自作の音声データベースと標準語音声データベースを用いて, 3 種類の結合モデリングを構築した. 実験結果より, 提案した手法は従来の識別手法や認識手法と比較して方言識別, 音声認識のどちらも性能を向上させたことを報告する.

2 従来法

2.1 End-to-End 日本語 DID

本節では E2E を用いた DID の従来法モデルについて説明する. このモデルは音響特徴量 $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_M\}$ から方言ラベル d の生成確率を予測するものである. ここで, $\boldsymbol{x}_m$ は, 音声フレーム中の m 番目の音響特徴量を, M は音声特徴量の総フレーム数を表す. DID モデルでは, 識別モデルのパラメータセット $\boldsymbol{\Theta}_{\text{did}}$ を用いた d の識別確率を $P(d|\boldsymbol{X}; \boldsymbol{\Theta}_{\text{did}})$ と定義する. 式 (1) に示される 音声発話とラベルのペアデータを用いてモデルパラメータは更新される.

$$\mathcal{D} = \{(\boldsymbol{X}^1, d^1), ..., (\boldsymbol{X}^T, d^T)\} \tag{1}$$

ここで T は学習データセットの発話数を表す. また, 識別モデルの目的関数は次のように定義される.

$$\mathcal{L}_{\text{did}}(\boldsymbol{\Theta}_{\text{did}}) = -\sum_{t=1}^{T} \log P(d^t|\boldsymbol{X}^t; \boldsymbol{\Theta}_{\text{did}}) \tag{2}$$

さらに, DID には異なる従来法があり, そのモデルは入力に音声ではなくテキスト列 $\phi(\boldsymbol{X}) = \phi(\{\boldsymbol{x}_1, ..., \boldsymbol{x}_M\})$ を用いて方言ラベル d の生成確率を予測している. ここで, ASR 関数 $\phi()$ によって音響特徴量 $\boldsymbol{X}$ から得られる $\phi(\boldsymbol{X})$ をテキスト列とする. テキストを入力とする場合, d の識別確率は $P(d|\phi(\boldsymbol{X}); \boldsymbol{\Theta}_{\text{did}})$ と定義され, 目的関数は次のように定義される.

$$\mathcal{L}_{\text{did}}(\boldsymbol{\Theta}_{\text{did}}) = -\sum_{t=1}^{T} \log P(d^t|\phi(\boldsymbol{X}^t); \boldsymbol{\Theta}_{\text{did}}) \tag{3}$$

日本語の方言は, 音響的・言語的に多様であるため DID において式 (2) のような音響的特徴量のみ用いる手法でも式 (3) のような言語的特徴量のみを用いる手法でも適切なモデル化には不十分であるといえる.

2.2 End-to-End 日本語 ASR

本節では, E2E 日本語 ASR のモデルについて説明する. このモデルは入力として音響特徴量 $\boldsymbol{X}$ が与え

*A comparison of joint modelings for End-to-End Japanese Dialect Speech Recognition and Dialect Identification, by Ryo Imaizumi (Tokyo Metropolitan University), Ryo Masumura (NTT), Sayaka Shiota, Hitoshi Kiya (Tokyo Metropolitan University)

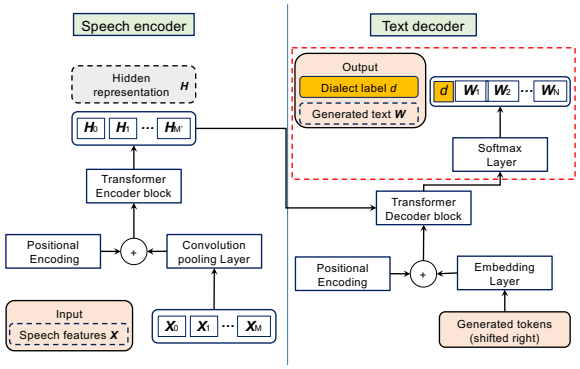

Fig. 1　DID, MD-ASR の順番で行う結合モデリング (DID2ASR) のネットワーク構造

られた時に発話内容のテキスト $\boldsymbol{W} = \{w_1, ..., w_N\}$ の生成確率を予測するものである．ここで w_n はテキストの n 番目のトークン，N はテキスト内のトークンの数を表す．自己回帰生成モデルの ASR では，認識モデルのパラメータセット $\boldsymbol{\Theta}_{\text{asr}}$ を用いて $\boldsymbol{W}$ の生成確率を次のように定義する．

$$P(\boldsymbol{W}|\boldsymbol{X};\boldsymbol{\Theta}_{\text{asr}}) = \prod_{n=1}^{N} P(w_n|\boldsymbol{W}_{1:n-1}, \boldsymbol{X};\boldsymbol{\Theta}_{\text{asr}}) \quad (4)$$

