Vol.2018-SLP-124 No.2
2018/10/10

データ拡張処理の非ネイティブ英語音声認識への効果

福田 隆[1]　ラウル フェルナンデス[2]　サミュエル トーマス[2]　アレキサンダー ソリン[3]　倉田 岳人[1]

概要：外国語なまりのアクセントを持つ話者（非ネイティブ，L2 話者）の音声認識は未だチャレンジングな状況にある．この課題に対する最も効果的なアプローチは，外国語なまりのアクセント音声を収集し，発話内容の書き起こしと共に学習データに含めることである．しかし，非ネイティブ話者の音声はネイティブ話者（L1）に比べて豊富に収集できる訳ではないため，収集されたデータがもたらすインパクトには限りがある．本報告は，非ネイティブ話者の音声に対する人工的データ拡張処理が，音声認識にどのような効果をもたらすかを実験的に検証する．実験では，ラテンアメリカ英語とアジア英語の 2 種類を対象に，声質変換（声帯振動と声道特性の変換），話速変形，雑音付与によって複数のコピーを生成し，教師あり学習と教師なし学習の両方のシナリオで，人工的に生成されたデータの効果を確認する．非ネイティブ話者の音声認識には話速変換，声質変換，雑音付与によるデータ拡張処理の順で効果があったことを述べる．特に外国語なまりアクセントの専用音響モデルをスクラッチから構築する場合にデータ拡張の効果が大きく，話速変換を用いたデータの生成によって，30%以上の相対的誤り削減が得られるケースもある．

キーワード：音声認識，音響モデル，非ネイティブ，アクセント，データ拡張

1. はじめに

音声認識性能は改善の一途にあり，一部のケースでは人間の性能に匹敵するまでに至っている．しかしながら，外国語なまり（非ネイティブ）のアクセントを持つ音声の認識精度は未だ低く，ネイティブ話者の認識性能と比べて大きな乖離がある．英語のようなグローバル言語では，非ネイティブ話者の人口がネイティブ話者の 3 倍にまで至っているとの推定もあり，実用上大きな問題となっている [1]．これはよく知られた事実であり，様々な角度からの研究が行われている．性能改善の上で効果的な方法は，非ネイティブ話者の音声を収集し，そのデータセットの単独利用でモデルの学習を行うか，もしくはネイティブ話者の音声と組み合わせて利用することであるが，時間的・金額的な収集コストは無視できないほど高い．

非ネイティブ話者の人口が多いにも関わらず，パブリックに利用できるコーパスや各研究機関が独自に所有しているデータのほとんどは，ネイティブ話者によるものである．また，一概に非ネイティブ話者といってもアクセントの様式は多種多様であり，母国語がどこであるかということもさることながら，英語を話した経験や熟練度によって顕著な差があることが問題をさらに難しくしている．例えば，英語の初学者は流暢な話者と比べて音響的な発話スタイルが大きく異なる．また容易に想像できるように，日本人による英語はスペイン語話者の英語と特性が大きく異なる．これらの多角的な要因が多くの音響的変形を生み出し，話者の違いや録音環境の違い以上に，非ネイティブ話者の認識処理に対して大きな影響を与える．そもそもの収集量が少ないことに加えて，母国語や英語学習の経験に基づく様々な要因をクリアしていかなければならない．

この問題に対する有望なアプローチの一つはデータ拡張処理（data augmentation）である[*1]．これは利用可能な音声データについて人工的な変形を加えて新たな音声を作る方法であり，発話内容を変えるわけではないので，書き起こしデータなどは既存のものをそのまま利用できる．また，音声データ長を変えるタイプの変形でなければ，音素アライメントも元々のデータから推定したものを活用してもよい．本報告では，少量の非ネイティブ話者の音声が利用可能である状況を想定し，データ時間領域でのデータ拡張処理を対象に，スクラッチから音響モデルを学習する場合と，何らかの既存のモデルに対して適応処理を行うケースについて効果を検証する．本報告で述べる技術的な貢献は以下のとおりである．

- データ拡張処理としての声道特性および声帯音源特性

[1] 日本 IBM 東京基礎研究所
[2] IBM T. J. Watson Research Center
[3] IBM Haifa Research Lab

