Vol.2016-SLP-112 No.5
2016/7/28

国際会議ICASSP2016参加報告

峯松 信明[1]　秋田 祐哉[2]　浅見 太一[3]　伊藤 信貴[3]　落合 翼[4]　郡山 知樹[5]　齋藤 大輔[1]
塩田 さやか[6]　篠崎 隆宏[5]　鈴木 雅之[7]　高木 信二[8]　俵 直弘[9]　橋本 佳[10]　樋口 卓哉[3]　福田 隆[7]

概要：2016 年 3 月 20 日から 25 日にかけ中国・上海で開催された IEEE 主催の ICASSP2016 に参加した．ICASSP は音声言語情報処理分野における top conference と位置づけられており，今後の本分野の動向に大きく影響を与えている．ここでは，海外からの発表を中心に，本会議における最新の研究動向や注目すべき発表について報告する．

1. はじめに

2016 年 3 月 20 日から 25 日にかけ中国・上海で開催された IEEE 主催の ICASSP2016 に参加した．ICASSP は，INTERSPEECH と並んで音声言語情報処理分野の top conference と位置づけられており，前者の方がより技術色の濃い会議となっている．通常論文の投稿数は 2682 件あり，採択数は 1265 件（受理率 47%）であった．本稿ではこれらの論文の中から，1) 音声認識，2) 話者認識・照合，言語・年齢推定，3) 音声合成・声質変換の 3 つのテーマに関し，筆者らが注目する研究をいくつか選択し，最新の技術動向についても言及する．

2. 音声認識

音声認識全体の研究動向としては，引き続きディープラーニングに関する研究発表が多く見られ，音響モデルだけでなく言語モデルやフロントエンド・音声強調処理の分野にも同技術が深く浸透し始めている．音響モデルについては DNN や CNN の検討が主流であるが，最近では LSTM の利用も増えてきており，今回の会議でも関連研究の発表があった．中でも，音声認識システムを単一のネットワークで表現し，全てを統一的に学習する End-to-End アプローチに関する発表が大きな注目を浴びた．一方，言語処理関連においては RNN・LSTM の検討が中心であり，実用的な進展も見受けられる．本節ではフロントエンド・音声強調処理，音響モデル，言語モデルを中心に注目すべき研究発表をまとめる．

2.1 フロントエンド・音声強調処理

近年では，従来の信号処理による音声強調処理を，深層学習に基づく音響モデル学習と関連付けて考える研究がいくつか行われている．これはシングルマイク入力に対する音声強調処理に限った話ではなく，マイクロホンアレイ分野にもニューラルネットワークを適用する動きが広まりつつあり，本会議でもこれに関連する数件の発表があった．今回の ICASSP における発表の中では，例えば Sainath ら [1] は，多チャンネル観測信号に対するビームフォーミングによる音声強調処理を，時間領域の信号に対する畳み込み処理として考え，CNN の一部として音響モデルと結合し事前学習する手法に対する研究を進めている．従来のビームフォーミングにおけるフィルタ (と特徴量抽出のためのフィルタ) は，CNN 内部のフィルタとして捉えられ，多チャンネルの時間領域信号に対して複数のフィルタを畳み込み，その出力にプーリング処理を施すことで，各時間フレームの特徴量を得ている．これらの処理を行う CNN は，音響モデルと結合し学習されるので，認識性能の規準によって最適化される．また次の例として，Wisdom ら [2] の研究では，生成モデル (GMM) を用いた音源分離手法において，生成モデルのパラメータ推定プロセスを “unfold” し，推定したパラメータによる音源分離プロセスと合わせて，深いネットワークによる処理として考え，教師信号を用いてパラメータを事前学習している．これにより，生成モデルの規準でパラメータ推定を行った場合と比べて，精度よい音源分離を実現している．この手法もまた，自然な形で音響モデルと結合することができると考えられ，今後

[1] 東京大学
[2] 京都大学
[3] 日本電信電話株式会社
[4] 同志社大学
[5] 東京工業大学
[6] 首都大学東京
[7] 日本 IBM
[8] 国立情報学研究所
[9] 早稲田大学
[10] 名古屋工業大学