ASR では，音声発話とテキストのペアデータから，以下の式のようにモデルのパラメータを更新する．

$$\mathcal{D} = \{(\boldsymbol{X}^1, \boldsymbol{W}^1), ..., (\boldsymbol{X}^T, \boldsymbol{W}^T)\} \quad (5)$$

ASR のモデルの目的関数は次のように定義される．

$$\mathcal{L}_{\text{asr}}(\Theta_{\text{asr}}) = -\sum_{t=1}^{T}\sum_{n=1}^{N^t} \log P(w_n^t|\boldsymbol{W}_{1:n-1}^t, \boldsymbol{X}^t;\Theta_{\text{asr}}) \quad (6)$$

ここで，w_n^t は t 番目の発話の n 番目のトークン，N^t は t 番目の発話のトークンの数を表す．日本語の方言と標準語の間には音響的・言語的特徴に大きな違いがあるため，日本語の ASR において標準語のみで構築された音声認識モデルに方言を入力すると認識性能が低下することが知られている．

3　結合モデル

本章では，3 種類の DID・MD-ASR 結合モデルの説明をする．

3.1　DID2ASR

結合モデルの 1 つである DID2ASR について説明する．このモデルでは DID を行い，推定された方言情報を用いて MD-ASR を行う．DID2ASR モデルは従来の E2E DID モデル $P(d|\boldsymbol{X};\boldsymbol{\Theta})$ にテキスト $\boldsymbol{W}$ を

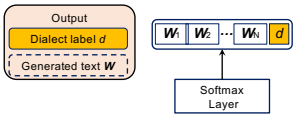

Fig. 2　MD-ASR, DID の順番で行う結合モデリング (ASR2DID) のネットワーク構造の一部 (Fig. 1 の赤枠に該当)

加えたものであり，テキストや方言ラベルの生成確率は次式のように再定義される．

$$\begin{aligned} P(\boldsymbol{W}, d|\boldsymbol{X};\boldsymbol{\Theta}) &= P(\boldsymbol{W}|\boldsymbol{X}, d;\boldsymbol{\Theta})P(d|\boldsymbol{X};\boldsymbol{\Theta}) \\ &= P(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\Theta}) \end{aligned} \quad (7)$$

$$P(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\Theta}) = \prod_{n=0}^{N} P(z_n|\boldsymbol{Z}_{1:n-1}, \boldsymbol{X};\boldsymbol{\Theta}) \quad (8)$$

ここで，方言ラベル・テキストの順番で連結させた出力列は $\boldsymbol{Z} = \{d, w_1, ..., w_N\}$ として定義される．Fig. 1 より，DID2ASR の Speech encoder は一般的な Transformer [2] に基づく E2E の構造と同様のものであり，Text decoder は，方言ラベル d とテキスト $\boldsymbol{W}$ の両方の生成確率に softmax 関数を適用する．モデルパラメータは音声，方言ラベル，テキストのセットを用いて最適化をする．

$$\mathcal{D}_{\text{comb}} = \{(\boldsymbol{X}^1, d^1, \boldsymbol{W}^1), ..., (\boldsymbol{X}^T, d^T, \boldsymbol{W}^T)\} \quad (9)$$

Speech encoder および Text decoder のパラメータを θ_{enc}, θ_{dec} とした時，目的関数は以下のように定義される．

$$\begin{aligned} \mathcal{L}_{\text{D2A}}(\theta_{\text{enc}}, \theta_{\text{dec}}) &= -\sum_{t=1}^{T} \log P(\boldsymbol{W}^t, d^t|\boldsymbol{X}^t;\theta_{\text{enc}}, \theta_{\text{dec}}) \\ &= -\sum_{t=1}^{T}\sum_{n=1}^{|\boldsymbol{Z}^t|} \log P(z_n^t|\boldsymbol{Z}_{1:n-1}^t, \boldsymbol{X}^t;\theta_{\text{enc}}, \theta_{\text{dec}}) \end{aligned} \quad (10)$$

