Vol.2017-SLP-115 No.7
2017/2/17

国際会議INTERSPEECH2016報告

浅見 太一[1]　小川 厚徳[1]　小川 哲司[2]　大谷 大和[3]　倉田 岳人[4]　齋藤 大輔[5]　塩田 さやか[6]
篠原 雄介[3]　鈴木 雅之[4]　高道 慎之介[5]　南條 浩輝[7]　橋本 佳[8]　樋口 卓哉[1]　増村 亮[1]
吉野 幸一郎[9]　渡部 晋治[10]

概要：2016 年 9 月 8 日から 12 日にかけて米国・サンフランシスコで ISCA 主催の国際会議 INTERSPEECH2016 が開催された．INTERSPEECH は，音声言語情報処理分野における一流の国際会議として位置づけられており，音声言語情報処理の最新の研究動向が報告される場である．本稿では海外からの発表を中心に，注目すべき発表について報告を行う．

1. はじめに

2016 年 9 月 8 日から 12 日にかけて米国・サンフランシスコで行われた国際会議 INTERSPEECH 2016（http://www.interspeech2016.org/）での発表について報告を行う．1,585 件の投稿があり，そのうち 779 件の論文が採択されたとのことである．

本稿では，（1）音響モデル・特徴量・対雑音，（2）言語モデル・言語理解・対話，（3）サーチ，（4）音声変換・合成，（5）話者認識，の 5 つの分野の発表について注目する研究をいくつか選択し，報告を行う．各分野で活躍する 15 名の研究者にサーベイを依頼し，内容をまとめた．（南條）

2. 音響モデル・特徴量・耐雑音

2.1 音響モデル

ニューラルネットとして，通常の全結合層だけでなく，CNN，RNN もしくは TDNN が用いられるようになった．特に state-of-the-art のシステムは，両者を組み合わせることで構築されている [1]．CNN に関しては，CNN の一種である VGG を Batch Normalization や系列識別学習と組み合わせることで，高い性能が報告されている [2]．RNN に関しては，ベーシックな単方向 LSTM だけでなく，双方向の LSTM もよく使われているほか，LSTM によく似た GRU を使う例も見られる．文献 [3] では，双方向 GRU のゲートに畳み込みを融合した双方向 GRCU を提案し，認識精度が僅かながら向上することを報告している．また TDNN に関しては，長いコンテキストをモデル化出来るうえ，RNN よりも効率的に学習出来ることから，再注目されている．文献 [4] では，通常入力特徴量のみに行われる context expansion を，途中の層においても行うことにより，改善を示している．従来の TDNN とは attention を導入している点と residual connection による深層化を行っている点が異なり，ベンチマークで高い認識精度 [5] を報告している．これらの他に，文献 [6] や [7] では，highway connection を導入してさらに複雑なネットワークを構築し，さらなる改善を示している．

新しいトレンドの一つである end-to-end 音声認識も注目を集めた．半年前の ICASSP では encoder-decoder を用いたものが提案され話題となったが，今回は CTC を用いたものの改良などが発表された．文献 [8] では，CTC を用いた end-to-end 音声認識において，RNN の代わりに深層の CNN を使ったモデルを提案した．CTC で通常使われる RNN には学習が重く時に不安定であるという課題があったが，これを克服出来るとしている．文献 [9] では，CTC と単語単位の言語モデルを組み合わせる際に，サブワード単位の言語モデルのスコアを逆に差し引くことで MAP に基づくデコーディングが実現できることを示し，日本語の音声認識で高い性能を実現している．また文献 [10] では，segmental CRF と RNN を組み合わせた segmental RNN を用いた end-to-end 音声認識が提案されている．さらに，CTC においてフレームシフトを 10 msec より長くするだけで性能が改善することが発表されるなど，この分野の発展にはまだまだ目が離せない [11]．

1 日本電信電話株式会社
2 早稲田大学
3 東芝
4 日本 IBM
5 東京大学
6 首都大学東京
7 京都大学
8 名古屋工業大学
9 奈良先端科学技術大学院大学
10 Mitsubishi Electric Research Laboratories

系列識別学習でも注目すべき進展があった．文献 [12] では，lattice-free MMI を用いた，事前のクロスエントロピー学習が不要な系列識別学習を提案している．5 つの LVCSR タスクで評価を行い，CE → sMBR と比べて最大 8% 単語誤りを削減したと報告している．

音響モデルの小型化・高速化の発表も増えている．文献 [13] では，順伝搬時のみ量子化する (逆伝搬時はしない) quantization aware 学習により，認識精度を落とさずに，LSTM-CTC の結合荷重を 32-bit 浮動小数点から 8-bit 整数に量子化出来ると報告している．また文献 [14] では，大規模なモデル (e.g. アンサンブルモデル) と同等の振る舞いをする小規模なモデルを作成する student-teacher 学習を改良し，系列識別学習にも適用出来るようにしたと報告している．(篠原，鈴木)

