Vol.2017-SLP-119 No.19
2017/12/22

敵対的学習を適用したEnd-to-end音声認識

増田 嵩志[1,a)]　齋藤 大輔[1]　峯松 信明[1]

概要：End-to-end 音声認識は DNN-HMM 音声認識と同等の精度を達成しており, 近年盛んに研究が行われている. しかし基本的に単一のコーパスのみで学習を行う End-to-end 音声認識は, コーパスの性質をより有してしまう傾向にあるといえるため, 汎化性能に留意する必要がある. そこで汎化性能をより向上させるための手法として, 敵対的学習の枠組みを End-to-end 音声認識に適用させることを考える. 敵対的学習は, モデルの出力結果を間違わせるような微小なノイズを考慮する手法であり, モデルをこのようなノイズに頑健にさせることによって汎化性能を向上させることができる手法である. WSJ0 および Aurora4 での音声認識実験では, 敵対的学習を適用することにより, エラー率を大きく改善させることに成功した. またノイズを考慮する敵対的学習の枠組みにより耐雑音性を向上させることも期待されたが, 単純に適用させるだけでは不十分であることが判明した.

キーワード：音声認識, End-to-end, Connectionist Temporal Classification, 敵対的学習, 耐雑音性

1. はじめに

近年, 音声特徴量から文字を直接出力する End-to-end 音声認識モデルが提案されている. 従来の音声認識システムは音響モデルや言語モデル, 発音辞書, デコーダー等複数の要素から構成されており, それぞれの構成要素は各々の学習基準によって最適化され, 複雑な工程を経てデコードが行われる [1]. それに対し End-to-end モデルは一つのニューラルネットワークで構成されており, 所望の誤差基準に近い基準でネットワーク全体が最適化されるだけでなく, デコード機構も非常にシンプルである. End-to-end 音声認識には, Connectionist Temporal Classification (CTC)[2] に基づく手法と, Attention mechanism に基づく手法の大きく分けて 2 つの手法が提案されており, 現在盛んに研究が行われている.

しかしながら, 構成要素を別々のコーパスで学習させていた従来の音声認識システムとは異なり, End-to-end 音声認識では一つのコーパスを元に学習が行われるため, コーパスに依存した性質を有してしまう傾向にあるといえる. 例えば clean な音声で構成されているコーパスを使って学習させた場合, 当然のように耐雑音性は著しく低くなる. したがって End-to-end モデルを構築する場合は汎化性能に留意する必要があると言える. 最もシンプルに汎化性能を向上させる手法として, 大量の音声データおよび書き起こし文を用意してネットワークを学習させるというデータドリブンな手法が考えられるが, 学習に非常に長い時間を要するだけでなく, 言語によっては十分な量の学習データが手に入らない, 等の問題点がある. また汎化性能の例として耐雑音性を挙げた場合, 雑音環境下での認識率は静音環境下の認識率と比べて劣悪であり [3], [4], 汎化性能が不十分であるといえる. [4] では学習データに雑音を付与することで学習データの水増しを行い耐雑音性を向上させているが, 先述のような学習時間の問題があるだけでなく, 雑音が混入した音声データを用いることでネットワークの最適化を難化させてしまう可能性がある. 加えて, この手法によって特定の種類の雑音に対して耐雑音性を向上させることが出来ても未知の雑音に対する性能は保証されないという欠点を持つ.

そこで本研究では, 耐雑音性を汎化性の一つとして捉え, 敵対的学習の枠組みを End-to-end 音声認識に応用することを提案する. Adversarial Training (AT)[5] は, 正解ラベルを出力する確率が最も下がるようなノイズを自動で算出し, そのノイズを付与されても出力ラベルが変わらないように目的関数に正則化項を付与して学習を行う手法である. Virtual Adversarial Training (VAT)[6] は, AT を半教師あり学習に応用した手法で, 出力ラベルではなく出力分布が変わらないようなノイズを用いる手法である. 両手法によって付与される雑音は種類を全く仮定する必要がなく, 学習の各ステップで毎度自動で算出されるため, 静音環

[1] 東京大学大学院 工学系研究科
Graduate School of Engeneering, The University of Tokyo, Bunkyo, Tokyo 113-8656, JAPAN
[a)] masuda@gavo.t.u-tokyo.ac.jp

