Vol.2016-SLP-110 No.4
2016/2/5

国際会議INTERSPEECH2015参加報告

浅見 太一[1]　大谷 大和[2]　小川 哲司[3]　木下 慶介[1]　倉田 岳人[4]　齋藤 大輔[5]　塩田 さやか[6]
太刀岡 勇気[7]　中村 静[8]　増村 亮[1]　渡部 晋治[9]

概要：2015年9月6日から10日にかけてドイツ・ドレスデンで開催されたISCA主催のINTERSPEECH2015に参加した．INTERSPEECHは音声言語処理分野で一流の国際会議である．ここでは海外からの発表を中心に，最新の技術動向，注目すべき発表について報告する．

1. はじめに

2015年9月6日から10日にかけてドイツ・ドレスデンで開催されたISCA主催のINTERSPEECH2015に参加した．INTERSPEECHは音声言語処理分野で一流の国際会議である．INTERSPEECHは音声，言語に関する研究を広く取り扱った本研究分野におけるトップレベルの会議である．通常論文の投稿数は1458件あり，採択数は746件（受理率51%）であった．本稿では筆者らが注目する研究をいくつかピックアップし，INTERSPEECH2015ならびに関連ワークショップについて最新の技術動向および注目すべき発表について紹介する．

2. 音声認識（フロントエンド・音響モデル）

近年，それぞれ個別に最適化されていた音声認識のフロントエンド処理と音響モデルを，Deep Neural Network (DNN) の枠組みで同時最適化する手法が注目されている．例えば文献 [1] では，1) マスク推定型の音声強調，2) 強調音声の対数メルフィルタバンクによる特徴抽出，3) 音響モデルを，一つのネットワークで表現し，音声強調ネットワーク，フィルタバンク係数を初期値とする特徴量の線形変換，及び音響モデルネットワークのパラメータを，クロスエントロピー基準で一括学習する手法を提案している．文献 [2] においても，同様のコンセプトを，マルチタスク学習及び反復学習の枠組みで実現する手法を提案している．両タスクともCHIME-2コーパスにおいてその有効性を示している．文献 [3] では，特徴量抽出過程をもニューラルネットワークでモデル化・一括最適化するCLDNNsが提案されている．CLDNNsはConvolutional Neural Network (CNN)，Long Short-Term Memory (LSTM)，全結合型DNNを重ねた形で表現されており，時間領域信号そのものを入力とし，HMM状態事後確率を出力する．入力信号として（対数メルスペクトルに加え）時間領域信号を用いることで，対数メルスペクトルのみを用いる場合よりも高い認識性能を達成できることが報告されている．文献 [4] では，変調周波数スペクトルをDNNの入力として用いる音声認識システム (TRAP) において，変調周波数フィルタと音響モデルパラメータをDNNで一括最適化する枠組みを提案している．この枠組みでは，DNNの前段に畳み込み層を入れ，帯域ごとの対数フィルタバンク出力のトラジェクトリをその入力とすることで，変調周波数フィルタの学習を実現している．DCTに基づくフィルタを用いて抽出した変調周波数スペクトルをDNNの入力とした場合と比較して性能の向上が報告されている．

以上のように，DNNを用いることによる要素技術の全体最適化に多くの注目が集まる中，DNNの動作をさらに深く分析・理解し，新たなモデルの提案につなげようとする試みも報告されている．例えば，文献 [5] では，DNNの最適化により，中間層では自律的に各音素や調音点，調音法に対応するノードが構成されることが報告されている．

また，音響モデルで標準技術になりつつあるLSTMなどのRecurrent Neural Network (RNN) と同等の性能を持つDNNを構築するために，LSTMで得られるsoft alignmentをターゲットにしたクロスエントロピー基準でDNNを学習する手法も提案されている [6]．これらは，モデル圧縮技術として機械学習で広く用いられている手法である．この圧縮されたDNNを用いて，認識時の音響スコア計算を，

[1] 日本電信電話株式会社
[2] 東芝
[3] 早稲田大学
[4] 日本IBM
[5] 東京大学
[6] 首都大学東京
[7] 三菱電機株式会社
[8] 京都大学
[9] Mitsubishi Electric Research Laboratories

高速且つ省メモリで行うことができる．RNN の学習時のコストを削減するための方式としては，標準的なフィードフォワードネットワークと同等の学習時間で長時間の依存関係をモデル化可能な time delay neural network (TDNN) が提案されている [7]．提案の枠組みは，サブサンプリングを用いることで標準的な TDNN の学習に対し 5 倍の高速化に成功し，Switchboard コーパスを用いた大語彙連続音声認識において RNN の誤りを削減したことが報告されている．（小川, 木下, 渡部）