2.2 特徴量・表現学習

近年，音響特徴抽出はニューラルネットワークを用いて行われることが多い．特に，音響特徴表現をデータから学習可能にするネットワークの構造や，音響特徴抽出器を音声認識のための識別基準で結合学習 (joint training) する枠組みに関する検討が多い．例えば，文献 [15] では，複素領域におけるフィルタバンクを音響モデルの最適化基準で学習する方式が提案されている．周波数分析のための基底関数を学習する枠組みについても近年盛んに検討がなされており，文献 [16] では，この基底関数を複数の解像度に対して同時に学習する試みがなされている．文献 [17] では，LSTM-HMM 音響モデルで時間・周波数パターンを捉えるための有効な構造について検証がなされている．

音素の識別に寄与する特徴パラメータを陽に抽出し音声認識の入力として用いる試みは依然として多い．近年では，特に埋め込み (embedding) に関する検討が盛んである．文献 [18] では，グローバルな多様体学習を自己符号化器を用いて実現し，LSTM-CTC の入力として用いている．文献 [19] では，音素と話者各々に関するトリプレット損失を最小化することで，音素の違いのみを強調する表現と話者の違いのみを強調する表現が得られるモデルを学習している．また，ローリソース音声認識や多言語音声認識への適用を目的とし，半教師あり学習や転移学習を用いて特徴表現を学習する試みも多い [20], [21]．調音特徴抽出器には従来よりニューラルネットワークが用いられるが，その学習において強制アライメイントを不要とする CTC を用いることで，最終的な音声認識誤りを大幅に削減できることが実証されている [22].

DNN 音響モデルにおける各層の役割やその最終的な識別性能に対する寄与について分析する試みもある [23], [24]．具体的には，DNN 音響モデルの各層の出力にソフトマックスを適用して得られた結果を MDS 法で可視化するとともに，それぞれの音素識別性能を調査している．(小川哲)

2.3 耐雑音

音声認識の前処理として音声強調を行うことで耐雑音性を向上させる研究が，近年盛んになされている．文献 [25] では，BLSTM を用いた時間周波数マスク推定に基づくビームフォーミング [26] にポストマスキングを導入することで，学習データとテストデータにミスマッチがある環境下で，音声認識性能を改善したことが報告されている．また，深層学習を用いた時間周波数マスク推定を複数人発話に適用する研究も，近年では行われている．文献 [27] では，BLSTM に直接マスクを出力させるのではなく，一度埋め込み空間に射影して学習させることで，話者間のパーミュテーションの問題を回避し，複数話者のマスク推定を実現している．この手法は，学習時にモデルの出力と話者間のアライメントをとる手法 [28] と合わせて，音声認識性能の改善が報告されている．これらの比較的新しいアプローチは，今後，音響モデルと音声強調に用いるモデルの結合学習により，さらなる音声認識性能向上が期待される．

また近年の傾向として，音声強調・認識両モデルをそれぞれニューラルネットワークで表現し，それらを音声認識の識別基準で結合学習するアプローチが盛んに研究されている（詳しくは [29]）．文献 [30] では，単一チャンネル denoising auto-encoder に基づく音声強調と音響モデルの結合学習，文献 [15], [31], [32] では，LSTM や CNN, 複素線形ネットワークを用いた複数チャンネルの音声強調と音響モデルの結合学習による耐雑音・遠隔発話音声認識における性能改善が報告されている．また文献 [33] では，複素混合分布を用いた音声強調と音響モデルの結合学習により，音声認識の前処理としての音声強調手法の改善が報告されている．(樋口，渡部)

CHiME2016 workshop

最後に，INTERSPEECH のサテライトワークショップとして開催された CHiME2016 workshop における，耐雑音音声認識の技術評価国際イベント第 4 回 CHiME challenge の技術動向について概説する．本チャレンジは前回行われた第 3 回 challenge のマイナーアップデートであり，前回と同一タスクである 6 チャンネルトラックに加え, マイク数を 6 から 1 及び 2 に制限することにより難易度を増加させた 1 チャンネル・2 チャンネルトラックの合計 3 トラックで構成される．世界各国 19 の研究グループが本イベントに参加し，3 トラック合計で 43 システムが提案された．上位チーム [34], [35], [36], [37], [38] を含む多くのシステムが前回 challenge のトップシステム [39] の性能を上回っている．またこれらの上位システムは全てマスク推定に基づくビームフォーミング (2.3 節にて説明) を用いており，実環境における同技術の有用性が広く示されたといえる．

またそれ以外でも，多チャンネルウィナーフィルタを用いた複数条件学習データの生成や [40]，マスク推定に基づくビームフォーミングと音声認識用音響モデルの統合学

習法 [41], 進化アルゴリズムを用いたネットワークの最適化 [42] などの興味深い技術も提案されている．

INTERSPEECH における耐雑音研究や CHiME workshop では，日本から興味深い研究や高性能システムの発表多く見られ (上記以外でも Adversarial Multi-task Learning の利用 [43] など), 本分野における日本のプレゼンスの高さを示しているといえる．(渡部)

