DDSupport: モデルとなる発音との差異・距離を提示する言語学習支援システム

河村 和紀[1,a)]　暦本 純一[2,1,b)]

概要：母語ではない言語を学ぶ際，自分がうまく話せているかどうかを学習者自身で判断することは難しい．さらに，自分がうまく話せていない場合に，自分の発音がネイティブスピーカーのようなモデルとする発話者の発音とどこがどの程度異なっているのかを判断することも困難である．そこで，本研究ではシステムとのインタラクションを通じて一人で言語のスピーキングを学ぶことのできる新しい言語学習支援システムを提案する．提案システムでは，深層学習に基づく音声処理を利用し，ユーザの発話がうまくいっているかどうかを判断し，学習者の発音がモデルの発音とどこが違うのか，どのくらい異なるのかを視覚的に分かりやすく提示する．学習者がシステムに示されたモデルとの差異を解消し，距離を近づけるように発音の修正を繰り返すことで，発音が徐々に改善されていくことが期待される．また，英語を母語としない学習者が英語を学習するアプリケーションを構築し，ユーザがアプリケーションを使用することで発話の分かりやすさが向上することを確認した．

1. はじめに

母語ではない第二言語の学習，特にスピーキングの学習は，語彙，文法，音韻，アクセントなど多くの点が母語とは異なるため，困難を伴う．新しい言語を効率よく学ぶためには，学習者は自分と学習対象の言語のネイティブスピーカーとの発音の違いを理解する必要がある．学習者が違いを理解するための最も一般的な方法は，ネイティブスピーカーや専門家が学習者のスピーチがどれだけ聞き取れるかを評価するディクテーションである [3], [7]．しかし，この手法は多くの時間と労力を必要とするため，本研究ではこの方法を自動化することを目的とする．本研究では，自分の発音がネイティブスピーカーの発音に近いかどうかを学習者が一人で判断することができる新しいシステムであるDDSupport を提案する．さらに，このシステムではユーザの発音とネイティブスピーカーの発音のどこが違うのか，どれくらい離れているのかを可視化することで，ユーザの言語学習を支援する．

本システムと同じようにコンピュータを利用することで発音の学習を支援するシステムは CAPT (Computer-Aided Pronunciation Training) と呼ばれ，盛んに研究がおこなわれている [14]．これらの既存の研究では，学習者の発音とネイティブスピーカーの発音の違いを，イントネーション，リズム，音素などの何らかの指標で詳細に計算している．しかし，このような指標での違いの表示は初学者にとって理解が困難であったり，特定の側面を修正したとしても，学習者の発音が分かりやすくなるとは限らないといった問題がある．そこで，我々の提案手法では，学習者とネイティブスピーカーの発音の違いをそのような特定の側面で評価しない．その代わり，学習者の発音とモデルの発音との差異や距離を，直感的かつ視覚的に分かりやすく学習者に表示する．図 1 のようにユーザの発話とお手本の発話を波形として並べて表示し，ユーザの発話がお手本の発話と異なる部分を波形上に赤く強調して表示する．また，ユーザの発話とお手本の発話をそれぞれ 2 次元座標上の点で表し，自分とお手本の発話がどの程度離れているか確認することができるようにする．学習者はシステムによって示された差異を無くすように，距離を縮めるように発音を修正することで発話がモデルに近づいていくことが期待される．本研究では，「この部分のイントネーションをこのように変えてください」というように，ユーザに何を修正すればよいかを事細かに伝えるのではなく，ユーザ自身の耳や目でどのように修正すればよいかを考える余地を残すことで学習者が感覚的に言語を習得できるようなシステムの構築を目指す．

[1] ソニーコンピュータサイエンス研究所京都研究室
[2] 東京大学大学院情報学環・学際情報学府
[a)] Kazuki.Kawamura@sony.com
[b)] rekimoto@acm.org