[*1] Data augmentation は直訳すればデータ増強であるが，本報告では用語をデータ拡張に統一する．

に基づく声質変換技術の導入
- 話速変換，声質変換，雑音付加を含む 3 つのデータ拡張処理の非ネイティブ音声に対する効果の比較
- 非ネイティブ英語音声を用いた教師なし音響モデル適応の効果の検証

本報告では，非ネイティブ話者の音声を用いた認識実験により，教師あり・教師なし学習の文脈で声質変換，話速変換が性能改善に大きく寄与することを述べる．また，驚くべきことに，外国語なまりのあるアクセントを持つ話者の音声認識において，最も簡便な話速変換処理が顕著な改善効果をもたらすことを示す．

2. 先行研究

様々な信号・データ処理技術を利用するデータ拡張処理は，学習データの多様性を人工的に増加させる技術として広く用いられている．人工的とは言え，学習データの増加はモデルの過学習を避けることに繋がり，結果としてより頑健なモデルを構築することができる．Ko らは様々な音声認識タスクで話速変換技術を用いたデータ拡張を比較し，効果を検証している [2]．Hartmann らは話速変換に加えて時間領域での雑音付加処理も含めたデータ拡張処理を用い，また fMLLR に基づく周波数領域での変換処理も併用して，幅広く効果の検討を行った [3]．類似研究として Cui らは，VTLP（Vocal Tract Length Perturbation）や SFM（Stochastic Feature Mapping）などの周波数領域におけるデータ拡張処理を低リソースの音声認識精度改善に用いた [4]．一方，Ragni らは教師なし学習やマルチリンガル音声認識の枠組みでデータ拡張処理の可能性を模索した [5]．また，多くの音声認識システムでは環境変化に対する頑健性が望まれるため，クリーン音声に様々な加法性・乗法性雑音を付加して学習データのバリエーションを増やす処理が従来からよく検討されている [6]．

本報告では，外国語なまりのアクセントを持つ非ネイティブ話者の英語音声認識率の改善を目的として，時間領域でのデータ拡張処理を比較検討し，これまで検討が行われてこなかった声質変換処理の音声認識に対する可能性も模索する．本報告で検討する声質変換処理は，元々 prosody-labelling タスクのために提案されたものであり，その根幹の部分を学習データの拡張に利用する [7]．我々の知る限り，非ネイティブ話者の音声認識を対象とした研究において，声質変換や話速変換処理を導入した事例はない．

3. データ拡張処理

本章では，声質変換（3.1 章），話速変換（3.2 章），雑音付加（3.3 章）の順でデータ拡張の詳細について説明する．

3.1 声質変換

本報告で検討する声質変換は，声帯音源と声道特性の変形操作により，ある話者の声を別の話者の特性に近づけようとするものである．この変換は発話の音素ラベルを保持し，時間方向の伸縮は行わないので，変換後の音声に対しても元々の音声に用いた書き起こしや音素アラインメントをそのまま利用できる．本節では，分析，合成と変換処理についての概観を説明する．詳細は文献 [8] を参照されたい．

3.1.1 分析

まず，入力信号からピッチの概形を 5 ms シフトで抽出する．分析は有声音区間についてのみ行い，3.5 ピッチサイクルの短区間窓で行う．無声音（子音）区間は分析対象外とする．反復処理による適応逆フィルタを用いて声帯音源波形（glottal source derivative wavefom）を推定し，3 つの要素を持つベクトル $\theta = [T_p, T_e, T_a]^T$ で表現する．これは，Liljencrants-Fant（LF）モデル [9] と呼ばれ，T_p, T_e, T_a はタイミングパラメータを表す．

声道の周波数特性は 40 次元の線スペクトル対（LSF: Line Spectral Frequencies）で表現し，7 フレームの移動平均窓で平滑化する．

3.1.2 合成

パラメータ群から時間領域の信号を復元において，有声音区間のフレームは連続性を保持するためにスタックし，その後無声音区間フレームを挟み込む．ただし，無声音区間の波形は修正の対象としない．有声音区間を合成するため，所望の F0 値に照らし合わせてピッチサイクルの始まり位置の系列を生成する．