3. 言語モデル・言語理解・対話

3.1 言語モデル

言語モデルにおいても，引き続きニューラルネットワークの利用に関する発表が多数行われていた．[44] では，LSTM，GRU，Highway Network を利用した言語モデルについて，音声認識実験を通じて比較が行われている．GRU と Highway Network の組み合わせとの比較を行った上で，LSTM がさらに優れているという実験結果は，参考になると考えられる．[45] では，RNN 言語モデルの適応方法として，適応データによって全パラメータを更新する方法と，隠れ層に対する要素ごとのスケーリングを学習する LHUC (Learning Hidden Unit Contribution) の比較が行われている．報告されている実験での改善はそれほど大きくなく，また双方の改善は同程度であったが，LHUC では推定するパラメータ数が少なく，実用上興味深い手法だと考えられる．

[46] では，多言語対応 RNN 言語モデルが提案されている．主にリソースが少ないマイナー言語の音響モデルを構築することを目的として，多言語対応の DNN 音響モデルが提案されている．多言語対応 DNN 音響モデルは，入力層と中間層を言語間で共有し，出力層のみ言語別に用意するという構造を持っている．入力層と中間層を言語間で共有して言語に共通な特徴量を抽出することで，メジャー言語のリソースをマイナー言語の音響モデリングに活用している．多言語対応 RNN 言語モデルは，中間層を言語間で共有し，入力層と出力層を言語別に用意するという構造を持っている．中間層で言語に共通な特徴量を抽出 (及び再帰結合で伝搬) することにより，複数言語のリソースをターゲット言語の言語モデリングに活用している．14 種類の性質の異なるマイナー言語データを用いた実験が行われている．効果はそれほど大きくないものの，性質の異なる言語間で言語モデリングを共通化できる可能性を示した意義は大きいと思われる．（小川厚，倉田）

3.2 言語理解・対話

音声言語理解では，スロットフィリングや発話意図理解などのタスクにおいて，Encoder-Decoder アプローチなど，近年注目を集める深層学習技術の適用が昨年に引き続き検討されていた．[47] では，過去のターンの情報も活かしたスロットフィリングと発話意図理解のために，End-to-End Memory Network を応用した手法を提案している．End-to-End Memory Network は長距離コンテキスト情報を選択的に利用するために有用な技術であり，今後は様々なタスクで適用が進むと考えられる．

また，特に Encoder-Decoder に関しては，RNN を用いた Sequence-to-Sequence の研究が注目されている．今回の会議においても，この Sequence-to-Sequence のモデルをユーザシミュレーションのタスクに適用する研究 [48] が発表され，Dialogue State Tracking Challenge (DSTC) 2 および 3 のタスク達成対話ドメインで適用が試みられている．この研究では，タスク対話におけるコンテキスト，具体的には直前にどのような対話行為が用いられたか，どのスロットが埋まっているかの情報などを入力として，次の行動を出力する Encoder-Decoder モデルを学習している．

音声対話の実用化に伴い，音声言語理解のパーソナライズ化についても検討が進んでいる．[49] では，個人情報は保護しつつ，ユーザごとに異なる音声言語理解を行う方式を検討している．報告されている実験では，発話中に含まれる人名やアプリケーション名など，ユーザごとに大きく異なる部分で，大幅な改善効果が確認されている．音声対話におけるパーソナライズ化は，対話管理や応答生成など，さらに検討が進むことが期待される．

一方で，対話のモデルに関する研究のトレンドとしては，ユーザの発話に対する同調現象（エントレインメント）が注目されている．同調が起こる現象は韻律，語彙選択，対話行為など多岐に渡るが，[50] はこれらのうち音響的・韻律的特徴における同調を行うような対話エージェントを提案している．具体的には，同調行為が必要であるとシステムが認定した場合に，パワー，ピッチにおいて音声合成のパラメータを調整し対話に利用している．今後は，こうしたユーザとより自然に対話を行うための現象について，実際に対話システムで用いるためのモデル化が進められると考えられる．（増村，吉野）

4. サーチ

サーチの分野においても，HMM を置き換える CTC ネットワークや，入力系列全体を利用する双方向 LSTM の音声認識への適用に伴い，新たなモデルの性質に合わせた新方式が提案された．

HMM 音響モデルを用いる音声認識では，音響モデルと言語モデルの独立性を前提として，HMM 音響モデルから得られる音響スコアと，言語モデルから得られる言語スコアを線形補間して求めた仮説のスコアを探索に用いていた．一方，CTC 音響モデルは，特に音素や音節などの HMM 状態よりも高次の出力シンボルをターゲットとして学習した場合，音響スコアの算出に（部分的な）言語情報が用いられることになる．そのため，従来のデコーディングが前提としていた音響モデルと言語モデルの独立性が崩れ，音