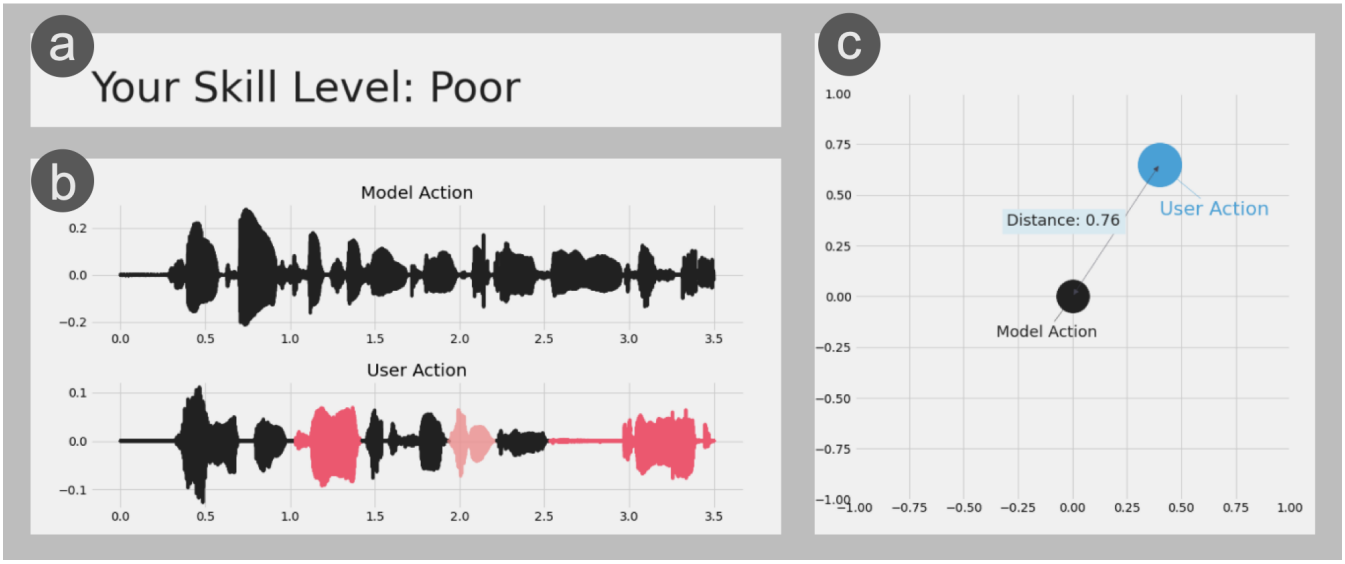


図 1: DDSupport システム概要

本システムは，深層学習の技術を用いて学習者の発音と学習対象の言語のネイティブスピーカーの発音を識別する．その上で，識別スコアを用いて学習者の発話がネイティブスピーカーのものに十分近いかを判定する．このような深層学習の識別器を構築するためには，通常ネイティブスピーカーの音声データだけでなく，学習者の言語体系に近い言語の非ネイティブスピーカーの音声データが大量に必要である．ネイティブスピーカーの発話データはラジオ，テレビ，動画配信サービスを通じて容易に収集することができる．一方で，非ネイティブスピーカー，すなわち，学習者の発話データを収集するのは非常に困難である．そこで，提案システムでは収集することの容易なネイティブスピーカーの音声データを効率的に利用するため近年の自己教師あり学習技術 [15] を使用し，非ネイティブスピーカーの発話データが少量しか手に入らないときであってもシステムを構築することを可能にする．非ネイティブスピーカーの発話データに，どこがネイティブスピーカーの発話と異なっているかというラベルや，ネイティブスピーカーの発話とどの程度異なっているかというスコアのラベルがあれば，教師あり学習の技術を用いることで，学習者とモデルの発音の差異や距離を算出することは容易である．しかし，このようなラベルが付与されたデータはほとんど存在しないため，このようなラベルがない場合でも差異や距離を算出するために深層学習モデルの判断基準を可視化することのできる attention とデータから距離関数を設計することのできる距離学習の技術を応用する．このように，本システムは大量のモデルの行動データと少量の初学者の行動データがあれば実装することができるように設計されているため，言語学習に閉じず様々な技能に関しての初学者の技能判定や技能獲得支援をおこなう技術基盤につながるものとなっている．

提案システムの技能獲得支援の効果を検証するために，英語を母語としないユーザが英語のスピーキングを学習するアプリケーションを実装した．このアプリケーションを実際にユーザに使用して発音の学習をおこなってもらい，ユーザに対してのアンケート評価と英語を母語にする評価者によるユーザの学習後の発話の評価をおこなった．評価の結果，提案手法が第二言語のスピーキング学習において，学習者の補助になり，学習者の発音の分かりやすさを向上させる効果があることが分かった．

以下に，本研究の貢献を示す．

- 学習者の発音がモデルの発音に近いかどうかを判定し，学習者の発話がモデルの発話とどこが違うのか，どのくらい違うのかを可視化するシステムを提案する．
- 英語を母語としない学習者が英語のスピーキングを学習するアプリケーションを構築し，学習を通じて学習者の発音が向上することを確認した．

2. 関連研究

2.1 第二言語学習

提案システムは，第二言語（L2）の音声学習を支援することを目的としている．第二言語学習における発音指導の分野では，何を目標にするかに関して大きく分けて二つの考え方がある [23]．一つ目は，対象言語のネイティブスピーカーにできる限り近い発音を身につけることを目標とするというものであり，二つ目は，多少外国語訛りがあったとしても，聞き手に分かりやすい発音であることを重視するものである．[27] では，この分かりやすさ（intelligibility）を，「聞き手が話し手のメッセージをどれだけ理解できるか」と定義している．昨今，対象の言語を話す多様なバックグラウンドを持つ人々との円滑にコミュニケーションするためには，特定のネイティブスピーカーの発音に限りな

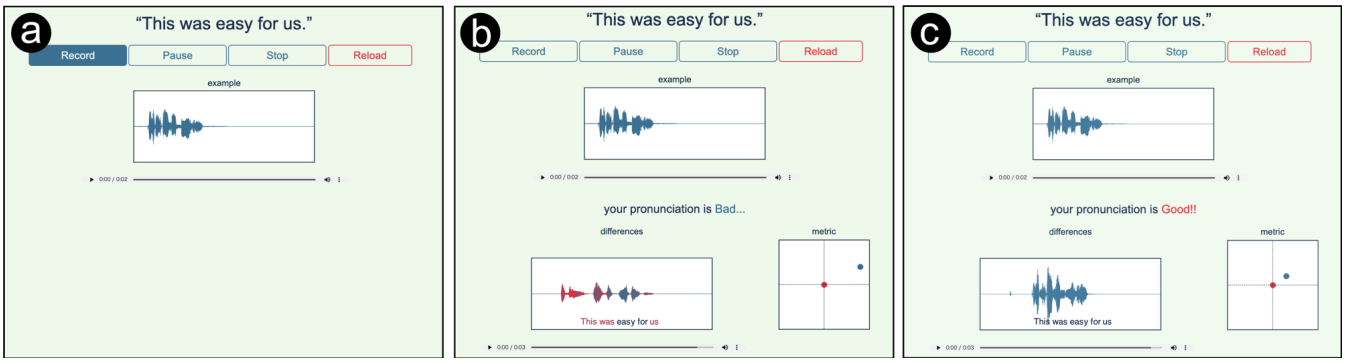


図 2: 言語学習支援システム画面

く近い発音よりもむしろ，様々な人に対して分かりやすく発音することが不可欠であると考えられている [19], [26]．つまり，特定のネイティブスピーカーの話し方を完璧に真似るのではなく，コミュニケーションに支障をきたさない範囲で本来の言語の音を学習者にとって発音しやすい音に置換してもよいという考え方である [8], [17]．この研究では後者の考え方に従い，対象の言語を第二言語として学ぶ学習者がその言語で円滑にコミュニケーションできるようになることを支援するシステムの構築を目的としている．

2.2 言語スキル評価

学習者が対象言語の分かりやすい発音を身につけるためには，分かりやすさをどのように評価すべきかが問題となる．最も一般的な評価方法はディクテーションと呼ばれるもので，ネイティブスピーカーや専門家が学習者の音声がどの程度聞き取れるかを評価するものである [3], [7]．この方法では評価に多くの時間と労力を要するため，本システムではこのプロセスを自動化する．また，2.1 節で述べたように我々の目的はコミュニケーションツールとしての第二言語の学習であり，対象言語の特定のネイティブスピーカーを完璧に真似ることを目的としているわけではない．そのため，本システムでは学習者の対象言語の音声が様々な背景を持つネイティブスピーカーの平均的な音声とどのように異なるかを可視化するだけで，イントネーション，リズム，音素などの点でどの程度異なるのかということを事細かに提示することはしない．このようにすることで学習者は自分の裁量でモデルを真似る度合いを調整し，自分の好みに合わせてスピーキングを学ぶことができるのである．