の動向が注目される．

その他，文献 [3] の雑音除去法では，各時間周波数点にて音声と雑音のいずれがより大きいパワーを持つかを表すマスクをニューラルネットワークを用いて推定し，前記マスクに基づいてビームフォーマを設計する．本手法により騒がしい実環境での音声認識性能を大幅に改善できたと報告している．文献 [4] では，ビームフォーマ・特徴抽出・音響モデルを一体化したネットワークを，共通の評価関数に基づいて全体最適化する方法が提案されている．文献 [5] の音源定位法では，音源位置の情報を含んだマルチチャネル複素スペクトルを扱える特殊な活性化関数が用いられている．

2.2 音響モデル

音響モデルに関する発表もディープラーニングに関するものが依然として主流である．音声認識に対してニューラルネットワークを適用する主流なアプローチとして，「ハイブリッドアプローチ」と「タンデムアプローチ」の 2 つが挙げられる．これらのアプローチは共に，音声データに含まれる時間的な特性を HMM に基づいてモデリングするアプローチであった．これに対して今回の ICASSP では，LSTM を代表とするリカレント構造を持ったニューラルネットワークを使用することで，ニューラルネットワークのみによって音声データに含まれる時間的な特性をモデリングし，音声認識システムを構築する End-to-End アプローチの研究が多く見られた．以前より報告されていた CTC アプローチを基礎とした研究 [6], [7] が報告されているのに加えて，特に今回の ICASSP では，音声翻訳の分野で提案された Encoder-Decoder モデルによる sequence to sequence learning の枠組みを基礎とした研究が，新たに同時多発的に報告されている [8], [9], [10]．Encoder-Decoder アプローチでは，入力フレーム系列から出力アルファベット系列へのマッピングをニューラルネットワークのリカレント構造によって直接的にモデリングする．このとき，入力と出力の対応関係 (ある種のアラインメント) 自体も，attention と呼ばれる機構に基づき，学習を通して自動に獲得される．これらの報告の実験結果からは，現状この Encoder-Decoder アプローチが，上記のハイブリッドアプローチや CTC アプローチ等と比較して，認識精度の面で優れた結果を示しているとは言い難い．しかしながら，音声認識問題に対するニューラルネットワークのより効果的な適用方法を探る上で，今後のさらなる研究が期待されるアプローチであることは間違いないであろう．

一方，ハイブリッド型の音響モデルでは，より深い階層を持つモデルが注目を集めている．10 層を超える深いニューラルネットワークを適切に学習する方法として，現在 2 通りの方法が知られている．1 つ目の方法は，小さな畳み込み層を用いてパラメータ数を削減する方法である [11], [12]．この方法は，画像認識のコンペティションで高い成績を収めたことで注目を集めるようになった [13]．[11] では，VGG Net と呼ばれる方法を音響モデルに適用し，非常に高い性能を実現している [14]．もう一つの方法は，通常の線形変換と活性化関数 $\phi(Ax+b)$ の代わりに，$\phi(Ax+b)+x$ のように活性化関数を適用しない項を加える方法である [15]．この方法には，複数のモデルのアンサンブル効果があるという解釈もある ．[16] では，合計のパラメータの数を固定した上で層を深くする実験を行っており，48 層で最も高い精度を実現している．これらに加え以前から，時間方向に非常に深いモデルと解釈できる LSTM や GRU を利用した RNN の利用も広がっており，今回の ICASSP でも多くの発表があった [17], [18]．

その他，耐雑音や残響環境での利用に焦点を当てた研究として，雑音・残響混じりの音声とクリーン音声の対からなるパラレルデータの活用や，個別に学習されたニューラルネットワークを最後に統合的に学習するアプローチが提案されている．パラレルデータの活用は，過去には SPLICE 関連の研究で盛んに検討されていたが，ニューラルネットワークにおいても Denoising autoencoder という形で雑音・残響音声からクリーン音声を推定する研究が進められている．文献 [19] では，残響環境に焦点を当てたパラレルデータの活用とともに，オートエンコーダの学習に音素事後確率を同時推定するマルチタスクラーニングの考えを取り入れ，残響環境の性能向上に成功した．また，文献 [20] では，Denoising autoencoder と並行して，話者識別，音素識別用のニューラルネットワークを個別に学習し，それらを話者性や音響環境を代表するネットワークとして捉え，最後に各ネットワークを統合して学習するいわゆる Joint Training の一手法を提案している．