響スコアと言語スコアの単純な線形補間は仮説のスコア計算方法として適切ではなくなる．[9] はこの問題に対して，CTC 音響モデルが持つ言語情報を明示的に考慮し，仮説のスコア計算時にキャンセル（正規化）する枠組みを提案している．CTC 音響モデルの出力シンボルの出現確率による正規化を，最大事後確率基準による音声認識の定式化 ($\hat{W} = argmax_W P(W|X)$) に基づいて導出し，WFST に導入する方式により，DNN-HMM ハイブリッド型を超える認識精度が達成されている．

また，CTC 音響モデルは，デコーディング時に多くのフレームで認識結果に寄与しない空シンボルに高い確率を与え，数フレームから数十フレーム程度の間隔で意味のある（音素などの）シンボルに高い確率を与えるという振る舞いをする．そのため，仮説の生成と枝刈りをフレーム毎に行うフレーム同期ビームサーチでは，空シンボルに由来する無駄な仮説が大量に生成され，探索効率が低下するという問題も指摘された [51]．この問題に対して [51] は，CTC が空でないシンボルを出力したフレームでのみ仮説を生成する音素同期ビームサーチの枠組みを提案している．はじめに CTC 音響モデルだけを用いて構築した音素ラティスを後段の WFST と合成してビームサーチを行う実現方法を採り，フレーム同期ビームサーチと同等の認識精度を保ちつつ 2〜3 倍の速度向上が得られている．

双方向 LSTM (BLSTM) が単方向 LSTM よりも高い精度を示すことはよく知られているが，BLSTM は入力系列の終端が決まるまで出力が得られないため，オンライン処理に適用できないという問題がある．[52] は，入力ストリームへの窓掛けによって先読み量を限定するアプローチで，BLSTM 音響モデルのオンラインデコーディングへの適用に挑戦している．500〜1000ms 程度の窓長，50〜100ms 程度のシフト幅で，各窓を入力系列全体とみなして BLSTM 音響モデルで HMM 状態事後確率を計算し，窓のオーバーラップ部分の事後確率を平均化するシンプルな方法により，発話全体を使った場合と同等の認識精度を得られることが報告された．

DNN-HMM 型音声認識を計算能力や記憶容量の限られる組み込みデバイスで動作させるための技術も発表された．[53] では，1MB 以下の音響モデルパラメータ，10MB/s 以内のメモリバンド幅で DNN-HMM 型音声認識のリアルタイム処理を実現するための工夫の数々が報告された．モデルパラメータの量子化や WFST のデータ構造の効率化に加え，ハードウェアにより規定されるメモリ上限を超えないように，探索中の仮説数の増減に応じて動的にビーム幅を制御する手法を導入することで，認識精度を低下させることなく計算量を 20%程度削減することに成功している．（浅見）

5. 音声変換・音声合成

音声変換並びに音声合成に関するセッションは，オーラルが 4 つ（スペシャルセッションを含む），ポスター 2 つで構成されていた．特に今回，音声変換分野において特筆すべき点が共通のデータセットを用いて声質変換技術を評価する Voice Conversion Challenge (VCC) のスペシャルセッションが開催されたことである [54]．VCC は，テキスト音声合成における国際的な評価チャレンジである Blizzard Challenge と同様に，共通のデータセットを用いて各声質変換技術を比較する事で，その理解を深めていくことを目的としている．今回のチャレンジでは入力話者 5 名，出力話者 5 名の総当たりのペアで変換システムを作成し，評価データに対して声質変換を行った上で大規模な主観評価実験を行った．チャレンジ全体には 17 チームの参加があり，スペシャルセッションではこれらのうちのいくつかの発表と，聴取実験結果のまとめが発表された [55]．今回のチャレンジにおける学習データ，評価データと提出されたサンプル，及び主観評価実験の結果は全て公開されており，今後新しい声質変換技術についても，このデータを共通の評価データとして用いた比較が可能となる．（齋藤）

5.1 ビッグデータの利用

高品質な音声合成を学習するためには，通常，十分に配慮された収音環境や音素バランスに基づいた音声コーパスの使用が不可欠である．故に，背景ノイズやチャネルノイズを含む超大規模コーパス（例えば YouTube-8M）や，十分な収音環境を確保できない希少言語の音声などの直接的な利用は，合成音声の品質を著しく劣化させる．[56] は，大量の音声データから音声合成の学習データを選択する基準（measure of goodness）について調査している．この論文では，背景ノイズやチャネルノイズを検出する基準として，メルケプストラム歪み，変調スペクトルや相関性などを利用する．また，特定の基準を最大化するようにデータを選択し，音声合成を反復的に学習する枠組みを提案している．関連研究として，[57] では，DNN 音声合成における音素アライメントの影響を調査している．学習時にアライメントを自動修正する HMM 音声合成とは異なり，典型的な DNN 音声合成は与えられたアライメントを固定して学習を行う．故に，学習時に与える音素アライメントは，DNN 音声合成における品質に強く影響を及ぼす．[57] では，アライメント用 HMM を準備し，HMM パラメータが合成音声品質に与える影響を実験的に調査している．この評価はクリーン音声のみで実施されているが，[56] と同様，背景ノイズやチャネルノイズを含む音声での評価及び品質改善法が期待される．（高道）