2.3 発音学習支援システム

CAPT (Computer-Aided Pronunciation Training) [14] すなわちコンピュータを利用することで発音の学習を支援するシステムは，言語学習を自動化することを目的にこれまで盛んに研究開発が行われてきた．特に近年では，様々な音声情報処理の分野で高い性能を示している深層学習の技術 [11], [35] を用いて発音のミスを検出するシステムも提案されている [16]．しかし，これらの手法の多くは，同じ文章に対して対象言語の学習者とネイティブスピーカーの発話のペアを必要とする．一方，我々の手法は発話が一致している必要はなく，単に学習者の発話とネイティブスピーカーの発話のデータがあれば構築することができる．つまり，我々のシステムではデータセットにない文も含めて学習したい文を学習者が自由に選ぶことができるのである．

従来の CAPT の研究は，音声情報処理技術を応用することでいかに正確にユーザの音声と見本となる音声の差異を計算するかに重きを置いており，それらが実際に発話学習に効果的かということやどのように表示すると学習に効果的かという議論はあまりされてこなかった [10], [22]．一方，インタラクションを重視する CAPT システムは言語学習と HCI の複合領域で研究がおこなわれるようになってきている．例えば，Tip Tap Tones [9] や SIAK [18] はユーザがゲームを通じて言語学習をおこなうシステムです．PTeacher [4] は，ユーザーに誇張された音声・映像によるフィードバックをし，Seiyuu-Seiyuu [6] はビデオ学習インターフェースを介して実用的な能力の開発を促進します．特に，Robertson ら [30] は，発音エラー検出の性能に関係なくユーザの発音は向上することを報告し，CAPT の研究は、工学的な問題ではなく、HCI の問題として扱うのが有益であると主張している．我々の手法の CAPT への応用も同様にして，システムとインタラクションを通じて言語学習することを重視しており，ユーザ実験を通して CAPT を HCI の問題として捉えることでよりユーザにとって学習効果の高い技術になることを裏付ける一つの結果を示す．

3. DDSupport

3.1 概要

このシステムは，学習者の発話がネイティブスピーカーの発話に近いかどうかを判断し，そうでない場合は，学習者の発話とネイティブスピーカーの発話との差異や距離を表示することで，母語ではない言語のスピーキングの学習

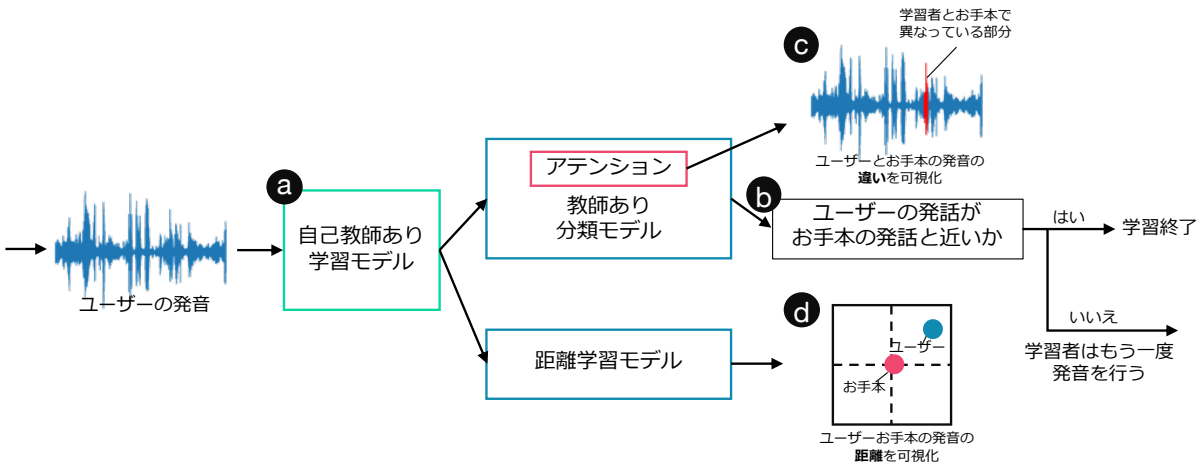


図 3: システム構成

を支援する．まず，図 2（a）に示すように，学習したい文とその文のモデルの発話をシステムが提示する．学習者はこの発音を聞いて，対象の文章をどのように発音すれば聞き手が理解しやすいかを確認する．次に，提案システムはユーザの発話を分析し，図 2（b）のように学習者の発話とネイティブスピーカーの発話の差異と距離を提示する．図 2（b）中に示すようにユーザの発話の波形とテキスト上では，その差異が赤で示されており，この赤い部分が濃いほど，モデルの発話との差が大きいことを示している．また，距離に関しては学習者の発話とネイティブスピーカーの発話それぞれをセンテンス単位で 2 次元座標上の点として表現する．ここで，赤色の点がネイティブスピーカーの音声，青色の点が学習者の音声で，これらの点の間の距離が大きいほどユーザの発話がネイティブスピーカーの発話と異なることを意味する．ユーザは，システムに示される差異を消すように，距離が縮まるように自分の発音を修正する．図 2（c）にはユーザが発話を修正したあとの例を示しており，音声を修正することで差異が小さくなり，距離が縮まっていることが分かる．このようにして音声を学習することで，ユーザの発話がネイティブスピーカーの発話に近づくことが期待できる．

3.2 システムを用いた言語学習

ここでは，ユーザがシステムを用いた学習により対象言語をうまく話すことができるようになるまでのプロセスについて説明する．図 5 は，ユーザが音声の修正を繰り返すことで，ユーザの発話とモデルの発話が近づいていく様子を示している．システムはユーザの発話を分析し，学習者の発話とモデルの発話との差異や距離をユーザに提示する．差異については，図 5(a) に示すようにユーザの発話の波形とテキストにおいて，ネイティブスピーカーと異なる部分が赤く示される．この例では，ユーザが最初に “this was easy for us” というセンテンスを発話したとき，“this”，“easy”，“for” の発音がネイティブスピーカーと異なっていることを表している．システムから指摘された赤い部分を修正するためにユーザがセンテンスを発音し直すと，“easy” と “for” の発音が改善される．次に，“for” の発音が，さらに，“this” の発音を修正すると，最後の図に示されているように，ユーザの発音とモデルの発音の差がなくなる．この時点で，このセンテンスのスピーキングの練習は終了となる．我々のシステムで発音の違いを文字上だけでなく波形上でも表示する理由は，違いがどこにあるかをより正確にユーザに示すためである．文字ベースの差異の表示では，表示の単位が最小レベルでも音節単位となる．しかし，もっと細かいレベルでの違いや単語間での違いがある場合もあり得る．そのようなわずかな違いをユーザーが知りたいこともあると考え，波形を採用して差異を表現している．