各ピッチサイクルに関連づけられる声帯音源と声道特性のパラメータは，そのサイクルのエッジフレームに対応するパラメータまでとの間で補間をとることによって生成する．生成された声門パルス系列は，ゲインファクターで調整する．加法性の Aspiration noise は，500Hz の高域通過フィルタで処理したガウス雑音とする．この振幅変調は雑音信号の包絡線を形成し，声門パルスのエネルギー変化を模擬する．

LSF パラメータは自己回帰係数に変換の後，声門パルス系列に変形を加えるフィルタの役割を果たす．フィルタ係数は各ピッチサイクルの始めに更新し，各有声音区間は隣接する無声音区間と重畳加算によって結合する．

3.1.3 変換

上記アルゴリズムは，ピッチ，声道特性，声帯振動パルスを修正する時間不変のグローバルな声質変換を実現するためのものである．ピッチの修正は，元々の音声の F0 軌跡を f_0^{shift} と f_0^{range} によって転調・伸縮させる．声道特性の変換はスプライン関数によるユーザ指定の変曲点で平滑化することによって実現する．一方，声門パルスは以下 2 つの戦略で変換する．

(i) ユーザによって与えられる声門パルスベクトル θ_{ref} を，混合重み $0 \leq \alpha \leq 1$ で変形．

$$\hat{\theta} = (1-\alpha)\theta + \alpha\theta_{ref} \quad (1)$$

(ii)“Lax” と “Tense” として記述される音質に関係する 2 つの定型アンカーパルスとの補間

$$\hat{\theta} = \begin{cases} (1-\beta_{lt})\theta + \beta_{lt}\theta_l & \text{if } \beta_{lt} > 0 \\ (1-|\beta_{lt}|)\theta + |\beta_{lt}|\theta_t & \text{otherwise} \end{cases} \quad (2)$$

ここで，$-1 \leq \beta_{lt} \leq 1$ は “Lax” と “Tense” のトレードオフに関するユーザ指定のパラメータである．

3.2 話速変換

話速変換は，サンプリングされたディジタル信号を 0.9〜1.1 の倍率でリサンプリングすることによって実現する．単なるリサンプリングの場合，発話長だけでなく，ピッチと周波数スペクトルにも影響を及ぼすが，ネイティブ話者が主体の音声認識タスクではその変換が有効であることがわかっている [2]．

3.3 雑音付加

雑音に頑健な音声認識を実現するために，元々のクリーンな学習データに雑音を付加してバラエティを増やし，マルチコンディションのデータセットを生成する戦略は実用上でも非常に有効な方法である．例えば，雑音を人工的に付加し，マルチコンディション学習データの効果を実証しているデータセットとして Aurora コーパスが有名である [10]．このデータセットは，路上，駅，車内，人混み，レストラン，空港で収集された雑音と，様々なマイク特性を反映した乗法性雑音をクリーン環境収録の音声に畳み込んでバラエティに富んだデータセットを構成している．本報告では雑音付加に FaNT（Filtering and Noise-adding Tool）を利用する．本報告で用いる後述のテストデータにはレベルの小さなわずかな雑音のみが含まれているので，雑音付加によるデータ拡張では SNR=20dB 以上になるように雑音レベルを調整した．これは単なる雑音データ拡張の側面だけでなく，元々の波形にわずかな変化を加えることによって（比較的クリーンな）アクセント音声学習セットのバリエーションを増やす効果も期待している．

4. 評価実験

4.1 実験データ

実験データとして，ラテンアメリカ英語とアジア英語から成る 42.8 時間の非ネイティブ話者の英語音声（以後 AD データ（Accented Data）と呼ぶ）を用いる．ラテンアメリカ英語データ（以後 LA データと呼ぶ）は話者数 94 名による 20.7 時間の音声であり，アジア英語データ（以後 AE データと呼ぶ）は話者数 96 名による 22.1 時間から構成される．このうち 38.8 時間をモデルの学習に用い，残りの 5 時間はヘルドアウトセットとする．発話内容は，数字，アルファベットの孤立発話，コマンド発話，対話システム向けの短い発話音声などである．