境下での認識精度だけでなく耐雑音性も向上させることが可能であると考えられる．また学習データの増量も不要である．

次章以降の本稿の構成を以下に示す．第 2 章で End-to-end 音声認識, 第 3 章で敵対的学習の概要について述べたのち, 第 4 章でそれらを End-to-end 音声認識に適用させた本研究の手法について述べる．第 5 章で評価実験について述べ、最後に第 6 章でまとめと今後の方針について述べる．

2. End-to-end 音声認識

End-to-end 音声認識モデルでは LSTM を用いた 1 つのネットワークに音響特徴量を時系列順に入力し, 対応する文字列 (あるいは音素列) を得ることで音声認識を行う．この際, 1 つの文字出力に対し複数フレームの入力が対応するため, この入力と出力のマッピングを実現させるために, 目的関数として CTC[2] を用いる手法 [4], [7], [8], [9] と, Encoder-Decoder モデルをベースとし Attention mechanism を用いる手法 [10], [11], [12] の大きく分けて 2 つの手法が提案されている．より高い精度を得るために言語モデルを用いたデコードをするのが一般的だが, Attention mechanism による手法は後述の独立性の仮定が不要であるため, 言語モデルを使用しない場合は CTC による手法よりも高い認識精度を出すことが可能である．しかし雑音環境下での音声認識に関しては, 雑音によって Attention mechanism の働きが悪化し精度が大きく落ちるとの報告がある [3] ため, 本研究では CTC を用いた End-to-end 音声認識を扱った．以下にその概要を示す．

CTC は, 音響特徴量系列 $\boldsymbol{x}$ と文字列あるいは音素列系列 $\boldsymbol{y}$ のアライメントを自動で学習する機構である．$\boldsymbol{y}$ が K 種類のラベルで構成されているとする．CTC は $\boldsymbol{x}$ と $\boldsymbol{y}$ とのマッピングを行う代わりに, $\boldsymbol{x}$ と同じ系列長の中間表現 $\boldsymbol{\pi}$ とのマッピングを行う．$\boldsymbol{\pi}$ は先程の K 種類とブランクラベルの $K+1$ 種類のラベルで構成されており, $\boldsymbol{y}$ に対して同一ラベルの繰り返しとブランクラベルの挿入が許されている．このような中間表現を用いることで, 系列長の違いを吸収したマッピングを実現させている．

このとき, 生起確率 $P(\boldsymbol{y}|\boldsymbol{x})$ は中間表現における複数のパスの生起確率の和であるとすることが出来る．すなわち, $\boldsymbol{y}$ に対してラベルの繰り返しやブランクラベルの挿入を行って作成できる中間表現の全集合を $\boldsymbol{\Phi}(\boldsymbol{y})$ とすると, 生起確率は以下で表される．

$$P(\boldsymbol{y}|\boldsymbol{x}) = \sum_{\pi \in \boldsymbol{\Phi}(\boldsymbol{y})} P(\boldsymbol{\pi}|\boldsymbol{x}) \tag{1}$$

$P(\boldsymbol{\pi}|\boldsymbol{x})$ は, 各時刻の出力の独立性を仮定することで, それぞれの時刻での出力の生起確率の積で近似される．

$$P(\boldsymbol{\pi}|\boldsymbol{x}) \approx \prod_{t=1}^{T} P(\pi_t|\boldsymbol{x}) = \prod_{t=1}^{T} q_t(\pi_t) \tag{2}$$

ここで $q_t(\pi_t)$ は, 時刻 t での RNN の出力 π_t に対する softmax 関数の出力結果である．CTC は損失関数として, 正解ラベル系列 $\boldsymbol{y}$ に対する負の対数尤度で定義され, 以下のように表される．

$$\mathcal{L}_{\mathrm{CTC}}(\boldsymbol{x}, \boldsymbol{y}) \equiv -\ln P(\boldsymbol{y}|\boldsymbol{x}) \tag{3}$$

この確率分布を求める際には HMM 同様に forward-backward アルゴリズムを用いて計算量を削減することが可能である．

デコードの際には, 各時刻について最も高い生起確率を与えるラベルを生起することで出力系列を簡単に求めることが可能である．さらにビーム探索を行うことで, より正確な出力系列を得ることが出来る．