他方，教師無し・半教師付き学習に関連した研究として，Neil 等は Deep Scattering Spectrum を ABnet の入力として用いて学習した特徴量について報告している [21]．ABnet は Siamese Network の一種であり，同じ音素に属する特徴量フレームが入力されたときに類似した特徴量を出力し，そうでない場合はそれらの差異が大きくなるように学習を行う．教師なし学習を行う際には，教師なし Spoken term discovery により類似音声セグメント対を求め，そらをフレームレベルでアライメントすることにより学習サンプルを得ている．実験では音声情報の損失の大きいメルフィルタバンク特徴量を入力とする場合よりも，より音素識別に有効な特徴量が得られることを示している．Ali 等は Active Learning における発話選択において，少量のラベル付きデータから学習した音声認識システムを使用する教師付き手法と，ラベル付きデータをまったく使用しない教師なし手法について検討を行っている [22]．ランダムな選択を行ったベースラインからの認識性能の向上はあまり大きいとは言えないものの，教師なし手法においても教師

付き手法に近い性能が得られることを示している．

2.3 HLT

言語関係（HLT, Human Language Technology）のセッションでは，口頭発表として低計算資源の音声認識，言語モデル，キーワード検索の 3 セッション，ポスター発表として音声言語理解（2 件）および言語獲得・対話の 3 セッションが構成された．

低計算資源の音声認識のセッションでは，モバイルデバイスに搭載できるように，パラメータの多いニューラルネットワークの圧縮・削減を行う研究が報告されており，実用上の課題への取り組みといえる．たとえば Google は，音響モデル（LSTM）の重み行列の特異値分解（SVD）による圧縮やパラメータの 8 ビット整数化（量子化），言語モデルの分割・圧縮などに基づく，スマートフォン端末で動作するフットプリント 20MB のシステムを報告している [23]．このシステムでは，端末にあるユーザの連絡先リストを辞書の拡張に利用しているが，このために必要な G2P モジュールも LSTM でコンパクトに構築している．

言語モデルの主要なトピックはニューラルネットワークのモデルであり，このうち RNN・LSTM モデル学習の方法論として，最小単語誤り基準に基づく識別学習法の提案があった [24]．ここでは，N-best 仮説の単語誤り数の期待値に対する各仮説の単語誤り数の大小によって，伝播される誤差信号に正負の強調が行われることとなる．AMI および CSJ コーパスにおける評価で，RNN・LSTM のいずれの場合も，通常のクロスエントロピー基準に基づくモデルに対して性能の改善が得られている．また，自然言語処理の分野で盛んに用いられるようになった分散表現を用いた言語モデルも報告された [25]．ニューラルネットワークにおいて，入力の単語履歴ベクトルから分散表現を求めて追加の特徴として利用する．あわせて隠れ層も拡張し，追加の重みを学習して出力に反映する．フィードフォワード型およびリカレント型のニューラルネットワークで放送ニュースタスクにおける評価実験が行われており，特に前者のモデルで単語誤り率に改善が見られた．一方，ニューラルネットワークでないモデルとしては，音声翻訳における音声認識の言語モデルを，機械翻訳のフレーズベース翻訳モデルを言語モデル確率に組み込むことで適応する手法が報告された [26]．

このほかの話題としては，言語資源が少ない環境でのキーワード検索や，これまでの会議に引き続いて音声対話の意図やドメイン等の検出を RNN・CNN などのニューラルネットワークで行う研究などが目立った．

以上，本節では音声認識関連の研究動向を総括したが，日本人研究者からも同分野に係る研究発表が数多くあった．詳細は [27], [28], [29], [30], [31], [32] を参照されたい．（福田，秋田，伊藤，落合，鈴木，篠崎，樋口）