声質変換として話者同一性と感情表現のバリエーションがよくなるように実験的に生成された 7 つの AD データ変換音声を用いる．その後，モデル適応実験のベースラインとして用いる汎用的音声認識システム（後述）でデコードした結果，発話単位の WER が 50%以下となる発話をランダムにピックアップし，オリジナルの AD データと加えて合計 114 時間（元の AD データの約 3 倍）になるように構成する．以後，声質変換された AD データを VT（Voice transformed）データと呼ぶ．話速変換は，0.9〜1.1 の間で変換倍率をランダムに選択し，リサンプリングを行うことによって変換音声を得る．これについても，オリジナルの AD データと加えて，データサイズが合計 114 時間になるように調節した．以後，話速変換された AD データを Speed データと表記する．最後に雑音付加は，DEMAND データベースから 12 種類の雑音を利用し，AD データに付加する形で新たな音声データを得る [11]．雑音の種類はレストラン，家屋，オープンスペース，会議室，バス，電車，車内などである．各雑音について約 10 時間分をデータ拡張に用い，AD データを合計 120 時間に増加させた．以後，雑音付加された AD データを Noise と表記する．

教師あり・なしのモデル適応の実験においてベースラインとなる音響モデルは，ネイティブ話者による 1200 時間の音声データを雑音付加によるデータ拡張で 3600 時間に増大させたもので学習した．オリジナルの 1200 時間分は，BN コーパスから 420 時間，Mixer 6 から 280 時間 [12]，AMI コーパスから 100 時間 [13]，プライベートデータから 450 時間で構成し，それを JEIDA 雑音データベース提供の雑音 [14] と，RWCP 提供の環境インパルス応答データ用いて拡張した [15]．雑音付加時の SNR は 5〜20dB とした．

4.2 音響モデル

実験で用いた音響モデルはすべて CNN（Convolutional Neural Network）に基づくモデルである．CNN に基づくベースライン音響モデルを上述のデータで学習する．音響特徴量は 40 次元の対数メル周波数スペクトル係数に 1 次と 2 次の動的特徴量を付加した 120 次元のベクトルとした [16]．対数メル周波数スペクトルは，フレーム窓長 25ms でシフト幅 10ms 毎に抽出する．そして，対数メル周波数スペクトルについて平均・分散正規化を行った後，前後 5 フレームを連結して計 11 フレームからなる特徴量に拡張する．

畳み込み層は隠れノード数 128 と 256 を持つ 2 層からなり，畳み込み層の後にノード数 2048 の全結合層を 4 層追加する．出力層は前後 2 音素のコンテキスト依存決定木に対応する 9300 ノードを持つ．畳み込み層の第一層は，前述の特徴量を入力とするサイズ 9×9 の畳み込みフィルタか

ら構成される．第 2 層はサイズ 3×4 のフィルタを持ち，第 1 層と第 2 層の両方で，マックスプーリングを行う．畳み込み層 2 層目の非線形出力は，全結合層へと接続される．本実験では，全ての隠れ層において活性化関数にシグモイド関数を採用する．なお，音声認識がサポートする語彙数は約 250K 単語であり，言語モデルは 200M のエントリーを持つ 4-gram LM である．

4.3 Weight-decay に基づく教師ありモデル適応

教師あり適応に基づく実験は，[17] で提案された方法に改良を加えたものを採用した [18]．このスキームは MAP 適応と類似しており，適応前のベースラインモデルと適応データからの更新量との重み付き結合で新しいモデルを構築する．[17] で検討されているような話者単位の適応ではなく，適応データプール全体を用い，次式のようのドメイン適応の枠組みで実行する．

$$\Delta \mathbf{w}_t = -\alpha \nabla_{\mathbf{w}} E(\mathbf{w}_t) - \beta(\mathbf{w}_{t-1} - \mathbf{w}_0), \qquad (3)$$

ここで，α は学習係数，β は正則化パラメータである．そして $E(\mathbf{w})$ は誤差関数，$\mathbf{w}_0$ は適応前のモデルのパラメータを表す．ネットワークはクロスエントロピー基準で適応学習される．