3. 敵対的学習

DNN は画像認識や音声認識等の分野で驚異的な精度を達成しているが, 非直感的な認識を行うことがある．[5] や [6] では, 学習済みの DNN を欺くように作成されたノイズを加えたサンプル (Adversarial example[13]) にも頑健になるように学習させることで汎化性能を向上させる敵対的学習を提案している．Goodfellow らが提案した Adversarial Training (AT)[5] は, 正解ラベルの尤度を下げるようなノイズを加えた場合の対数尤度を正則化項として設定しており, 正解ラベルを必要とする．宮戸らが提案した Virtual Adversarial Training (VAT) [6] は AT を半教師あり学習に応用した手法であり, 正則化項は教師ラベルを必要としない．これらの手法は画像認識の分野で適用され, どちらの手法も認識精度の向上を実現している．

3.1 Adversarial training

x を入力, y を正解ラベル, θ をモデルのパラメータとする．AT は, 正解ラベルの尤度が最も小さくなるようなノイズを加えた場合の対数尤度を正則化項として目的関数に加えて学習させる手法である．

$$\mathcal{L}_{\mathrm{AT}} \equiv -\ln P(y|x + r_{at}; \theta) \tag{4}$$

$$r_{at} = \arg\min_{r, |r| \leq \epsilon} \ln P(y|x + r; \hat{\theta}) \tag{5}$$

ここで r は入力に対するノイズであり, $\hat{\theta}$ は学習時のモデルパラメータである．AT では毎学習時このようなノイズを用いて学習を行うが, 式 (5) を満たすノイズを解析的に求めるのは難しい．そこで Goodfellow ら [5] は, 線形近似によってこのノイズの値を以下のように近似している．

$$r_{at} \approx \epsilon \mathrm{sign}(\nabla_x P(y|x; \hat{\theta})) \tag{6}$$

ニューラルネットワークの誤差逆伝播法によって, 式 (6) 中の勾配は簡単に求めることが可能である．

3.2 Virtual adversarial training

AT が正解ラベルの尤度を下げるノイズを用いるのに対し, VAT ではモデルの出力分布が最も変わるようなノイズを用いる. このようなノイズを入れた前と後の KL ダイバージェンスを小さくするように, 以下で表される項を目的関数に加える.

$$\mathcal{L}_{\mathrm{VAT}} \equiv \Delta_{\mathrm{KL}}(r_{vat}, x) \tag{7}$$

ただし

$$\Delta_{\mathrm{KL}}(r, x) = \mathrm{KL}[P(y|x;\hat{\theta})|P(y|x+r;\hat{\theta})] \tag{8}$$

$$r_{vat} = \arg\max_{r,|r|\leq\epsilon} \Delta_{\mathrm{KL}}(r, x) \tag{9}$$

ノイズの混入に対してモデルの出力分布が変わらないように学習されるため, VAT の利用によって予測分布がなめらかになることが期待される. また VAT によって算出される正則化項は正解ラベルを必要としないため, 半教師あり学習に応用することが可能である.

AT 同様, r_{vat} を解析的に求めるのは難しいため, 宮戸ら [6] は以下のような効率的な近似手法を用いて近似している. テイラー展開を用いた 2 次近似により,

$$\Delta_{\mathrm{KL}}(r, x, \theta) \simeq \frac{1}{2} r^T H(x, \theta) r \tag{10}$$

となる. ここで H はヘッセ行列である. この近似のもとでは, r_{adv} は H の最大固有値に対応する固有ベクトルとなるが, H を求めるための計算コストが大きいため, べき乗法および有限差分法を用いた近似によって H を近似する.

AT 同様, r_{vat} を解析的に求めるのは難しいため, 宮戸ら [6] は以下のような効率的な近似手法を用いている. テイラー展開を用いた 2 次近似から

$$\Delta_{\mathrm{KL}}(r, x, \theta) \simeq \frac{1}{2} r^T H(x, \theta) r \tag{11}$$

となる. ここで H はヘッセ行列であり, $H(x,\theta) = \nabla\nabla_r \Delta_{KL}(r, x, \theta)|_{r=\mathbf{0}}$ である. 一般に $|x| = 1$ のとき, $x^T A x$ の最大値は行列 A の最大固有値となるので, H の最大固有値に対応する固有ベクトルを u と表せば,

$$r_{vat} = \arg\max_{r} \{\Delta_{\mathrm{KL}}(r, x, \theta); ||r||^2 < \epsilon\} \tag{12}$$

$$\simeq \epsilon\overline{u(x,\theta)} \tag{13}$$

と近似できる ここで $\bar{x}$ は x の単位ベクトル化したものである.