図 5(b) は，ユーザが発話の修正を繰り返すことでユーザの発話とモデルの発話の距離が近づいていく様子を表している．ネイティブスピーカーの発音と学習者の発音は，図の二次元座標上の点で表現され．赤い点はネイティブスピーカー，青い点は学習者の発音である．これらの点の間の距離が大きいほど，ユーザの発話がネイティブスピーカーの発話と異なることを意味する．ユーザはこれらの点の間の距離が近くなるように発音を修正する．ここで，なぜ「差異の可視化」だけでなく「距離の可視化」が必要なのか説明する．差異の可視化は，主に差異がどこにあるかをユーザに伝えるためのものである．既存の多くの手法のように，イントネーション，リズム，音素などの細かい要因で差異を評価すれば，ユーザが発音を変化させるたびにどのように発音が変化したかをある程度把握することはできる．しかし，学習者はこれらの特定の側面で発音を矯正しても必ずしも分かりやすさが向上するわけではないため，我々のシステムではこのような指標を用いていない．そうするとユーザは発音を変えたときに自分の発音がどのよう

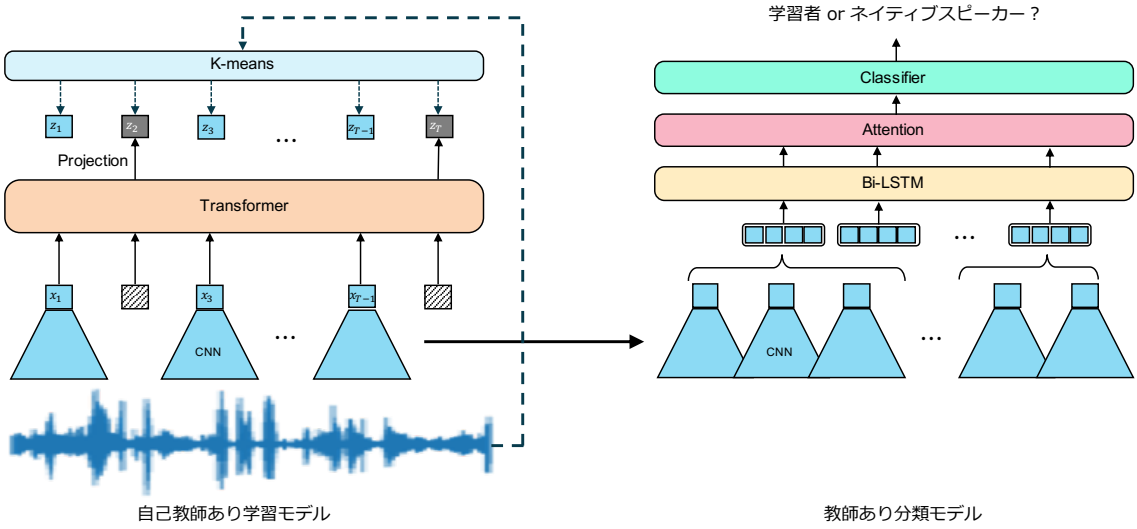
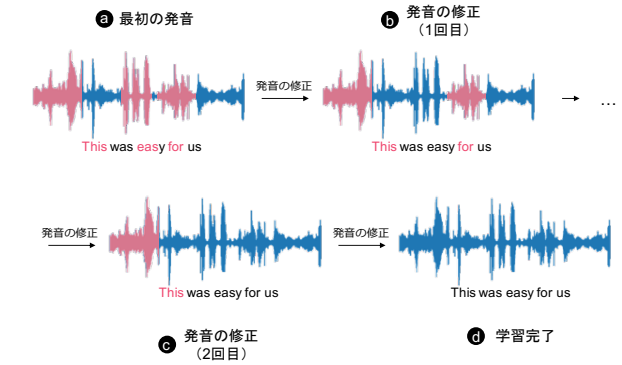
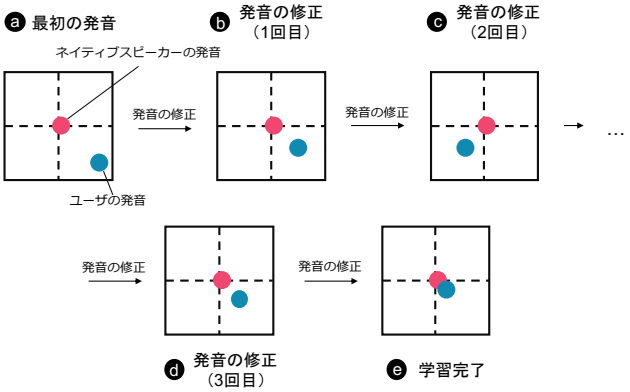


図 4: 表現学習及び音声分類アーキテクチャ

に変化したのかを，テキストや波形のみから判断する必要があるが，これらだけではユーザの音声がどのように変化しているのか十分な情報を得ることができない．そこで，ユーザの発話とネイティブスピーカーの発話をそれぞれ 2 次元上の点で表現すると，ユーザは発音を繰り返すたびに自分の発音の点がどのように変化するのか軌跡を視覚的に確認することで，発話の変化を理解することができる．例えば，ユーザが発話を繰り返すたびにある一定の直線や曲線に沿ってユーザの点がネイティブスピーカーの点に近づいていけば，その発音の修正方法は正しい，すなわちその修正方法で発音を修正していけば効率的に発音を修正することができることが分かる（図 5(b)ⓐ → ⓑ，ⓓ → ⓔ）．一方，発音を変えたときにそれまでの軌道上とは異なる場所に急に点が移動した場合は，間違った方法で発話を修正している可能性があることを示す（図 5(b)ⓑ → ⓒ）．



(a) モデルの発音との差異を消すプロセス



(b) モデルの発音との距離を近づけるプロセス

図 5: ユーザが発音を修正するプロセス

3.3 システム構成

本システムは，図 3 に示すように，4 つの主要な要素からなる．

(a) 差異や距離を計算するための音声表現を学習
(b) 学習者とネイティブスピーカーの音声を分類
(c) 学習者とネイティブスピーカーの音声の差異を計算
(d) 学習者とネイティブスピーカーの音声の距離を計算