3. ロバスト音声認識

ロバスト音声認識に関して特筆すべきものとして，今回の INTERSPEECH では「Robust speech processing using observation uncertainty and uncertainty propagation」と題する special session (SS) が企画された．SS では 9 件のポスター発表があり，そのうち 2 件話者認識で，残りの 7 件が音声認識に関するものであった．GMM においては特徴量の分布において uncertainty を理論的に扱えることから広く使われていたものの，DNN においては理論的な扱いが難しい．これに対していくつかの提案がされた．代表的なのは分布をモデル化することをあきらめて，サンプリングを使う手法である [8], [9], [10]．[9] では Monte Carlo sampling と unscented 変換により，DNN の隠れ層における平均と共分散を推定する試みが行われている．[10] では特徴量のサンプリングを行いその平均を認識することで，uncertainty を考慮したデコーディングを行っている．このほかに，DNN ならではの方法として，フロントエンドと連携し DNN による音声強調の uncertainty を DNN の音響モデルで扱う試み [11] もあった．

そのほかにロバスト音声認識に関わる特徴量・遠隔・残響および適応化のセッションがあった．[12] では長いコンテキストを扱うために，TDNN を使っており，大規模データで効果を確認していた．DNN 関連の研究では，以前検討されていた事柄を大規模データに適用することで効果を得ることがよくあるが，これもその温故知新の一例といえよう．適応化では，従来通り fMLLR を用いたもの [13] や i-vector を用いたもの [14] がみられた．（太刀岡）

4. 言語モデル・音声言語理解

言語モデルにおいても，深層学習による高精度なモデル化に関する研究が主流であったが，音声認識システムの普及に伴い，実際のアプリケーションで生じている問題への対応に関する研究発表も多く行われていた．深層学習の利用では，単語とサブワードをモデル化の単位として，LSTM で同時にモデル化する手法が提案されていた [15]．初期段階の実験で大きな改善は報告されていないが，モデル化の単位を検討することで，日本語への適用も可能と考えられる．認識時のレイテンシの問題などから，n-gram モデルは引き続き幅広く利用されているが，その枠組みの中で，音声認識システムの実運用時に生じる問題を検討している論文を 3 本紹介する．アプリケーションやユーザの状況に依存した発話は，1 種類の大規模な言語モデルで対応することは困難である．それに対して，アプリケーション・状況に依存した表現・単語を，on-the-fly で重み付けする方法が提案されていた [16], [17]．また，多くのアプリケーションで重要となる数字，時間，住所のような表現は，クラス n-gram モデルで対応することが一般的であるが，適切なクラス情報を持つコーパスを用意することは困難であった．事前に CRF に基づく系列ラベリングを学習コーパスに対して適用することで，コンテキストを考慮したクラス情報を持つコーパスを用意し，当該クラスに属する表現の音声認識率を向上させる方法が提案されていた [18]．深層学習による高精度なモデル化と，実際のアプリケーションで生じている問題への対応の両方が，今後も重要な研究テーマとなるだろう．

音声言語理解においては近年の傾向に引き続き，音声対話のための発話意図識別やスロットフィルタリングなどの分野において，RNN や LSTM，CNN の適用が検討されていた．その中でも，単に深層学習のモデルを適用する流れは終わりを迎えつつあり，様々な拡張が検討されていた．[19] では，発話意図識別で RNN や LSTM を適用する際に，珍しい単語や未知語が含まれている場合でも頑健な処理を行うために，部分的に文字ベースのモデル化を行う方法が検討されていた．一方 [20] では，テキスト分類タスクにおいて，音声認識誤りに頑健な処理を行うため，音声認識結果のラティスを直接 CNN の入力として利用できるように拡張する方法が提案されていた．さらに，単に一つの識別タスクをモデル化するだけではなく，複数タスクを同時にモデル化するマルチタスクラーニングに関する報告も見られた．[21] では，RNN による系列のモデル化の際に，次の単語の予測とラベルの予測を同時にモデル化する方法を提案しており，単語系列と教師ラベルのペアを疑似的に生成するために利用していた．また [22] では，リカレント構造を持つ CNN において，意図決定とスロットフィルタリングを同時にモデル化することで，性能向上を報告していた．これらに加えて，話し言葉を扱う研究を 1 つ紹介したい．自然言語処理の領域では，Word2Vec 等の単語ベクトルを，大量の書き言葉テキストから教師なしで獲得して利用することが一般的になりつつある．一方で音声言語の領域では，話し言葉に適した単語ベクトルが必要になるが，話し言葉テキストを大量に集めることは困難であるため同様の方法は利用できない．それに対して [23]，大量の書き言葉テキストから獲得した単語ベクトルを少量の話し言葉テキストで適応する方法が提案されていた．現状，音声言語理解の研究では書き言葉を扱うことが一般的であるが，話し言葉に対する深層学習を用いた技術検討が今後さらに