方言ラベル，音声，テキストの生成確率を用いて最適化することで，方言特有の特徴が陽に示されることで，方言および標準語の適切な推定が可能になる．特に，DID の情報が既知の時に MD-ASR に有用な情報が与えられるため，性能が高くなると考えられる．

3.2　ASR2DID

次に DID2ASR の DID と ASR の順序を逆にして MD-ASR, DID の順番で行う ASR2DID を提案する．ASR2DID ではテキストや方言ラベルの生成確率を次のように定義する．

$$\begin{aligned} P(\boldsymbol{W}, d|\boldsymbol{X};\boldsymbol{\Theta}) &= P(\boldsymbol{W}|\boldsymbol{X};\boldsymbol{\Theta})P(d|\boldsymbol{X}, \boldsymbol{W};\boldsymbol{\Theta}) \\ &= P(\boldsymbol{Y}|\boldsymbol{X};\boldsymbol{\Theta}) \end{aligned} \quad (11)$$

モデルパラメータの更新は式 (9) を用いて最適化しており，目的関数はテキスト・方言ラベルの順番で連

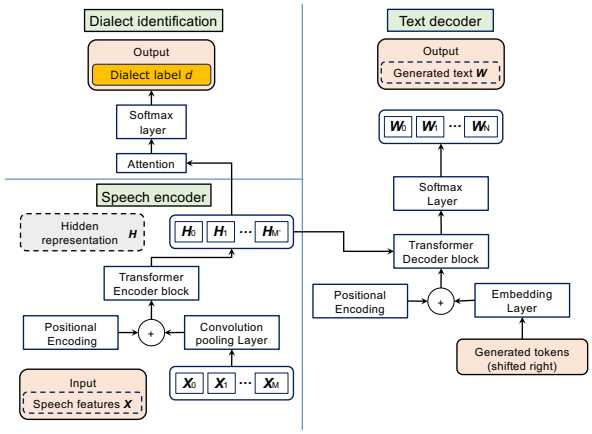

Fig. 3　DID, ASR を同時に行う結合モデリング (DID+ASR) のネットワーク構造

結させた $\boldsymbol{Y} = \{w_0, ..., w_{N-1}, d\}$ を用いて次のように定義される.

$$\mathcal{L}_{\mathrm{A2D}}(\theta_{\mathrm{enc}}, \theta_{\mathrm{dec}}) = -\sum_{t=1}^{T} \log P(\boldsymbol{W}^t, d^t | \boldsymbol{X}^t; \theta_{\mathrm{enc}}, \theta_{\mathrm{dec}})$$
$$= -\sum_{t=1}^{T} \sum_{n=1}^{|\boldsymbol{Y}^t|} \log P(y_n^t | \boldsymbol{Y}_{1:n-1}^t, \boldsymbol{X}^t; \theta_{\mathrm{enc}}, \theta_{\mathrm{dec}}) \quad (12)$$

Fig. 2 より, ASR2DID は DID2ASR における Text decoder の部分を変更して, DID と MD-ASR の順序を入れ替えることで実現している. ここで ASR2DID は DID を MD-ASR よりも後に行うため言語的特徴量を陽にとらえて方言ラベルの推定を行うことができる. そのため, DID2ASR と比較して DID の性能が高くなると考えられる.

3.3　DID+ASR

3 つ目の結合モデルは, DID と ASR を同時に推定するものである. DID2ASR や ASR2DID が方言ラベルを一意に決まるのに対し, DID+ASR の方言情報は登録された各方言の確率分布で表される. 方言ラベル d とテキスト $\boldsymbol{W}$ の生成確率は次のように定義される.

$$P(\boldsymbol{W}, d | \boldsymbol{X}; \boldsymbol{\Theta}) = P(\boldsymbol{W} | \boldsymbol{X}; \boldsymbol{\Theta}) P(d | \boldsymbol{X}; \boldsymbol{\Theta}) \quad (13)$$

DID+ASR では, 方言ラベル d と $\boldsymbol{W}$ の生成確率が独立しており, 入力音声から音素配列と方言の確率分布を予測することができる. Fig. 3 に示す通り, DID+ASR の構造は Speech encoder, Text decoder, Dialect identification の組み合わせになっている. モデルパラメータは式 (9) を用いて最適化し, 目的関数は Dialect identification のパラメータ θ_{id} を用いて次のように定義される.