5.2 多言語・複数話者

[58] では LSTM-RNN を用いた多言語・複数話者音声合成を提案している．本手法は，3 層の LSTM で構築された平均タワーと基底タワー，言語コードからなる言語ブロックと，単純な RNN と話者コードからなる話者ブロックで構成される．合成時には，言語特徴ベクトルを各タワーに入力し，これらの出力を言語コードに基づいて線形結合する．これを話者コードが示す目標話者の RNN に入力することで目標話者の音声を合成する．また，新言語のデータを用いて言語コードと平均タワーのパラメータを更新することで言語適応を実現している．実験では，単一言語・単一話者の音声合成と比較して良好な結果を示した．（大谷）

5.3 省リソース

[59] では LSTM-RNN を用いた音声合成システムの携帯デバイスでの利用のために，ネットワークの重みパラメータの量子化によるディスクフットプリントの削減，複数フレームの同時推定による計算量の削減，ϵ-contaminated Gaussian loss function を用いた頑健なパラメータ推定を行っている．実験結果から，約 70%の省メモリ化，約 40%の実行時間の削減を実現すると同時に，波形接続型音声合成システムと同程度の高い自然性を示した．[60] では，音声認識用の不特定話者 DNN 音響モデルが出力する音素状態（senone）事後確率を利用することで，パラレルデータを用いることなく声質変換システムを構築する方法を提案している．DNN から出力される音素状態事後確率は話者性への依存度が少ないと仮定し，音素状態事後確率の KL 距離を利用して元話者から目標話者へとマッピングを行う．実験結果から，DNN に基づく声質変換システムよりも高い自然性と話者性を示した．（橋本）

6. 話者認識

話者認識分野の発表を大別すると，なりすまし攻撃・短発話・実環境ノイズの 3 つに分けられる．

まず，なりすまし攻撃に関する発表では，フィルタバンクを工夫して得られる特徴量（e.g., IMFCC, RFCC, SSFC）や Local Binary Pattern (LBP) を用いた [61]，識別に有効なサブバンドの選択 [62] などが挙げられる．注目すべき点としては，なりすまし攻撃を考慮したデータベースが ASVspoof[63], AVspoof[64], SAS[65] と立て続けに公開されたため，また各研究機関がそれぞれのデータベースについて調査をしている段階であることが挙げられる．

また，話者認識関連で 2 つのスペシャルセッションが開催された．1 つは Interspeech2016 で公開された RedDots データベースを用いた RedDots Challenge で，短発話を用いたテキスト依存型話者照合における精度向上を目指したものである．[66] は，テキスト依存話者照合システムを発話照合と話者照合という 2 つの照合システムの組み合わせで実現することを提案している．発話照合の手法としては，DTW や HMM/DNN 音声認識によるアライメント推定などを用いており複数の発話照合と話者照合を組み合わせることで高い照合性能が得られることを報告している．

もう 1 つのスペシャルセッションは SITW データベース [67] を用いたコンペティションである．SITW も新しく公開されたデータベースで，ノイズや残響などを制御せずリアルな環境でデータ収集したことが特徴となっている．投稿された論文の傾向としては，既存の手法である DNN/i-vector による手法をベースに工夫したものが多かった．

話者認識全体の傾向としてはより実用に近いデータベースの使用と実用に近いシナリオ（短発話，なりすまし攻撃）に対する頑健性向上を目標としており，学習アルゴリズムの改善よりも抽出する特徴量を増やす方向に進んでいるといえる．（塩田）

7. おわりに

INTERSPEECH 2016 での発表について報告を行った．最新の研究動向がうかがえると思う．本稿が，音声研究における日本のプレゼンスを高める一助になれば幸いである．

次回の INTERSPEECH は，2017 年 8 月 20 日から 24 日の日程で，スウェーデン・ストックホルムで開催予定である（http://www.interspeech2017.org/）．（南條）