3. 話者認識・照合，言語・年齢推定

話者認識分野全体の傾向としては，i-vector と PLDA に基づくアプローチが依然として主流であり，頑健性の向上を目的として本アプローチを改良する研究が多く見られた．例えば，[33] では転移学習の考え方に基づき，ターゲットとは異なるドメインで学習された PLDA モデルからの KL divergence をターゲットドメインでの PLDA モデル学習の正則化項として用いることで，少量データに対する過学習を防ぐ手法が提案されている．特徴量の音素クラス（senone）を識別する DNN を i-vector 抽出に用いるアプローチも依然として活発に研究が続けられている．例えば [34] では，DNN から各フレームの特徴量の音素クラスが得られることを活用し，音素クラス依存の i-vector 抽出器を学習する手法が提案され，テキスト依存型／テキスト非依存型いずれにおいても話者照合精度の向上が確認されている．Senone-based i-vector を話者の年齢推定に用いる試みもなされた [35]．NIST SRE 2008 および 2010 の電話会話音声において，LDA で次元圧縮した i-vector からサポートベクトル回帰で年齢の対数を推定するシンプルな方法でも実年齢と高い相関を示す推定値が得られ，さらに特徴量正規化手法を組み合わせることにより $r > 0.9$ の非常に高い相関で年齢を推定できることが確認されている．また，DNN を用いた話者照合に関する研究の中でも特筆すべき試みとして，end-to-end の話者照合システムに関する報告があった [36]．これは登録音声と照合音声の音響系列を入力とする LSTM をそれぞれ構築し，これら LSTM の隠れ層の出力ベクトル同士のコサイン類似度から accept/reject の 2 値へロジスティック回帰することで単一ネットワーク構造での end-to-end の話者照合を実現する．“OK Google” をキーフレーズとしたテキスト依存型話者照合において，従来の i-vector/DNN + PLDA に基づく手法よりも高い精度を達成した．その他の話題として，[37] では，話者照合の新たなオープンソースツールキットである SIDEKIT（Speaker IDEntification toolKIT）を紹介している．Python で書かれているため内容理解および修正が容易でインストール等も簡単である．加えて，他のツールへの依存性が低く，アルゴリズムの実装における制約が少ない．今後の展望として，言語識別や話者ダイアリゼーションへの拡張，Theano との連携によるニューラルネットワークの利用も予定している．

言語認識分野では NIST 2015 language recognition evaluation (LRE) i-vector challenge を対象とした報告が 2 件あった．本チャレンジは共通の評価基盤として対象言語の i-vector のみが与えられ，そのバックエンド処理で言語認識精度を競う試みである．本年度は識別対象の 50 言語の他に複数の未知言語が含まれており，これら未知言語の検出に重点がおかれている点に特色があったため，いずれの

報告もこの問題にフォーカスしたものだった．例えば [38] では，既知言語の識別器に対するスコアを利用して未知言語を検出するアプローチと，未知言語クラス識別器を構築し未知言語を検出するアプローチとを比較し後者が最良の性能を与えることを示している．一方，i-vector に基づく言語推定システムの front-end を対象とした研究として，言語間の類似度の階層性を考慮した言語認識システムが提案された [39]．この手法では，学習データに含まれる各言語を i-vector のコサイン類似度に基づき階層的にクラスタリングすることで，言語間の類似度の階層的な構造を事前に推定する．得られた階層構造の各レベルにおいて下位クラスタの識別器をそれぞれ異なる基準で作成した i-vector の組合せを用いて作成することで最上位階層の識別では最も大雑把なクラス基準での識別が行われ下位層になるにつれ具体的な言語のペアを識別するような階層的な識別を可能とした．未知語を含まない NIST LRE 2007 で提案法は従来法よりも高い識別率を達成したことが示されている．（俵，浅見，塩田）

4. 音声合成・声質変換

音声合成・声質変換分野はオーラルセッション 3 つ，ポスターセッション 3 つから構成された．音声合成については他分野と同様に DNN などの深層学習が注目を集めており，特にテキスト音声合成に関する研究では，半数以上が何らかの形で深層学習を利用していた．一方，声質変換においては深層学習を利用した手法は少なく，Exemplar-based な手法や学習データにパラレルデータを用いない手法などが提案された [40], [41]．以下では深層学習を利用した手法についていくつかを紹介する．