4.4 Teacher-student 学習に基づく教師なしモデル適応

Teacher-student 学習は，複雑な構成の教師ネットワークの振る舞いを，コンパクトかつシンプルな生徒ネットワークに転移学習させることができるフレームワークである [19]．一般的なニューラルネットワーク音響モデルの構築には，（正確な）書き起こしに基づくハードターゲットが用いられる一方，Teacher-student 学習では，ハードターゲットの代わりに次の損失関数が定義される．

$$\mathcal{L}(\theta) = -\sum_i q_i \log p_i, \qquad (4)$$

ここで q_i は教師モデルから生成されるソフトラベルであり，擬似ラベルとして用いられるものである．p_i は生徒モデルの音素コンテキストクラス i の出力確率である．各学習サンプルのソフトラベル q_i は，値は小さいものの競合するクラスにおいて 0 ではない確率値を持つ．VGG や Resnet のような強力なネットワークを一度学習すると，この枠組みで教師ネットワークの振る舞いを模倣する生徒ネットワークを構築することができる．Teacher-student 学習では，教師ネットワークから生成されるソフトラベルを生徒ネットワークのターゲットデータとして利用するので，対応する書き起こしデータは必ずしも必要ない．これは，Teacher-student 学習の枠組みで教師なし適応を実現可能なことを意味している．本報告では，適応データを教師ネットワークに入力して得られるソフトラベルを用いて，CNN ベースラインモデルを非ネイティブアクセント音声のドメインに適応する．

5. 実験結果

5.1 スクラッチからのモデル学習

本節では，AD データとその拡張セットのみで構築した音響モデルの比較を行う．学習はランダムな初期化による重みからスタートし，書き起こしには人手で用意したものを利用する．音響モデルはラテンアメリカ，アジア英語について別々に学習する．ここで用いる学習データサイズ（非ネイティブ話者の音声）は，近年の実用的な音響モデル構築のために用いられるセットと比べて極めて小さい．しかし，AD データがマイナー言語であると仮定すると，20 時間程度の音声＋書き起こしの組み合わせは現実的な仮定と言える．ここでは，まずは適応学習ではなく，スクラッチからモデルの構築を行うことを検討する．

表 1 はラテンアメリカ，アジア英語の専用モデルを，それぞれ AD データのみおよび AD+拡張データで構築した結果である．表 1 に示すように，スクラッチからの音響モデル構築では，声質変換と話速変換が共に大きな改善を与えていることがわかる．雑音付加による拡張データはノイズロバストを目指したシステムでは極めて大きな改善を与えるが，今回の設定では性能改善に貢献しなかった．すなわち，声質変換や話速変換が単に学習データのサイズによる寄与ではないことが示唆される．

表 1 スクラッチから学習した音響モデルの比較

Augmentation Scheme	LA Training and test WER	Asian Training and test WER
AD Baseline	23.20	26.18
AD+Noise	29.75	31.64
AD+VT	19.13	18.99
AD+Speed	17.57	18.00

5.2 教師ありモデル適応

次に，ネイティブ話者の音声で構築されたベースライン音響モデルを，教師あり適応で非ネイティブ音声にマッチさせることを検討する．ここで，音響モデルの適応はラテンアメリカとアジア英語で個別に行う．ベースラインモデルは英語のネイティブ話者音声のみで構成される 3600 時間の学習データから構築されているので，比較的頑健なシステムであると言える．（ネイティブ話者によるクリーンな読み上げ+自由発話発声の WER はおよそ 2〜10%である．）

しかしながら，表 2 に示すとおり，ベースラインシステムの非ネイティブ音声に対する性能は悪い．非ネイティブ話者による英語発声が如何に音響的変形を含んでいるかが伺える．その一方で，わずか 20 時間程度の人手による

書き起こし付き音声を用いて適応処理するだけで，誤り率は約半分になる．ここからさらに話速変換によるデータ拡張処理を行うことで，1〜3%の相対的誤り削減が実現できる．声質変換データによる性能は本節の実験ではほぼ変わらず，ラテンアメリカのテストセットでごくわずかな改善が得られるにとどまった．雑音付加による方法は劣化する結果となった．

表 2 教師ありモデル適応による音響モデルの比較

Augmentation Scheme	LA Supervised Adaptation WER	Asian Supervised Adaptation WER
Unadapted Baseline	28.30	26.70
AD	14.25	13.69
AD+Noise	15.42	14.95
AD+VT	14.22	13.85
AD+Speed	14.06	13.29