$$\mathcal{L}(\theta_{\mathrm{enc}}, \theta_{\mathrm{did}}, \theta_{\mathrm{dec}}) =$$
$$\alpha \mathcal{L}_{\mathrm{asr}}(\theta_{\mathrm{enc}}, \theta_{\mathrm{dec}}) + \gamma \mathcal{L}_{\mathrm{did}}(\theta_{\mathrm{enc}}, \theta_{\mathrm{id}}) \quad (14)$$

ここで, α および γ は MD-ASR と DID のロスの重みである. DID+ASR では先に紹介した 2 つのモデルと異なり, 方言ラベルを確率分布と見なすため, 方言の情報の度合いを認識することができ, より柔軟に方言同士の特徴量の違いを識別し, ASR に利用することが可能であると考えられる.

4　実験

4.1　データベース

本実験で使用するデータベースとして自作の日本語方言音声データベース [3] と標準語音声データベースの 2 つを用いた. 方言データベースは, 青森, 広島, 熊本, 名古屋, 札幌, 仙台の 6 地方の方言から構成されており, 標準語データベースには, 日本語話し言葉コーパス (CSJ) を用いた. 各方言と CSJ の発話数は [3] と同様のものを用いている. 方言データにおける方言ごとの男女比は偏りがない. 各方言発話は iPhone5 または XperiaZ1 を用いて収録されており, 日常会話をメインとした 7 秒程度のものとなっている. 方言音声データベースのテキストおよび方言ラベルは人手で付与されている. 全データベースのサンプリング周波数は 16kHz, 量子化ビットは 16bit となっている.

4.2　実験条件

提案する DID および MD-ASR の結合モデリングの詳細な条件を示す. Speech encoder のエンコーダーブロック数は $I = 8$, Text decoder のデコーダーブロック数は $J = 6$ とした. Transformer ブロックの構成については, 出力連続表現を 256 次元, 位置ごとの feed forward ネットワークの内部出力を 2,048 次元, Multi-head attention のヘッド数を 4 とした. Speech encoder では入力に音響特徴量として 40 次元のログメルスケールフィルターバンクにデルタおよび加速係数を追加して使用した. フレーム長は 25 ms, フレームシフトは 10 ms とした. また音響特徴量は, ストライドが 2 の 2 つの畳み込み層とマックスプーリング層を通過したため, 時間軸に沿って 1/4 にダウンサンプリングする. DID+ASR では, 6 つの方言と標準語を表す 7 次元の one-hot ベクトルでラベルを用いて識別し, 目的関数には cross entropy loss を用いた. この時, 認識・識別に対するロスの重みである α, γ はそれぞれ 1 と 0.01 とした. Text decoder では, 256 次元の単語埋め込みを使用し, ビームサイズが 20 に設定されたビーム検索アルゴリズムを使用した. ネットワークの最適化には学習率 0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-9}$ の radam オプティマイザーを用いた. ミニバッチサイズは 16, Transformer ブロックのドロップアウト率は 0.1 を設定した. また, データ拡張として SpecAugment, 過学習を防ぐためにラベルスムージングを用いた. 本実験は方言識別も同時に行うため評価指標は次式による文字誤り率 (CER) と方言識別正解率 (ACC) の 2 つを用いた.

$$\mathrm{CER} = (1 - \frac{\text{文字正解率} - \text{挿入語数}}{\text{全文字数}}) \times 100(\%) \quad (15)$$

Table 1　従来法および結合モデルにおける方言識別の ACC(%)

		ACC
従来法	音声	71.2
	テキスト	53.3
結合モデル	DID2ASR	83.1
	ASR2DID	86.5
	DID+ASR	81.8

Table 2　従来法および結合モデルにおける多方言音声認識の CER(%)

		テストデータ		
		方言	標準語	方言+標準語
従来法		8.0	14.9	11.1
結合モデル	DID2ASR	8.4	14.2	11.0
	DID2ASR (oracle)	7.3	14.2	10.4
	ASR2DID	7.0	14.5	10.4
	DID+ASR	7.2	13.4	10.0

$$\mathrm{ACC} = \left(\frac{\text{正解ファイル数}}{\text{全ファイル数}}\right) \times 100(\%) \qquad (16)$$