以上より H および $\boldsymbol{u}$ を算出することで近似的に $\boldsymbol{r}_{vat}$ を求めることが出来るが, 計算コストが大きいため, べき乗法および有限差分法を用いて再び近似を行う.

べき乗法は行列の最大固有値および対応する固有ベクトルを求める方法である. 適当なベクトル $\boldsymbol{d}$ に以下の演算を繰り返し適用することを考える.

$$d \leftarrow \overline{Hd} \tag{14}$$

このようにすることで d は最大固有値に対応する固有ベクトルに漸近する. 以上によって u を求める.

一方で Hd そのものの導出には以下のようにして有限差分法を用いることで計算コストを抑える.

$$Hd \simeq \frac{\nabla_r \Delta_{KL}(r, x, \theta)|_{r=\xi d} - \nabla_r \Delta_{\mathrm{KL}}(r, x, \theta)|_{r=\mathbf{0}}}{\xi d} d$$

ここで, 分子第二項ではノイズ r がないときの KL ダイバージェンスであり, この時 KL ダイバージェンスとその微分値は 0 であるから,

$$Hd \simeq \frac{\nabla_r \Delta_{\mathrm{KL}}(r, x, \theta)|_{r=\xi d}}{\xi} \tag{15}$$

となる. 残った分子の項もネットワークの順伝播, 逆伝播を行うことで求めることができる.

以上より, べき乗法によって

$$d \leftarrow \overline{\nabla_r \Delta_{KL}(r, x)|_{r=\xi d}} \tag{16}$$

を繰り返し行なって得られた d を用いて,

$$r_{adv} = \epsilon d \tag{17}$$

として近似値を得ることが可能である. また宮戸ら [6] によると, べき乗法の適用回数は 1 回で十分とされており, 結果として少ないハイパーパラメータ数で VAT を適用することが可能である.

4. End-to-end 音声認識における敵対的学習の適用

本研究では, CTC の枠組みに敵対的学習として AT および VAT を適用させることによって End-to-end 音声認識の正則化を行なった. AT および VAT は元々画像認識の分野で考案された手法だが, これを音声認識という系列タスクに適用させた. CTC への適用は毎時刻ノイズを生成することによって成立し, 以下のように目的関数を変更することで可能となる.

$$\mathcal{L} = \mathcal{L}_{\mathrm{CTC}}(\boldsymbol{x}, \boldsymbol{y}) + \alpha\mathcal{L}_{\mathrm{AT}}(\boldsymbol{x}, \boldsymbol{y}) \tag{18}$$

$$= -\ln P(\boldsymbol{y}|\boldsymbol{x}) - \alpha \ln P(\boldsymbol{y}|\boldsymbol{x} + \epsilon \mathrm{sign}(\nabla_{\boldsymbol{x}} \ln P(\boldsymbol{y}|\boldsymbol{x}))) \tag{19}$$

$$\mathcal{L} = \mathcal{L}_{\mathrm{CTC}}(\boldsymbol{x}, \boldsymbol{y}) + \alpha\mathcal{L}_{\mathrm{VAT}}(\boldsymbol{x}) \tag{20}$$

$$= -\ln P(\boldsymbol{y}|\boldsymbol{x}) - \alpha \sum_t \Delta_{\mathrm{KL}}(\epsilon\overline{\nabla_r \Delta_{KL}(r, x_t)|_{r=\xi d}}, x_t) \tag{21}$$

ここで α は正則化項の重み付けのための値, ϵ はノイズの規模を表す値, d はランダム値を要素とする単位ベクトルである. α と ϵ はハイパーパラメータである.

敵対的学習は前述のようにモデルの性能を悪化させるノイズを考慮して学習を行う手法であるため, 耐雑音性のような性能を向上させることができると期待される.

5. 評価実験

TIMIT のような小規模コーパスにおけるタスクでは敵対的学習の有効性が確認できている [14] ため, 今回は, より規模の大きいコーパスにおいて有効であるかの検証を行った.