5.3 教師なしモデル適応

次に，教師なし適応による改善を見ていく．ここでも，モデルの適応はラテンアメリカとアジア英語で個別に行う．Teacher-student 学習によるモデル適応において，教師ネットワークは，ベースライン CNN と同じネイティブ話者による 3600 時間のデータで学習した VGG ネットワークとした．学習済みの教師 VGG ネットワークに AD データを入力し，結果として得られるソフトラベルを生徒 CNN の適応処理に用いる．教師と生徒ネットワークの出力層の構成は同じである．

教師 VGG ネットワークは 10 層の畳み込み層から構成され，マックスプーリング層が 3 層おきに挿入されている．そして，ベースライン CNN と同様に，最後の畳み込み層の後にノード数 2048 を持つ 4 層の全結合層を追加している．ここでは，活性化関数として全隠れ層で ReLU 非線形関数を利用している．また，バッチ正規化処理を全ての全結合層で行っている．VGG ネットワークは交差エントロピー基準によって学習した後，さらにシーケンストレーニングによる追加学習を行っている．教師と生徒ネットワーク間で（出力層以外の）トポロジーの違い，活性化関数の違いや，サンプリング周波数の違いがあったとしても，生徒ネットワークはうまく教師の振る舞いを学習できることがわかっている [19]．Teacher-student 学習における事後確率は上位 50 個のみを用いた．本節の実験では，声質変換と話速変換のみの比較を行い，雑音付加による結果を省いた．実験結果を表 3 に示す．表から声質変換と話速変換によるデータ拡張で一貫した結果が得られていることがわかる．ここでは，AD データのみで適応した場合と比べて，2.8〜4% の相対的改善が得られた．

表 3 教師なしモデル適応による音響モデルの比較

Augmentation Scheme	LA Unsupervised Adaptation WER	Asian Unsupervised Adaptation WER
Unadapted Baseline	28.30	26.70
AD	24.75	22.03
AD+VT	23.74	21.68
AD+Speed	23.89	21.37

5.4 複数のデータ拡張法の利用

前節までの実験結果から，声質変換と話速変換が外国語なまりのあるアクセント音声に対して有効なデータ拡張法であることを見出した．本節では，それらが相補的な関係にあり，さらなる改善が得られるかどうかを検証する．雑音付加によるデータ拡張は効果が見られないので検証から省くこととし，スクラッチからの音響モデル学習での効果を比較する．

Merged training:

声質変換（VT）データと話速変換（Speed）データをそれぞれ半分のサイズに分割し，AD データと組み合わせて新たな 114 時間のデータセット（AD + 50% VT + 50% Speed）を構成する．つまり，前節までの実験と同等のデータサイズである．しかしこの場合，話速変換単体の性能と比べて WER が悪く，LA テストセットで 18.24%，AE テストセットで 19.69%となった．一方で，全ての拡張データを用いた場合は，話速変換単体と比べて性能改善があり，それぞれ LA テストセットで 17.39%，AE テストセットで 17.71%となった．

Posterior combination:

次に 5.1 節で構築した音響モデル（AD+VT と AD+Speed）について，0.25 vs 0.75 の結合重みで事後確率レベルのシステムコンビネーションを試した．その結果，LA テストセットで 17.34%，AE テストセットで 17.57%となり，AD+Speed 単体と比較して性能が上回る結果となった．

以上までのように，複数のデータ拡張処理が非ネイティブ話者による音声の認識性能改善に有効であることを示した．特筆すべきは，話速変換が様々な条件下で顕著な改善を示した点である．

6. おわりに

本報告では，データ拡張処理がラテンアメリカ英語とアジア英語の音声認識に顕著な効果があることを示した．評価実験は Weight-decay に基づく教師あり適応，Teacher-student 学習の枠組みを利用した教師なし適応，スクラッチからのモデル構築の 3 つのシナリオで行い，各種データ拡張処理の効果を比較した．これらの実験において，1) 話速変換が最も効果的な手法であること，2) 声質変換は性能改善の助けになるが，改善度合いはそれほど顕著ではな