3.3.1 表現学習

我々のシステムは機械学習，特に深層学習の技術を用いてユーザの音声とネイティブスピーカーの間の差異や距離を計算する．そのため，システムを構築するためには学習する対象の言語の非ネイティブスピーカーとネイティブスピーカーの両方の音声データが大量に必要となる．しかし，学習対象言語を話す非ネイティブスピーカーの音声データや学習者の発話データを収集することは困難であり，通常そのようなデータの大規模なデータセットは手に入らない．

一方，ネイティブスピーカーの音声データは，ラジオやテレビ，動画共有サイトなど様々な情報源から入手でき，かつ大規模なデータセットが多数存在する．そこで，本システムでは少量の非ネイティブスピーカーの音声データだけでシステムを構築するために，このような大量のネイティブスピーカーの音声データを有効に活用する．ネイティブスピーカーの音声データは，非ネイティブスピーカーもし

くは学習者の音声データとは異なるが，少なくとも他の言語の音声データよりは単語の発音や文構造は似通っている．そのため，大量のネイティブスピーカーのデータを使って学習対象の言語の一般的な音声表現を学習し，少量の非ネイティブスピーカーのデータを使ってそれを微調整することで，少量のデータでも非ネイティブスピーカーの良い音声表現を得ることができる．この良質な音声表現を用いることで，非ネイティブスピーカーの音声データが少量であってもネイティブスピーカーと非ネイティブスピーカーの音声の分類や，ネイティブスピーカーと非ネイティブスピーカーの差異や距離を算出することができる．

近年，音声データから音声表現を学習するための自己教師あり学習手法は急速に進化しており wav2vec [34] や wav2vec 2.0 [1]，最近新たに発表された HuBERT [15] など，効率的なモデルが多く生み出されている．本論文では，音声表現を学習するための最新の自己教師あり手法である HuBERT を用いて，学習対象の言語の音声に共通する表現を獲得する．図 4 左に示すように，このモデルは，CNN エンコーダー，Transformer，Projection 層の 3 つの主要要素から構成される．このモデルに音声波形を入力すると，まず音声波形を CNN を介して音声表現 $X = [x_1, \cdots, x_T]$ にマッピングする．この音声表現を Transformer に与えることで，音声全体の文脈表現 $Z = [z_1, \cdots, z_T]$ を得る．学習中は，この CNN を介して得られた特徴は一部がランダムにマスクされ，マスクされた時刻の文脈表現 z_t が，k-means [25] を用いて別途入力音声をクラスタリングしたクラスのいずれに属するかを Projection 層によって予測する．我々のシステムでは，ラベルのないネイティブスピーカーの音声データでこの HuBERT を学習した後，得られた文脈表現 Z を用いて次の音声分類タスクを行う．

3.3.2 音声分類

HuBERT を事前学習した後，このモデルを特徴抽出器として使用し入力音声データをネイティブの発話に近いか非ネイティブの発話に近いか二値分類する．ネイティブスピーカーの大量の音声データで学習された HuBERT を特徴抽出器として使用し，入力音声 X から音声表現 $Z = [z_1, z_2, \ldots, z_T]$ を得る．

得られた音声表現を各ステップ k 時刻ごとに結合して得られた時系列データ $Z' = [z'_1, z'_2, \ldots, z'_S]$ が，Bidirectional LSTM (Bi-LSTM) [13], [32] に入力される．ここで，各 z'_s は，Z を k 時刻ずつ結合したベクトル $z'_t = [z_{k(s-1)+1} \parallel z_{k(s-1)+2} \parallel \ldots \parallel z_{ks}]$ である．そして，Bi-LSTM から各時間ステップの隠れ状態 h_t を得る．

最終的にこれらの隠れ状態 $H = [h_1, h_2, \ldots, h_T]$ をソフトマックス分類器を用いて，以下のようにラベル $\hat{y} \in \{\text{native}, \text{non-native}\}$ を予測する．

$$\hat{p}(y \mid X) = \text{softmax}\left(W_c(H\alpha^{\mathrm{T}}) + b_c\right),$$
$$\hat{y} = \arg\max_y \hat{p}(y \mid X),$$

ここで W_c と b_c は学習可能なパラメータベクトルであり，α は 3.3.3 節で導入される attention の重みである．

3.3.3 差異の可視化

ここでは，attention 機構を用いてユーザとネイティブスピーカーの音声の違いを可視化する手法の実装について説明する．attention 機構は近年，質問応答，機械翻訳，音声認識などの様々なタスクで成功を収めており，これらの分野で attention 機構はモデルの性能向上だけでなく，ニューラルネットワークモデルの解釈性を向上させるためにもよく用いられている [2], [5], [12]．我々の手法では，音声データを分類する際にシステムがどこに注意を向けているかを可視化するために，単純な attention 機構を用いる．各タイムステップにおける音声の重要度を示す attention の重みは以下のようにして得られる：

$$\alpha = \text{softmax}\left(W_{a2} \tanh(W_{a1}H)\right),$$

ここで，W_{a1} と W_{a2} は学習可能なパラメータベクトルで，attention の重みは $[0, 1] \in \mathbb{R}$ の値をとる．図 6 のように，閾値 th よりも大きな値が得られたタイムステップでは，音声波形上で赤く強調して表示する．

3.3.4 距離の可視化

ここでは，ユーザとネイティブスピーカーの発音間の距離を可視化する手法の実装について説明する．ユーザとネイティブスピーカーの音声間の距離を計算する前にまず，音声間の距離を計算するのに最適な距離を設計する必要がある．本研究では，物体間の距離関数を学習する距離学習 [21] を用いて距離を設計する．今回，距離学習の中で最も主要なアプローチの一つである Triplet Loss[31] を使用する．Triplet Loss は以下の 3 つの音声のセットで学習される：

- 任意の音声サンプルである anchor
- anchor と同じクラスの音声サンプルの positive
- anchor と異なるクラスの音声サンプルの negative