これまで DNN を音響モデルとして用いた音声合成が多く提案されてきたが，DNN 音声合成と HMM 音声合成との相違点は様々あり，具体的にどの要素が音声合成の性能の向上に寄与しているかは明らかにされていなかった．[42] では，コンテキストから HMM の状態を推定するのに決定木ではなく DNN を使用した方が，DNN を使用する場合でも状態単位ではなくフレーム単位の予測を行った方が，自然性がそれぞれ大きく向上すること実験的に示している．[43] では，DNN に基づく音声合成におけるスペクトル特徴量のモデル化において，最終層を Conditional Restricted Boltzmann Machine (CRBM) とした Deep Conditional RBM を提案し，スペクトル包絡をより精緻にモデル化する事を可能とした．[44] では，LSTM を用いた音声合成において，忘却ゲートの機能のみを残した Simplified-LSTM を提案し，パラメータ数を約半分にしたまま同等の合成品質を維持できることを示している．[45] では，DNN 音声合成における話者適応において適応文のテキスト情報を用いない教師なし適応を実現するため，多数話者で学習した DNN のモデルから得られた出力と適応データをフレーム単位で比較し，最近傍のデータに相当するラベルをラベル情報とすることで教師なし適応の実現を測っている．クロスリンガル話者適応の実験において，教師あり適応にある程度近い品質を得る事が可能になっている．また，[46] では，DNN 音声合成と素片接続型合成のハイブリッドシステムが提案された．提案手法では，DNN 音声合成の出力あるいは中間層の bottleneck feature である context embedding から得られる分布間の KL ダイバージェンスをコストとして素片選択を行うことで，従来の素片接続型音声合成に比べ自然性が向上している．

[47] では，テキスト音声合成における継続長モデルに DNN を利用しており，継続長モデルのロバストな推定法を提案している．大規模データを用いる際には，学習データに外れデータが含まれていることが想定される．このようなデータに対するロバストなモデル推定には，例えば，出力分布のピークを重視し Minimum Generation Error を学習基準として用いることや，GMM を用い合成時に適切なミクスチャーを選択することが挙げられる．本論文では，最尤推定を用いるのではなく，目的関数へ β ダイバージェンスの導入を行うことで，分布のピークを重視し，かつ，外れデータへの影響が少ないモデル推定が可能となることを紹介している．[48] では，テキスト音声合成のための女性話者のデータを対象とした Glottal flow の予測について検討している．既存手法では自然音声から計算された Glottal flow を保持・選択して合成に用いられていたが，本論文では，DNN を用い音響特徴量から時間領域の Glottal flow の予測が直接行われている点が興味深い．また，入力の F0 を変更することでその F0 に対応する Glottal Flow の予測が精度良く行われている．[49] では，クロスリンガル HMM 音声合成において，不特定話者 DNN-HMM 音声認識システムが出力する事後確率を用いた状態マッピングを提案している．各言語の HMM 音声合成システムの状態毎に DNN 事後確率を求め，事後分布の KL 距離が最小となるように言語間の状態マッピングを行う．不特定話者音声認識用の DNN を利用することで，元言語と目標言語の話者間の違いを吸収することが可能となり，従来法から音質，話者性を改善した．（橋本，郡山，齋藤，高木）

参考文献

[1] Sainath, T. N., Weiss, R. J., Wilson, K. W., Narayanan, A. and Bacchiani, M.: Factored spatial and spectral multichannel raw waveform CLDNNs, *Proc. ICASSP*, pp. 5075–5079 (2016).
[2] Wisdom, S., Hershey, J., Le Roux, J. and Watanabe, S.: Deep unfolding for multichannel source separation, *Proc. ICASSP*, pp. 121–125 (2016).
[3] Heymann, J., Drude, L. and Haeb-Umbach, R.: Neural network based spectral mask estimation for acoustic beamforming, *Proc. ICASSP*, pp. 196–200 (2016).
[4] Xiao, X., Watanabe, S., Erdogan, H., Lu, L., Hershey, J.,