4.3　実験結果

Table 1 に, 従来法および結合モデルにおける方言識別実験の ACC を示す. 従来法は, 入力を音声とする手法と, テキストとする手法の 2 種類あり, それぞれ ACC が 71.2%, 53.3%と十分な性能を得ることはできなかった. これは, 2.1 節で示したように, 日本語方言は音響的にも言語的にも固有の情報を持っているためどちらか一方の特徴を用いただけでは良い結果が得られなかったためだと考えられる. 一方, 3 種類の結合モデリングは従来法と比較して全てのモデルにおいて性能の改善が見られた. 特に, ASR2DID の ACC は 86.5%と結合モデルの中で最も良い性能を示した. ASR2DID は方言ラベルの予測において, MD-ASR によるテキストの予測結果を DID に有効活用できたことが理由の一つだと考えられる. DID2ASR の ACC は 83.1%と ASR2DID より若干低い値となった. DID2ASR および DID+ASR では先, もしくは同時に方言ラベルを推定するため, ASR2DID に比べて MD-ASR の予測結果を DID に十分に活用することはできていないことが原因だと考えられる. 以上の結果より, DID において提案する結合モデルのなかでは ASR2DID を用いることが最も有用であることが確認できた.

Table 2 に, 従来法および結合モデルにおける多方言音声認識の CER を示す. 学習データは方言と標準語を混ぜた混合データ, テストデータには方言のみ, 標準語のみ, 方言と標準語を混ぜた混合データを用いている. DID2ASR と従来法を比較すると, 方言のみのテストケースでは, DID2ASR の CER は 8.4%であり, 従来法の CER, 8.0%よりも高かった. 一方, 標準語のみのテストケースでは, DID2ASR の CER は 14.2%となり, 従来法の CER よりも低下した. DID2ASR の方言テストケースでの性能が悪化した理由は, DID での誤識別の影響があると考えられる. 性能の悪化の原因を調査するために DID2ASR の方言ラベルを既知, つまり ACC が 100%の時の場合の CER を調査した. その結果が DID2ASR (oracle) であり, 方言テストケースでの CER が 7.3%となった. 以上の結果より DID2ASR は MD-ASR の性能が DID の精度に依存することが確認できた. 次に, 方言のみのテストケースにおける ASR2DID では, CER は 7.0%まで改善した. ASR2DID では先に方言認識を行い, 認識で得られたテキスト情報を考慮して方言ラベルを識別する. そのため方言ラベルの誤りの影響が少なく, 認識のテキストに対して方言ラベルが補正をかけたため性能が改善したと考えられる. 一方, 標準語のテストケースでは CER が 14.5%と DID2ASR の CER よりも高くなっている. これは ASR2DID が方言に過剰に適応してしまったことで標準語の認識に不必要な情報が増えたためと考えられる. 最後に DID+ASR の場合, 方言のみのテストケースでは 2 番目の性能, 標準語のみのテストケースでは最高の性能を示した. DID+ASR では ASR2DID と異なり, 方言や標準語のラベル情報が確率分布で示されているため方言に過剰に反応することなく, 方言・標準語の両方に対して性能が改善したと考えられる. 以上の結果より, MD-ASR において最も有用な結合モデルは DID+ASR であることが確認できた.

5　まとめ

本報告では, 日本語 DID と E2E の MD-ASR の順序を変更した結合モデリングを比較検討した. 実験結果より, 3 つの結合モデルは識別・認識の全ての条件において従来法を上回るものであった. DID において, ASR2DID が最も良い性能を示し, MD-ASR においては DID+ASR が最も良い性能を示した. この結果より, ASR2DID は, 音声対話や方言を意識したアプリケーションの前処理システムとして採用するのに有効であり, DID+ASR は標準語も含む多方言音声認識をする際に有効であることが示された. 今後の課題として, CTC/Attention hybrid system など、他のネットワークについても実験やそれぞれの結合モデルを組み合わせた実験も行う予定である。

参考文献

[1] Ryo Imaiuzmi, et al. Dialect-aware modeling for end-to-end japanese dialect speech recognition. In *proc. APSIPA ASC*, pp. 297–301, 2020.

[2] Ashish Vaswani, et al. Attention is all you need. In *proc. NIPS*, pp. 5998–6008, 2017.

[3] Ryo Imaizumi, et al. End-to-end japanese multi-dialect speech recognition and dialect identification with multi-task learning. *APSIPA Transactions on Signal and Information Processing (accepted)*, 2021.