Triplet Loss では，anchor 音声から positive 音声までの距離を最小にし anchor 音声から negative 音声までの距離が最大になるように学習される．我々のシステムでは，ネイティブスピーカーの音声を anchor 音声とした場合，positive 音声は対象のセンテンスの他のネイティブスピーカーの音声，ネガティブ音声はそのセンテンスの非ネイティブスピーカーの音声となる．逆に，学習者の音声を anchor 音声とした場合，positive 音声は対象のセンテンスの他の非ネイティブスピーカーの音声，negative 音声はそのセンテンスのネイティブスピーカーの音声となる．このようにして，ネイティブスピーカーとネイティブスピーカー，非ネイティブスピーカーと非ネイティブスピーカーの音声

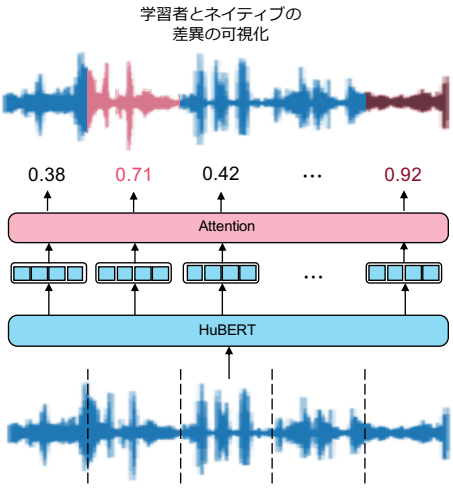


図 6: 差異の可視化アーキテクチャ

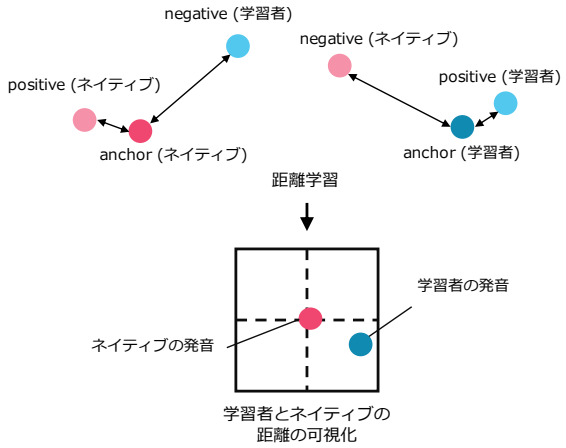


図 7: 距離の可視化アーキテクチャ

のペアは近づき，ネイティブスピーカーと非ネイティブスピーカーの音声のペアは遠ざかるように学習される．3.3.2 節で紹介した非ネイティブスピーカーとネイティブスピーカーの分類は音声データ全体の二値分類であるのに対し，今回の距離学習ではセンテンスごと，ネイティブスピーカー（非ネイティブスピーカー）ごとの音声を近づけるように学習する．これらの距離学習のための各入力音声は，音声特徴抽出器として 3.3.2 節で説明した事前学習済みの HuBERT を使用し，得られた特徴量は線形層に供給される．anchor 音声と positive 音声の間の距離を d_{ap}，anchor 音声と negative 音声の間の距離を d_{an} として，以下の Triplet Loss を最小化するように学習する：

$$L_{\text{triplet}} = [d_{ap} - d_{an} + m]_+ .$$

ここで m はマージンであり，anchor 音声と positive 音声の距離と anchor 音声と negative 音声の距離の差はコサイン類似度を用いて求める．距離学習が完了すると，図 7 の下に示すように，あるセンテンスに対するネイティブスピーカーの発話とユーザの発話の平均値を点で描画するごとでユーザにネイティブスピーカーとの距離を示す．

4. 実装

提案したシステムを評価するために，英語を母語としない日本人が英語（アメリカ英語）のスピーキングを学習するのを支援するシステムを構築した．

4.1 データセット

英語を発話している音声データの普遍的な表現を得るために，表現学習用の英語会話の音声データとして TIMIT Acoustic-Phonetic Continuous Speech Corpus[*1] を使用した．TIMIT には 6300 文の発話データが収録されており，米国の 8 つの主要地域から集まった 630 人の話者がそれぞれ 10 文ずつ発話している．また，学習者の音声と英語を母語とする人の音声を分類したり，学習者の音声と英語を母語とする人の音声の差異や距離を計算するためのデータセットとして UME-ERJ（English Speech Database Read by Japanese Students）[*2] を用いる．UME-ERJ には，英語を母語としない日本人学生の発話と，英語を母語とする人の発話が含まれおり，日本人 202 人，英語を母語とする人 20 人が合計 806 文を発話している．すべての音声ファイルは，シングルチャンネル 16000Hz のフォーマットに標準化し，データ長は 4 秒に揃えた．

4.2 詳細設定

表現学習には，FAIRSEQ [28] による HuBERT の実装を使用し，デフォルトの設定に従った．このモデルを TIMIT データを用いて学習し，英語の一般的な音声表現を獲得した．ユーザの発話とモデルの発話の差異を可視化するための attention 層は 32 次元の隠れ層を持つものを，距離を可視化するための距離学習には，128 次元の隠れ層を持つ Bi-LSTM と，512 次元の隠れ層を持つ線形層を使用した．また，分類モデルと差異・距離表示モデルの学習には，Dropout [33] を 0.5 の確率で使用することで過学習を抑制した．KaME-ERJ は英語を母語とする人の音声データよりも日本人の音声データの方が多いという不均衡データである．そこで，不均衡の影響を軽減するために誤差関数として Focal Loss [24] を使用した．最適化アルゴリズムには Adam [20] を用い，バッチサイズは 32，学習率は 10^{-4} とした．さらに，ノイズへの耐性を高め，データ不足を解消するために，以下のデータ拡張を行った．

- 背景にガウスノイズを加える
- 音声の音量をランダムに増減させる
- 音高をランダムに上下させる
- ランダムに選択した音を消す

*1 https://catalog.ldc.upenn.edu/LDC93S1
*2 http://research.nii.ac.jp/src/en/UME-ERJ.html

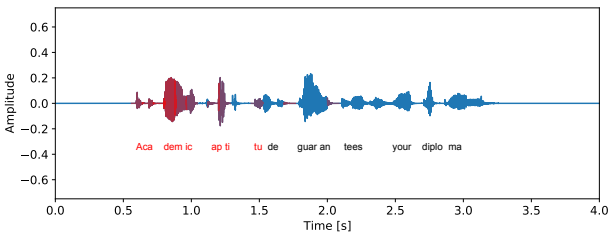


図 8: 差異の可視化の例

4.3 技術評価

まず，非ネイティブとネイティブの発話音声の分類に関する評価について述べる．UME-ERJ の音声サンプルを，同一の話者が混ざらないように 8 : 1 : 1 で訓練セット，検証セット，テストセットに分け，テストセットで評価を行った．非ネイティブとネイティブの発話の分類精度は F-measure で 0.99 であった．