Seltzer, M., Chen, G., Zhang, Y., Mandel, M. and Yu, D.: Deep beamforming networks for multi-channel speech recognition, *Proc. ICASSP*, pp. 5745–5749 (2016).
[5] Takeda, R. and Komatani, K.: Sound source localization based on deep neural networks with directional activate function exploiting phase information, *Proc. ICASSP*, pp. 405–409 (2016).
[6] Miao, Y., Gowayyed, M., Na, X., Ko, T., Metze, F. and Waibel, A.: An empirical exploration of CTC acoustic models, *Proc. ICASSP*, pp. 2623–2627 (2016).
[7] Rao, K., Senior, A. and Sak, H.: Flat start training of CD-CTC-SMBR LSTM RNN acoustic models, *Proc. ICASSP*, pp. 5405–5409 (2016).
[8] Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P. and Bengio, Y.: End-to-end attention-based large vocabulary speech recognition, *Proc. ICASSP*, pp. 4945–4949 (2016).
[9] Chan, W., Jaitly, N., Le, Q. and Vinyals, O.: Listen, attend and spell: a neural network for large vocabulary conversational speech recognition, *Proc. ICASSP*, pp. 4960–4964 (2016).
[10] Lu, L., Zhang, X. and Renals, S.: On training the recurrent neural network encoder-decoder for large vocabulary end-to-end speech recognition, *Proc. ICASSP*, pp. 5060–5064 (2016).
[11] Sercu, T., Puhrsch, C., Kingsbury, B. and LeCun, Y.: Very deep multilingual convolutional neural networks for LVCSR, *Proc. ICASSP*, pp. 4955–4959 (2016).
[12] Yoshioka, T., Ohnishi, K., Fang, F. and Nakatani, T.: Noise robust speech recognition using recent developments in neural networks for computer vision, *Proc. ICASSP*, pp. 5730–5734 (2016).
[13] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A.: Going deeper with convolutions, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9 (2015).
[14] Sercu, T. and Goel, V.: Advances in Very Deep Convolutional Neural Networks for LVCSR, *arXiv preprint arXiv:1604.01792* (2016).
[15] He, K., Zhang, X., Ren, S. and Sun, J.: Identity mappings in deep residual networks, *arXiv preprint arXiv:1603.05027* (2016).
[16] Ghahremani, P., Droppo, J. and Seltzer, M. L.: Linearly augmented deep neural network, *Proc. ICASSP*, pp. 5085–5089 (2016).
[17] Liu, C., Wang, Y., Kumar, K. and Gong, Y.: Investigations on speaker adaptation of LSTM RNN models for speech recognition, *Proc. ICASSP*, pp. 5020–5024 (2016).
[18] Tang, Z., Wang, D. and Zhang, Z.: Recurrent neural network training with dark knowledge transfer, *Proc. ICASSP*, pp. 5900–5904 (2016).
[19] Qian, Y. and Tan, T.: An investigation into using parallel data for far-field speech recognition, *Proc. ICASSP*, pp. 5725–5729 (2016).
[20] Qian, Y., Tan, T., Yu, D. and Zhang, Y.: Integrated adaptation with multi-factor joint-learning for far-field speech recognition, *Proc. ICASSP*, pp. 5770–5774 (2016).
[21] Zeghidour, N., Synnaeve, G., Versteegh, M. and Dupoux, E.: A deep scattering spectrum - Deep Siamese network pipeline for unsupervised acoustic modeling, *Proc. ICASSP*, pp. 4965–4969 (2016).
[22] Syed, A. R., Rosenberg, A. and Kislal, E.: Supervised and unsupervised active learning for automatic speech recognition of low-resource languages, *Proc. ICASSP*, pp. 5320–5324 (2016).
[23] McGraw, I., Prabhavalkar, R., Alvarez, R., Arenas, M. G., Rao, K., Rybach, D., Alsharif, O., Sak, H., Gruenstein, A., Beaufays, F. and Parada, C.: Personalized speech recognition on mobile devices, *Proc. ICASSP*, pp. 5955–5959 (2016).
[24] Hori, T., Hori, C., Watanabe, S. and Hershey, J. R.: Minimum word error training of long short-term memory recurrent neural network language models for speech recognition, *Proc. ICASSP*, pp. 5990–5994 (2016).
[25] Audhkhasi, K., Sethy, A. and Ramabhadran, B.: Semantic word embedding neural network language models for automatic speech recognition, *Proc. ICASSP*, pp. 5995–5999 (2016).
[26] Pelemans, J., Vanallemeersch, T., Demuynck, K., Verwimp, L., Van hamme, H. and Wambacq, P.: Language model adaptation for ASR of spoken translations using phrase-based translation models and named entity models, *Proc. ICASSP*, pp. 5985–5989 (2016).
[27] Fukuda, T., Ichikawa, O. and Tachibana, R.: Convolutional neural network pre-trained with projection matrices on linear discriminant analysis, *Proc. ICASSP*, pp. 5345–5349 (2016).
[28] Li, S., Akita, Y. and Kawahara, T.: Data selection from multiple ASR system's hypotheses for unsupervised acoustic model training, *Proc. ICASSP*, pp. 5875–5879 (2016).
[29] Ito, N., Araki, S. and Nakatani, T.: Modeling audio directional statistics using a complex Bingham mixture model for blind source extraction from diffuse noise, *Proc. ICASSP*, pp. 465–468 (2016).
[30] Ochiai, T., Matsuda, S., Watanabe, H., Lu, X., Kawai, H. and Katagiri, S.: Bottleneck linear transformation network adaptation for speaker adaptive training-based hybrid DNN-HMM speech recognition, *Proc. ICASSP*, pp. 5015–5019 (2016).
[31] Suzuki, M., Kurata, G. and Tachibana, R.: Speech recognition robust against speech overlapping in monaural recordings of telephone conversations, *Proc. ICASSP*, pp. 5685–5689 (2016).
[32] Higuchi, T., Ito, N., Yoshioka, T. and Nakatani, T.: Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise, *Proc. ICASSP*, pp. 5210–5214 (2016).
[33] Hong, Q., Zhang, J., Li, L., Wan, L. and Tong, F.: A transfer learning method for PLDA-based speaker verification, *Proc. ICASSP*, pp. 5455–5459 (2016).
[34] Chen, L., Lee, K. A., Chng, E.-S., Ma, B., Li, H. and Dai, L. R.: Content-aware local variability vector for speaker verification with short utterance, *Proc. ICASSP*, pp. 5485–5489 (2016).
[35] Sadjadi, S. O., Ganapathy, S. and Pelecanos, J. W.: Speaker age estimation on conversational telephone speech using senone posterior based i-vectors, *Proc. ICASSP*, pp. 5040–5044 (2016).
[36] Heigold, G., Moreno, I., Bengio, S. and Shazeer, N.: End-to-end text-dependent speaker verification, *Proc. ICASSP*, pp. 5115–5119 (2016).
[37] Larcher, A., Lee, K. A. and Meignier, S.: An extensible speaker identification sidekit in python, *Proc. ICASSP*, pp. 5095–5099 (2016).
[38] Yu, C., Zhang, C., Ranjan, S., Zhang, Q., Misra, A., Kelly, F. and Hansen, J. H. L.: UTD-CRSS system for