5.1 実験設定

音声コーパスには WSJ0[15] を用いた. WSJ0 は 15 時間ほどの静音環境下における音声データで, 学習データの発話数は 7,138 発話である. 開発データおよび評価データにはそれぞれ WSJ0 の “dev93” セット, 評価データには “eval92” セットを用いた.

一方, 耐雑音性を評価するため, Aurora4 を用いた. このコーパスは以下の 6 種類の (環境) 雑音を WSJ0 に人工的に付与したものである.

- Car
- Crowd of people (babble)
- Restaurant
- Street
- Airport
- Train station

学習データは上記 6 種類からランダムに選ばれた 1 種類の雑音が 10dB〜20dB で付与されている. 一方開発データ “dev330” および評価データ “test166” に関しては, 学習データ同様 6 種類からランダムに選ばれた 1 種類の雑音が, 5dB〜15dB で付与されいる.

以上のコーパス中の音声データから, Kaldi[16] を用いて 40 次元のメルフィルタバンク特徴量を抽出し, Δ特徴量および ΔΔ特徴量を加えた計 120 次元をモデルの入力とした. フレーム幅は 25msec, シフト長は 10msec である. 出力はアルファベット 26 種類にアポストロフィ, ピリオド, ダッシュ, スペース, noise ラベル, 文終了ラベル, そしてブランク文字を加えた計 33 クラス分類を毎フレーム行った.

モデルのネットワーク構造は全て, 隠れ層が 256 ユニットの 4 層の Bidirectional-LSTM を用いた. モデル中のパラメータは全て [-0.1, 0.1] の範囲で初期化を行った. 勾配のノルムの絶対値が 10 を超えないようにクリッピングを行った. バッチサイズは 32 とした. 学習率は 0.001 とし, Adam[17] による最適化を行った. デコードの際にはビーム幅 20 のビームサーチを行なった.

式 (19) および式 (21) 中のノイズの規模 ϵ を決めるにあたり, 開発データを用いてグリッドサーチを行なったところ, それぞれ 0.3, 5.0 としたときに最も高い性能が得られた. 式 (19), 式 (21) 中の正則化項の重み α は 1.0 に設定した.

なお, 今回の実験では言語モデルや語彙情報等は使用しなかった.

表 1 WSJ0 および Aurora4 における Character Error Rate (CER)

学習データ/モデル	CER(開発データ)	CER(評価データ)
WSJ0	dev93	eval92
CTC[3]	31.23	24.69
CTC	33.30	25.58
CTC + AT	31.28	22.90
CTC + VAT	**29.34**	**21.87**
Aurora4	dev330	test166
CTC	35.71	35.91
CTC + AT	33.67	33.94
CTC + VAT	**33.36**	**33.14**
WSJ0	dev330	test166
CTC	67.05	69.49
CTC + AT	63.86	66.51
CTC + VAT	**62.17**	**64.19**

5.2 結果および考察

各モデルの認識エラー率は**表 1** のようになった. WSJ0 や Aurora4 の各コーパスで学習させた結果, CTC 単体のベースラインよりも AT や VAT による敵対的学習を用いた枠組みの方が低いエラー率が得られた. 一方, Aurora4 が WSJ0 に人工的にノイズを付与して作成されたコーパスであることを利用して, WSJ0 で学習させた後に Aurora4 の開発データ “dev330” および評価データ “test166” でのエラー率を測定することで, 敵対的学習による耐雑音性の向上を検証した. その結果, 同じ条件のもとで学習させたベースラインよりも AT や VAT の方がエラー率が低くはなったものの, この条件下でのエラー率は全般的に悪い結果となり, Aurora4 で学習させたベースラインと比較させた場合, 大きく劣る結果となった. 以上から, 敵対的学習の枠組みを導入した場合, 汎化性能は向上するが, 耐雑音性を大きく向上させるまでには至らず, 雑音を付与して人工的に作成したコーパスによる学習が現段階ではより効果的であると言える.