ここでは，非ネイティブとネイティブの音声の差異を可視化するシステムの評価について述べる．UME-ERJ で評価するためにあらかじめ選んでおいた 50 の日本語発話サンプルを 15 人の英語を母語とする人に評価してもらい，システムが指摘した部分と比較した．英語を母語とする人には，日本語音声の中でネイティブスピーカーと発音が異なると感じた部分を，音節レベルで指摘してもらった．英語を母語とする人が指摘した音節の総数のうち，システムが違いを指摘できた数を精度とした．この精度の算出方法の例として，日本人が “Academic aptitude guarantees your diploma” と話している際の評価の様子を図 8 に示す．図中の赤い部分が，システムがユーザの発音がネイティブスピーカーの発音と異なっていると指摘した部分である．一方，英語を母語とする人の評価者が指摘した部分（音節）は，*aca*, *dem*, *ap*, *ti*, *tude*, *diplo* と *ma* である．したがって，この例では，*aca, dem, ap, ti, tude,* はシステムにおいて正しく指摘できているが，*ic* は正しく指摘することができず，*diplo, ma* も指摘できていないということになる．attention の重みの閾値 th を 0.3 に設定したところ，英語を母語とする人が指摘した 84 部分のうち 50 個を指摘することができ，結果として精度は 59.5% となった．

4.4 ユーザ評価

DDSupport が第二言語のスピーキング学習に役に立つのかを検証するために，ユーザーテストを実施した．ユーザには英語のセンテンス 10 文の発話練習をアプリケーションでおこなってもらった．実験の被験者には，英語を母国語としない日本人 3 名を採用した．練習が 終わった後にその手法に関するアンケートを記入してもらった．質問内容は，

Q1 アプリケーションでの学習のしやすさを 5 段階で評価してください
Q2 アプリケーションが示した熟練度の判定および差分・距離の可視化は学習に役立つかを 5 段階で評価してください
Q3 アプリケーションでの学習後発音が向上したと思うかを 5 段階で評価してくだい

の 3 つである．アンケート結果は，Q1，Q2，Q3 それぞれ平均値が 4.0，3.6，4.0 であった．さらに，実験中の被験者の発話は録音しており，学習前と学習後の発音が明瞭になっているかを評価した．各センテンスにおいて最初の発話と最後の発話をピックアップし，英語を母国語とする 5 人に学習前と学習後のどちらの発話が分かりやすいかブラインドテストで判断してもらった．被験者 3 人，センテンス 10 文，採点者 5 人の全 150 発話のうち，学習後の発話の方が分かりやすいと評価された発話は 121（=81%）であった．

5. 議論

5.1 評価

音声分類の評価実験では，分類，差異・距離の可視化のモデルを学習するために用いた英語の非ネイティブスピーカーである日本人のラベルありデータが少ないにも関わらず，高精度に非ネイティブスピーカーとネイティブスピーカーを分類できていることが分かる．これは，主に大量のラベルなしデータを用いて自己教師あり学習をおこなった結果，良質な英語の音声表現を獲得できたために分類精度が向上したものと考えれられる．差異の可視化の評価実験では，プロトタイプとしてのシステムの有効性は確認できたが，まだ精度は十分とはいえず，今後システムがより高い性能を持つようにモデルを改良することが必要である．より高い精度を実現するためには，注目度の閾値をはじめとするモデルのパラメータの調整や，差異の可視化に適した attention 機構の改良が必要と考えられる．また，今回扱ったデータより多くのサンプルでテストを行い，システムが正しく差異を指摘できない要素の共通点をより詳細に分析する必要がある．ユーザ評価では，対象としたユーザが少ないため，今後より詳細な評価が必要であるものの，提案手法が第二言語のスピーキング学習において，学習者の補助になり，学習者の発音の分かりやすさを向上させる効果があることが分かる．発話の分かりやすさを評価する手法は様々提案されいるため，それらの指標でも学習者の発話を評価することで，アプリケーションの効果をより正確に検証できると考えられる．例えば，今回は学習者の学習初期の発話と学習後の発話の 2 つのうち分かりやすい方を，評価者に選択してもらう方法をとったが，他にも，学習初期の発話と学習後の発話を評価者に書き取りしてもらい，どちらがより正確に書き取れるかで評価することもできる．この評価方法は今回の評価方法と違い，システムを

通じて学習がうまくいく部分とそうでない部分を明らかにできるという利点もある．

5.2 技能獲得支援への展開

本論文では，学習者とネイティブスピーカーを識別し，それらの差異と距離を提示する言語学習アプリケーションとしてシステム構築をおこなった．しかし，本質的には本システムの機能は学習者と専門家（プロフェッショナル）の動作の識別及びそれらの差異と距離の提示をおこなうことである．また，ユーザの動作を表す時系列メディア情報であればその形は何であっても対応することができるため，本システムの入力としては言語の音声データに限る必要はない．例えば，本システムで楽器学習支援アプリケーションを構築したければ，システム構成はそのままで，自己教師あり学習モデルでラベルなしのプロの演奏家の演奏音声で学習をすればよい．また，ユーザの動作を表す動画を入力としたければ，動画表現を学習する自己教師あり学習モデル [29] に変更し，ラベルなしの動画データで学習をすれば対応することができる．将来的にはこのシステムを応用し，スポーツ，クッキング，医療行為，演説など多種多様な分野において初学者の学習を支援するアプリケーションを構築することが期待される．

6. 結論

本論文では，ユーザーの発話がうまくいっているかを判断しユーザの発話とネイティブスピーカーの発話の差異や距離を提示することで言語学習を支援するシステムを提案した．本システムでは深層学習の技術の一つである自己教師あり学習を応用することで，学習者の技能レベルを判定する．また，深層学習モデルの解釈性を向上させるattention 機構を用いて，ユーザの音声とネイティブスピーカーの音声の差異を提示する．さらに，ユーザーの音声とネイティブスピーカーの音声の間の距離を，距離学習によって設計された距離を用いて提示する．学習者は，これらの差異や距離を自分の目と耳でモデルの音声と比較し，システムとインタラクションすることで徐々に発話を改善することが可能となる．今後，学習者の技能レベルを判定し，学習者と専門家の差異や距離を提示する提案手法は，本論文で対象とした言語学習だけでなくスポーツや楽器などさまざまな分野でのスキル習得支援システムへ応用されることが期待される．

謝辞 査読者には貴重な助言を頂きました．心より感謝申し上げます．

参考文献

[1] Baevski, A., Zhou, H., Mohamed, A. and Auli, M.: wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations, *arXiv*, pp. 1–19 (2020).