く，またケースに依存すること，3) 雑音付加はあまり効果がなく，劣化の方向にさえ動く可能性があることを確認した．これらの実験は事前に非ネイティブ話者が発話するタイミングがわかるという前提のもと，全てのケースで非ネイティブ専用モデル（ネイティブ話者の認識性能は意識しない）として構築した音響モデルにより比較実験を行った．データ拡張の効果はスクラッチからモデル構築を行う場合において顕著であり，ネイティブ話者の音声のみで構築したベースラインシステムと比較して，相対的誤り削減率が 30%を超えるケースもあった．これと同時に教師あり・なし適応の枠組みでも改善があることを確認したが，スクラッチからモデル構築を行う場合と比べて改善は少なかった．

参考文献

[1] Crystal, D.: *English as a Global Language*, Cambridge University Press (2003).
[2] Ko, T., Peddinti, V., Povey, D. and Khudanpur, S.: Audio augmentation for speech recognition., *INTERSPEECH*, pp. 3586–3589 (2015).
[3] Hartmann, W., Ng, T., Hsiao, R., Tsakalidis, S. and Schwartz, R. M.: Two-Stage Data Augmentation for Low-Resourced Speech Recognition, *INTERSPEECH*, pp. 2378–2382 (2016).
[4] Cui, X., Goel, V. and Kingsbury, B.: Data augmentation for deep neural network acoustic modeling, *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, Vol. 23, No. 9, pp. 1469–1477 (2015).
[5] Ragni, A., Knill, K. M., Rath, S. P. and Gales, M. J.: Data augmentation for low resource languages, *Fifteenth Annual Conference of the International Speech Communication Association* (2014).
[6] Deng, L., Acero, A., Plumpe, M. and Huang, X.: Large-vocabulary speech recognition under adverse acoustic environments., *INTERSPEECH*, pp. 806–809 (2000).
[7] Fernandez, R., Rosenberg, A., Sorin, A., Ramabhadran, B. and Hoory, R.: Voice-Transformation-Based Data Augmentation for Prosodic Classification, *Proc. ICASSP*, New Orleans, Louisiana, USA, pp. 5530–5534 (2017).
[8] Sorin, A., Shechtman, S. and Rendel, A.: Semi Parametric Concatenative TTS with Instant Voice Modification Capabilities, *Interspeech*, Stockholm, Sweden, pp. 1373–1377 (2017).
[9] Fant, G., Liljencrants, J. and Lin, Q.: A Four-Parameter Model of the Glottal Flow, *STL-QPSR*, Vol. 26, No. 4, pp. 1–13 (1985).
[10] Pearce, D. and Picone, J.: Aurora working group: DSR front end LVCSR evaluation AU/384/02, *Inst. for Signal & Inform. Process., Mississippi State Univ., Tech. Rep* (2002).
[11] Thiemann, J., Ito, N. and Vincent, E.: DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments, *Proceedings of Meetings on Acoustics* (2013).
[12] Brandchain, L.: The Mixer 6 Corpus: Resource for Cross-Channel and Text Independent Speaker Recognition, *LREC* (2010).
[13] Carletta, J.: Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus, *Language Resources and Evaluation*, Vol. 41, No. 1, pp. 181–190 (2007).
[14] Itahashi, S.: Recent Speech Database Projects in Japan, *Proc. ICSLP* (1990).
[15] Nakamura, S., Hiyane, K., Asano, F., Nishiura, T. and Yamada, T.: Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition, *LREC* (2000).
[16] Fukuda, T., Ichikawa, O. and Nishimura, M.: Long-term Spectro-temporal and Static Harmonic Features for Voice Activity Detection, *IEEE Journal of Selected Topics in Signal Processing*, Vol. 4, No. 5, pp. 834–844 (2010).
[17] Liao, H.: Speaker adaptation of context dependent deep neural networks, *Proc. IEEE ICASSP*, pp. 7947–7951 (2013).
[18] Suzuki, M., Tachibana, R., Thomas, S., Ramabhadran, B. and Saon, G.: Domain Adaptation of CNN based Acoustic Models under Limited Resource Settings, *Proc. Interspeech*, pp. 1588–1592 (2016).
[19] Fukuda, T., Suzuki, M., Gakuto, K., Cui, J., Thomas, S. and Ramabhadran, B.: Efficient Knowledge Distillation from an Ensemble of Teachers, *Proc. Interspeech*, pp. 3697–3701 (2017).