TIMIT コーパスにおける先行研究 [14] でも同様だが, このような敵対的学習によって生成されるノイズのノルムは非常に小さい値に設定するのが一般的かつもっとも効果的であり, 単純に大きな値に設定するだけでは精度が悪化してしまう. これは敵対的学習の考え方が, 「入力の微妙な揺れに対して頑健なモデルを生成する」ということに起因する. より雑音耐性を向上させる枠組みへ変化させるためには, ノイズの探索範囲を工夫する必要があると考えられる. またノイズの加え方に関しても, 今回は対数メルフィルタバンク特徴量にノイズを加算する形で付与させているが, これでは汎化性能を向上させる効果はあるとしても, 耐雑音性を向上させる手段としては不適切である可能性が高いと考えられる. より耐雑音性を向上させるためには, 対象とする雑音の加わり方に倣った定式化が必要となると言える.

6. おわりに

本研究では End-to-end 音声認識に adversarial training および virtual adversarial training を適用した．WSJ0 や Aurora4 といった中規模なデータセットで音声認識を行ったところ, それぞれのコーパスでの学習では精度が大きく向上した．しかし耐雑音性に関しては, 学習コーパスに雑音を付与する方法と比較すると大きく精度が劣っているため, 単純に敵対的学習を適用しても有効でないことが判明した．

今後の課題としては, 5.2 章で述べたように, AT および VAT で生成するノイズの探索空間やノイズの加え方を, より耐雑音性を向上させることを目的としたアプローチに変更させることを考えている．より具体的には, 敵対的学習では正解ラベルを変えないように微小なノイズを考慮するようにしているが, 例えば話者性による入力特徴量の分布の揺れを排除させることによってより大きなノイズを考慮しても問題ないような枠組みに変更することがあげられる．

参考文献

[1] Hinton, G. E., Deng, L., Yu, D. et al.: Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups, *Signal Processing Magazine*, Vol. 29, No. 6, pp. 82–97 (2012).
[2] Graves, A., Fernández, S., Gomez, F. and Schmidhuber, J.: Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks, *Proceedings of the 23rd International Conference on Machine Learning*, pp. 369–376 (2006).
[3] Kim, S., Hori, T. and Watanabe, S.: Joint CTC-Attention based End-to-End Speech Recognition using Multi-task Learning, *2017 IEEE International Conference on Acoustics, Speech and Signal Processing* (2017).
[4] Amodei, D., Anubhai, R., Battenberg, E. et al.: Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin, *Proceedings of The 33rd International Conference on Machine Learning*, Vol. 48, pp. 173–182 (2016).
[5] Goodfellow, I. J., Shlens, J. and Szegedy, C.: Explaining and harnessing adversarial examples, *International Conference on Learning Representations* (2015).
[6] Miyato, T., Maeda, S., Koyama, M., Nakae, K. and Ishii, S.: Distributional Smoothing with Virtual Adversarial Training, *International Conference on Learning Representations* (2016).
[7] Graves, A. and Jaitly, N.: Towards End-To-End Speech Recognition with Recurrent Neural Networks, *Proceedings of the 31st International Conference on Machine Learning*, Vol. 32, No. 2, pp. 1764–1772 (2014).
[8] Miao, Y., Gowayyed, M. and Metze, F.: EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding, *2015 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 167–174 (2015).
[9] Zhang, Y., Pezeshki, M., Brakel, P., Zhang, S., Laurent, C., Bengio, Y. and Courville, A.: Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks, *Interspeech 2016*, pp. 410–414 (2016).
[10] Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K. and Bengio, Y.: Attention-based Models for Speech Recognition, *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pp. 577–585 (2015).
[11] Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P. and Bengio, Y.: End-to-end attention-based large vocabulary speech recognition, *2016 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4945–4949 (2016).
[12] Chan, W., Jaitly, N., Le, Q. V. and Vinyals, O.: Listen, attend and spell: A neural network for large vocabulary conversational speech recognition, *2016 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4960–4964 (2016).
[13] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R.: Intriguing properties of neural networks, *International Conference on Learning Representations* (2014).
[14] 増田嵩志，齋藤大輔，峯松信明：耐雑音性の向上を目的とした End-to-end 音声認識における Virtual Adversarial Training の適用，日本音響学会秋季講演論文集 (2017).
[15] John Garofalo, David Graff, D. P. and Pallett, D.: CSR-I (WSJ0) Complete, Linguistic Data Consortium, Philadelphia, USA (2007).
[16] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G. and Vesely, K.: The Kaldi Speech Recognition Toolkit, *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, pp. 1–4 (2011).
[17] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, *International Conference on Learning Representations* (2014).