[2] Bahdanau, D., Cho, K. and Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate (2014).

[3] Brodkey, D.: Dictation as a Measure of Mutual Intelligibility: A Pilot Study, *Language Learning*, Vol. 22, No. 2, pp. 203–217 (1972).

[4] Bu, Y., Ma, T., Li, W., Zhou, H., Jia, J., Chen, S., Xu, K., Shi, D., Wu, H., Yang, Z. et al.: PTeacher: a Computer-Aided Personalized Pronunciation Training System with Exaggerated Audio-Visual Corrective Feedback, *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–14 (2021).

[5] Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K. and Bengio, Y.: Attention-Based Models for Speech Recognition (2015).

[6] Culbertson, G., Shen, S., Jung, M. and Andersen, E.: Facilitating development of pragmatic competence through a voice-driven video learning interface, *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 1431–1440 (2017).

[7] Derwing, T. M. and Munro, M. J.: Accent, Intelligibility, and Comprehensibility. Evidence from Four L1s, *Studies in Second Language Acquisition*, Vol. 19, No. 1, p. 1–16 (1997).

[8] Derwing, T. M. and Munro, M. J.: Second Language Accent and Pronunciation Teaching: A Research-Based Approach, *TESOL Quarterly*, Vol. 39, No. 3, pp. 379–397 (2005).

[9] Edge, D., Cheng, K.-Y., Whitney, M., Qian, Y., Yan, Z. and Soong, F.: Tip tap tones: mobile microtraining of mandarin sounds, *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services*, pp. 427–430 (2012).

[10] Eskenazi, M.: An overview of spoken language technology for education, *Speech Communication*, Vol. 51, No. 10, pp. 832–844 (2009). Spoken Language Technology for Education.

[11] Graves, A. and Jaitly, N.: Towards End-To-End Speech Recognition with Recurrent Neural Networks, Vol. 32, No. 2, pp. 1764–1772 (2014).

[12] Hermann, K. M., Kociský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M. and Blunsom, P.: Teaching Machines to Read and Comprehend, pp. 1693–1701 (2015).

[13] Hochreiter, S. and Schmidhuber, J.: Long Short-term Memory, *Neural computation*, Vol. 9, pp. 1735–80 (1997).

[14] Hönig, F., Batliner, A. and Nöth, E.: Automatic Assessment of Non-Native Prosody — Annotation, Modelling and Evaluation，pp. 21–30 (2012).

[15] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R. and Mohamed, A.: HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units (2021).

[16] Hu, W., Qian, Y. and Soong, F.: A New DNN-Based High Quality Pronunciation Evaluation for Computer-Aided Language Learning (CALL). (2013).

[17] Jenkins, J.: The Phonology of English as an International Language (2000).

[18] Karhila, R., Ylinen, S. P., Enarvi, S., Palomäki, K., Nikulin, A., Rantula, O., Viitanen, V., Dhinakaran, K., Smolander, A.-R. M., Kallio, H. H. et al.: SIAK–a game for foreign language pronunciation learning, *Proceedings of INTERSPEECH 2017 Interspeech: Annual Conference of the International Speech Communication Asso-*

ciation, International Speech Communications Association (2017).
[19] Kenworthy, J.: Teaching English Pronunciaton, pp. 4–8 (1987).
[20] Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization (2014).
[21] Kulis, B. et al.: Metric learning: A survey, *Foundations and Trends® in Machine Learning*, Vol. 5, No. 4, pp. 287–364 (2013).
[22] Lee, A. and Glass, J.: Pronunciation assessment via a comparison-based system, *Speech and Language Technology in Education* (2013).
[23] Levis, J. M.: Changing Contexts and Shifting Paradigms in Pronunciation Teaching, *TESOL Quarterly*, Vol. 39, No. 3, pp. 369–377 (2005).
[24] Lin, T.-Y., Goyal, P., Girshick, R., He, K. and Dollár, P.: Focal Loss for Dense Object Detection (2017).
[25] Lloyd, S. P.: Least squares quantization in pcm, *IEEE Transactions on Information Theory*, Vol. 28, pp. 129–137 (1982).
[26] Marianne, C.-M., Brinton, D. M. and Goodwin, J. M.: Teaching Pronunciation Paperback with Audio CDs (2): A Course Book and Reference Guide (2010).
[27] Munro, M. J. and Derwing, T. M.: Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners, *Language Learning*, Vol. 45, No. 1, pp. 73–97 (1995).
[28] Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D. and Auli, M.: fairseq: A Fast, Extensible Toolkit for Sequence Modeling, *NAACL HLT*, pp. 48–53 (2019).
[29] Qian, R., Meng, T., Gong, B., Yang, M., Wang, H., Belongie, S. J. and Cui, Y.: Spatiotemporal Contrastive Video Representation Learning, *CoRR* (2020).
[30] Robertson, S., Munteanu, C. and Penn, G.: *Designing Pronunciation Learning Tools: The Case for Interactivity against Over-Engineering*, p. 1–13 (2018).
[31] Schultz, M. and Joachims, T.: Learning a Distance Metric from Relative Comparisons, Vol. 16 (2004).
[32] Schuster, M. and Paliwal, K.: Bidirectional Recurrent Neural Networks, *Signal Processing, IEEE Transactions on*, Vol. 45, pp. 2673 – 2681 (1997).
[33] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting, *Journal of Machine Learning Research*, Vol. 15, No. 56, pp. 1929–1958 (2014).
[34] Steffen, S., Alexei, B., Ronan, C. and Michael, A.: wav2vec: Unsupervised Pre-training for Speech Recognition, *INTERSPEECH*, pp. 3465–3469 (2019).
[35] Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q. V., Agiomyrgiannakis, Y., Clark, R. and Saurous, R. A.: Tacotron: Towards End-to-End Speech Synthesis, pp. 4006–4010 (2017).