the NIST 2015 language recognition i-vector machine learning challenge, *Proc. ICASSP*, pp. 5835–5839 (2016).
[39] Irtza, S., Sethu, V., Bavattichalil, H., Ambikairajah, E. and Li, H.: A hierarchical framework for language identification, *Proc. ICASSP*, pp. 2820–2824 (2016).
[40] Ming, H., Huang, D., Xie, L., Zhang, S., Dong, M. and Li, H.: Exemplar-based sparse representation of timbre and prosody for voice conversion, *Proc. ICASSP*, pp. 5175–5179 (2016).
[41] Agiomyrgiannakis, Y.: The matching-minimization algorithm, the INCA algorithm and a mathematical framework for voice conversion with unaligned corpora, *Proc. ICASSP*, pp. 5645–5649 (2016).
[42] Watts, O., Henter, G. E., Merritt, T., Wu, Z. and King, S.: From HMMs to DNNs: Where do the improvements come from?, *Proc. ICASSP*, pp. 5505–5509 (2016).
[43] Yin, X., Ling, Z.-H., Hu, Y.-J. and Dai, L.-R.: Modeling spectral envelopes using deep conditional restricted Boltzmann machines for statistical parametric speech synthesis, *Proc. ICASSP*, pp. 5125–5129 (2016).
[44] Wu, Z. and King, S.: Investigating gated recurrent networks for speech synthesis, *Proc. ICASSP*, pp. 5140–5144 (2016).
[45] Fan, Y., Qian, Y., Soong, F. K. and He, L.: Unsupervised speaker adaptation for DNN-based TTS synthesis, *Proc. ICASSP*, pp. 5135–5139 (2016).
[46] Merritt, T., Clark, R. A. J., Wu, Z., Yamagishi, J. and King, S.: Deep neural network-guided unit selection synthesis, *Proc. ICASSP*, pp. 5145–5149 (2016).
[47] Henter, G., Ronanki, S., Watts, O., Wester, M., Wu, Z. and King, S.: Robust TTS duration modelling using DNNs, *Proc. ICASSP*, pp. 5120–5124 (2016).
[48] Juvela, L., Bollepalli, B., Airaksinen, M. and Alku, P.: High-pitched excitation generation for glottal vocoding in statistical parametric speech synthesis using a deep neural network, *Proc. ICASSP*, pp. 5130–5134 (2016).
[49] Xie, F., Soong, F. and Li, H.: A KL divergence and DNN approach to cross-lingual TTS, *Proc. ICASSP*, pp. 5515–5519 (2016).