参考文献

[1] Saon, G., Sercu, T., Rennie, S. and Kuo, J. H.-K.: The IBM 2016 English Conversational Telephone Speech Recognition System, *Proc. INTERSPEECH*, pp. 7–11 (2016).
[2] Sercu, T. and Goel, V.: Advances in Very Deep Convolutional Neural Networks for LVCSR, *Proc. INTERSPEECH*, pp. 3429–3433 (2016).
[3] Nussbaum-Thom, M., Cui, J., Ramabhadran, B. and Goel, V.: Acoustic Modeling Using Bidirectional Gated Recurrent Convolutional Units, *Proc. INTERSPEECH*, pp. 390–394 (2016).
[4] Yu, D., Xiong, W., Droppo, J., Stolcke, A., Ye, G., Li, J. and Zweig, G.: Deep Convolutional Neural Networks with Layer-Wise Context Expansion and Attention, *Proc. INTERSPEECH*, pp. 17–21 (2016).
[5] Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D. and Zweig, G.: Achieving human parity in conversational speech recognition, *arXiv:1610.05256* (2016).
[6] Hsu, W.-N., Zhang, Y., Lee, A. and Glass, J.: Exploiting Depth and Highway Connections in Convolutional Recurrent Deep Neural Networks for Speech Recognition, *Proc. INTERSPEECH*, pp. 395–399 (2016).
[7] Lu, L. and Renals, S.: Small-footprint Deep Neural Networks with Highway Connections for Speech Recognition, *Proc. INTERSPEECH*, pp. 12–16 (2016).
[8] 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, *Proc. INTERSPEECH*, pp. 410–414 (2016).
[9] Kanda, N., Lu, X. and Kawai, H.: Maximum A Posteriori based Decoding for CTC Acoustic Models, *Proc. INTERSPEECH*, pp. 1868–1872 (2016).
[10] Lu, L., Kong, L., Dyer, C., Smith, N. A. and Renals, S.: Segmental Recurrent Neural Networks for End-to-End Speech Recognition, *Proc. INTERSPEECH*, pp. 385–389 (2016).
[11] Pundak, G. and Sainath, T. N.: Lower Frame Rate Neural Network Acoustic Models, *Proc. INTERSPEECH*, pp. 22–26 (2016).
[12] Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y. and Khudanpur, S.: Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI, *Proc. INTERSPEECH*, pp. 2751–2755 (2016).
[13] Alvarez, R., Prabhavalkar, R. and Bakhtin, A.: On the Efficient Representation and Execution of Deep Acoustic Models, *Proc. INTERSPEECH*, pp. 2746–2750 (2016).
[14] Wong, J. H. and Gales, M. J.: Sequence Student-Teacher Training of Deep Neural Networks, *Proc. INTERSPEECH*, pp. 2761–2765 (2016).
[15] Variani, E., Sainath, T. N., Shafran, I. and Bacchiani, M.: Complex Linear Projection (CLP): A Discriminative Approach to Joint Feature Extraction and Acoustic Modeling, *Proc. INTERSPEECH*, pp. 808–812 (2016).
[16] Zhu, Z., Engel, J. H. and Hannun, A.: Learning multiscale features directly from waveforms, *Proc. INTERSPEECH*, pp. 1305–1309 (2016).
[17] Sainath, T. N. and Li, B.: Modeling time-frequency patterns with LSTM vs. convolutional architectures for LSVCSR tasks, *Proc. INTERSPEECH*, pp. 813–817 (2016).
[18] Liu, Y. and Kirchhoff, K.: Novel front-end features based on neural graph embeddings for DNN-HMM and LSTM-CTC acoustic modeling, *Proc. INTERSPEECH*, pp. 793–797 (2016).
[19] Zeghidour, N., Synnaeve, G., Usunier, N. and Dupoux, E.: Joint learning of speaker and phonetic similarities with siamese netowrks, *Proc. INTERSPEECH*, pp. 1295–1299 (2016).
[20] Mitra, V., Vergyri, D. and Franco, H.: Unsupervised learning of acoustic units using autoencoders and Kohonen nets, *Proc. INTERSPEECH*, pp. 1300–1304 (2016).
[21] Xu, H., Su, H., Ni, C., Xiao, X., Huang, H., Chng, E.-S. and Li, H.: Semi-supervised and cross-lingual knowledge transfer leraning for DNN hybrid acoustic models under low-resource conditions, *Proc. INTERSPEECH*, pp. 1315–1319 (2016).
[22] Abraham, B., Umesh, S. and Joy, N. M.: Articulatory feature extraction using CTC to build articulatory classifiers without forced frame alignments for speech recognition, *Proc. INTERSPEECH*, pp. 798–802 (2016).
[23] Pellegrini, T. and Mouysset, S.: Inferring phonemic classes from CNN maps using clustering, *Proc. INTERSPEECH*, pp. 1290–1294 (2016).
[24] Nagamine, T., Seltzer, M. L. and Mesagarani, N.: On the role of nonlinear transformations in deep neural network acoustic models, *Proc. INTERSPEECH*, pp. 803–807 (2016).
[25] Erdogan, H., Hershey, J., Watanabe, S., Mandel, M. and Roux, J. L.: Improved MVDR Beamforming using Single-Channel Mask Prediction Networks, *Proc. INTERSPEECH*, pp. 1981–1985 (2016).
[26] Heymann, J., Drude, L. and Haeb-Umbach, R.: Neural Network Based Spectral Mask Estimation for Acoustic Beamforming, *Proc. ICASSP*, pp. 196–200 (2016).
[27] Isik, Y., Roux, J. L., Chen, Z., Watanabe, S. and Hershey, J. R.: Single-Channel Multi-Speaker Separation using Deep Clustering, *Proc. INTERSPEECH*, pp. 545–549 (2016).
[28] Dong Yu, Morten Kolbak, Z.-H. T. and Jensen, J.: Permutation Invariant Training of Deep Models for Speaker-Independent Multi-talker Speech Separation, *arXiv preprint (https://arxiv.org/abs/1607.00325)* (2016).
[29] Delcroix, M. and Watanabe, S.: Recent Advances in Distant Speech Recognition, *INTERSPEECH Tutorial*, (online), available from ⟨http://www.kecl.ntt.co.jp/icl/signal/dsr-tutorial/⟩ (2016).
[30] Mimura, M., Sakai, S. and Kawahara, T.: Joint Optimization of Denoising Autoencoder and DNN Acoustic Model Based on Multi-target Learning for Noisy Speech Recognition, *Proc. INTERSPEECH*, pp. 3803–3807 (2016).
[31] Li, B., Sainath, T. N., Weiss, R. J., Wilson, K. W. and Bacchiani, M.: Neural Network Adaptive Beamforming for Robust Multichannel Speech Recognition, *Proc. INTERSPEECH*, pp. 1976–1980 (2016).
[32] Sainath, T. N., Narayanan, A., Weiss, R. J., Variani, E., Wilson, K. W., Bacchiani, M. and Shafran, I.: Reducing the Computational Complexity of Multimicrophone Acoustic Models with Integrated Feature Extraction, *Proc. INTERSPEECH*, pp. 1971–1975 (2016).
[33] Higuchi, T., Yoshioka, T. and Nakatani, T.: Optimization of Speech Enhancement Front-end with Speech Recognition-level Criterion, *Proc. INTERSPEECH*, pp. 3808–3812 (2016).
[34] Du, J., Tu, Y.-H., Sun, L., Ma, F., Wang, H.-K., Pan, J., Liu, C., Chen, J.-D. and Lee, C.-H.: The USTC-iFlytek System for CHiME-4 Challenge, *CHiME 2016 Workshop*, pp. 36–38 (2016).
[35] Menne, T., Heymann, J., Alexandridis, A., Irie, K., Zeyer, A., Kitza, M., Golik, P., Kulikov, I., Drude, L., Schluter, R., Ney, H., Haeb-Umbach, R. and Mouchtaris, A.: The RWTH/UPB/FORTH System Combination for the 4th CHiME Challenge Evaluation, *CHiME 2016 Workshop*, pp. 39–44 (2016).
[36] Erdogan, H., Hayashi, T., Hershey, J. R., Hori, T., Hori, C., Hsu, W.-N., Kim, S., Le Roux, J., Meng, Z. and Watanabe, S.: Multi-Channel Speech Recognition: LSTMs All the Way Through, *CHiME 2016 Workshop*, pp. 45–48 (2016).
[37] Heymann, J., Drude, L. and Haeb-Umbach, R.: Wide Residual BLSTM Network with Discriminative Speaker Adaptation for Robust Speech Recognition, *CHiME 2016 Workshop*, pp. 12–17 (2016).
[38] Tachioka, Y., Watanabe, S. and Hori, T.: The MELCO/MERL System Combination Approach for the Fourth CHiME Challenge, *CHiME 2016 Workshop*, pp. 1–3 (2016).
[39] Yoshioka, T., Ito, N., Delcroix, M., Ogawa, A., Kinoshita, K., Fujimoto, M., Yu, C., Fabian, W. J., Espi, M., Higuchi, T., Araki, S. and Nakatani, T.: The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices, *ASRU*, pp. 436–443 (2015).
[40] Fujita, Y., Homma, T. and Togami, M.: Unsupervised network adaptation and phonetically-oriented system combination for the CHiME-4 challenge, *CHiME*

2016 Workshop, pp. 49–51 (2016).
[41] Xiao, X., Xu, C., Zhang, Z., Zhao, S., Sun, S. and Watanabe, S.: A Study of Learning Based Beamforming Methods for Speech Recognition, *CHiME 2016 Workshop*, pp. 26–31 (2016).
[42] Tanaka, T., Shinozaki, T., Watanabe, S. and Hori, T.: Evolution Strategy Based Neural Network Optimization and LSTM Language Model for Robust Speech Recognition, *CHiME 2016 Workshop*, pp. 32–35 (2016).
[43] Shinohara, Y.: Adversarial Multi-task Learning of Deep Neural Networks for Robust Speech Recognition, *Proc. INTERSPEECH*, pp. 2369–2372 (2016).
[44] Irie, K., Tuske, Z., Alkhouli, T., Schluter, R. and Ney, H.: LSTM, GRU, Highway and a Bit of Attention: An Empirical Overview for Language Modeling in Speech Recognition, *Proc. INTERSPEECH*, pp. 3519–3523 (2016).
[45] Gangireddy, S. R., Swietojanski, P., Bell, P. and Renals, S.: Unsupervised Adaptation of Recurrent Neural Network Language Models, *Proc. INTERSPEECH*, pp. 2333–2337 (2016).
[46] Ragni, A., Dakin, E., Chen, X., Gales, M. J. and Knill, K. M.: Multi-Language Neural Network Language Models, *Proc. INTERSPEECH*, pp. 3042–3046 (2016).
[47] Chen, Y.-N., Hakkani-Tur, D., Tur, G., Gao, J. and Deng, L.: End-to-End Memory Networks with Knowledge Carryover for Multi-Turn Spoken Language Understanding, *Proc. INTERSPEECH*, pp. 3245–3249 (2016).
[48] Asri, L. E., He, J. and Suleman, K.: A Sequence-to-Sequence Model for User Simulation in Spoken Dialogue Systems, *Proc. INTERSPEECH*, pp. 1151–1155 (2016).
[49] Liu, X., Sarikaya, R., Zhao, L., Ni, Y. and Pan, Y.-C.: Personalized Spoken Language Understanding, *Proc. INTERSPEECH*, pp. 1146–1150 (2016).
[50] Levitan, R., Benuš, Š., Gálvez, R. H., Gravano, A., Savoretti, F., Trnka, M., Weise, A. and Hirschberg, J.: Implementing Acoustic-Prosodic Entrainment in a Conversational Avatar, *Proc. INTERSPEECH*, pp. 1166–1170 (2016).
[51] Chen, Z., Deng, W., Xu, T. and Yu, K.: Phone synchronous decoding with CTC lattice, *Proc. INTERSPEECH*, pp. 1923–1927 (2016).
[52] Zeyer, A., Schluter, R. and Ney, H.: Towards Online-Recognition with Deep Bidirectional LSTM Acoustic Models, *Proc. INTERSPEECH*, pp. 3424–3428 (2016).
[53] Price, M., Chandrakasan, A. and Glass, J.: Memory-efficient modeling and search techniques for hardware ASR decoders, *Proc. INTERSPEECH*, pp. 1893–1897 (2016).
[54] Toda, T., Chen, L.-H., Saito, D., Villavicencio, F., Wester, M., Wu, Z. and Yamagishi, J.: The Voice Conversion Challenge 2016, *Proc. INTERSPEECH*, pp. 1632–1636 (2016).
[55] Wester, M., Wu, Z. and Yamagishi, J.: Analysis of the Voice Conversion Challenge 2016 Evaluation Results, *Proc. INTERSPEECH*, pp. 1637–1641 (2016).
[56] Baljekar, P. and Black, A. W.: Utterance Selection Techniques for TTS Systems Using Found Speech, *Proc. in SSW9*, pp. 199–204 (2016).
[57] Li, M., Wu, Z. and Xie, L.: On the impact of phoneme alignment in DNN-based speech synthesis, *Proc. in SSW9*, pp. 212–217 (2016).
[58] Li, B. and Zen, H.: Multi-Language Multi-Speaker Acoustic Modeling for LSTM-RNN based Statistical Parametric Speech Synthesis, *Proc. INTERSPEECH*, pp. 2468–2472 (2016).
[59] Zen, H., Agiomyrgiannakis, Y., Egberts, N., Henderson, F. and Szczepaniak, P.: Fast, Compact, and High Quality LSTM-RNN Based Statistical Parametric Speech Synthesizers for Mobile Devices, *Proc. INTERSPEECH*, pp. 2273–2277 (2016).
[60] Xie, F.-L., Soong, F. and Li, H.: A KL Divergence and DNN-based Approach to Voice Conversion without Parallel Training Sentences, *Proc. INTERSPEECH*, pp. 287–291 (2016).
[61] Korshunov, P. and Marcel, S.: Cross-database evaluation of audio-based spoofing detection systems, *Proc. INTERSPEECH*, pp. 1705–1709 (2016).
[62] Sriskandaraja, K., Sethu, V., Le, P. N. and Ambikairajah, E.: Investigation of Sub-Band Discriminative Information between Spoofed and Genuine Speech, *Proc. INTERSPEECH*, pp. 1710–1714 (2016).
[63] Wu, Z., Kinnunen, T., Evans, N., Yamagishi, J., Hanilci, C., Sahidullah, M. and Sizov, A.: ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge, *Proc. INTERSPEECH*, pp. 2037–2041 (2015).
[64] Ergunay, S. K., Khoury, E., Lazaridis, A. and Marce, S.: On the Vulnerability of Speaker Verification to Realistic Voice Spoofing, *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pp. 1–6 (2015).
[65] Wu, Z., Khodabakhsh, A., Demiroglu, C., Yamagishi, J., Saito, D., Toda, T. and King, S.: SAS: A speaker verification spoofing database containing diverse attacks, *Proc. ICASSP*, pp. 4440–4444 (2015).
[66] Kinnunen, T., Sahidullah, M., Kukanov, I., Delgado, H., Todisco, M., Sarkar, A., Thomsen, N. B., Hautamaki, V., Evans, N. and Tan, Z.-H.: Utterance Verification for Text-Dependent Speaker Recognition: a Comparative Assessment Using the RedDots Corpus, *Proc. INTERSPEECH*, pp. 430–434 (2016).
[67] McLaren, M., Ferrer, L., Castan, D. and Lawson, A.: The Speakers in the Wild (SITW) Speaker Recognition Database, *Proc. INTERSPEECH*, pp. 818–822 (2016).