重要になるであろう．（倉田, 増村）

5. 話者認識・話者照合

INTERSPEECH2015 では，話者照合に対する音声合成技術を用いたなりすまし攻撃への対策方法を比較したスペシャルセッション ASVspoof2015 が行われた．この企画では，合成音声によるなりすまし攻撃と登録者本人の音声を識別することがタスクとなる．なりすまし攻撃としてモデル学習法やデータ数，特徴量，手法の既知／未知などの条件を変えた 10 種類の音声合成手法が用意され，それぞれのなりすまし攻撃に対する性能を競うものとなっている．詳細および各機関の論文については http://www.spoofingchallenge.org で確認できる．

なりすまし攻撃全種類に対する EER の総合第 1 位は DA-IICT のシステムであった．[24] より，特徴量に MFCC だけでなく蝸牛フィルタを用いた cochlear filter cepstral coefficients (CFCC) と CFCC に位相情報の Instantaneous Frequency (IF) を組み合わせた CFCCIF を用い，モデル化手法には従来の GMM を用いたシステムであることが紹介されている．これらの特徴量を合わせて使うことで詐称攻撃の中でもっとも難しい波形接続型の攻撃に対しても最も高い EER を得られることが報告されている．総合第 4 位の NTU のシステムは，波形接続型を除いた攻撃に対しては EER がほぼ 0%であり，波形接続型を除いた結果では第 1 位であった．[25] より，Log Magnitude Spectrum (LMS), Residual Log Magnitude Spectrum (RLMS), 群遅延, 修正群遅延，Instantaneous Frequency Derivative (IF)，Baseband Phase Difference (BPD)，Pitch Synchronous Phase (PSP) の 7 種類を特徴量として用い，特徴量毎に MLP を学習し，スコア統合したシステムとなっている．総合第 5 位の CRIM のシステムも上記のシステム同様に特徴量を工夫したものであった．[26] より，特徴量に MFCC と Cosine Normalized Phase-based Cepstral Coefficient (CNPCC) を結合した特徴量 MFCC-CNPCC を用いたものと Linear Prediction Residual Cepstral Coefficients (LPRCC) を特徴量とした 2 手法についてそれぞれ高い精度が得られていることが報告されている．同機関の論文である [27] においても，同じ枠組みである LP residual phase cepstra (LPRPC) を特徴量として用いることで Short utterance の話者認識においても高い性能を得ることが報告されていた．全体な傾向としては，モデリングで性能向上を目指すよりも特徴量に様々な手法を用いることによる性能向上を図る機関が多く，特に位相情報を様々な手法で抽出していることが特徴として挙げられる．

近年成功を収めている，話者認識向けの特徴量抽出にニューラルネットを適用する手法についても多くの報告があった．ニューラルネットの利用方法は，音声認識用の DNN から得られる HMM 状態（senone とも呼ばれる）の事後確率を i-vector 抽出過程の統計量計算に用いるアプローチと，話者を識別するように学習した DNN の隠れ層の出力を特徴量として用いるアプローチに大別される．DNN を i-vector 抽出過程の統計量計算に用いるアプローチは既に多くの成功報告があり，[28] では様々な実験条件で詳細な精度比較を行った結果が報告された．話者適応（fMLLR）の利用は効果がない，UBM には対角共分散行列よりも全共分散行列を使った方が精度が高い，senone の数は（GMM の混合数とは振る舞いが異なり,）増やせば増やすほど精度が向上する，といった知見が実験により示されている．ニューラルネットの構造の高度化も検討され，[29] では，DNN を LSTM に置き換えることにより，NIST 2008 のコアテスト 8 条件のうち 6 条件で性能が大きく改善することが確認されている．DNN の隠れ層の出力を特徴量として用いるアプローチでは，出力層の自由度が高いため，マルチタスク学習の枠組みによって様々な性質を併せ持つ特徴量を構成できる．[30] では，発声されたフレーズを話者 ID と同時に識別する DNN をマルチタスク学習し，この DNN の隠れ層の出力を特徴量として用いる手法を提案し，テキスト依存型話者照合での大きな精度改善を得ている．[31] では，spoofing detection において，攻撃か否かと同時に攻撃に使われた手法を識別する DNN を用いることで，ASVspoof 第 3 位の性能を達成している．教師ラベル付き学習データが必要となるものの，ラベルの付与方法次第で様々な工夫の余地があり，今後の発展が期待されるアプローチだと考えられる．（浅見, 塩田）

6. 言語教育応用（SLaTE 報告）

SLaTE（Speech and Language Technology in Education）は，教育への音声言語情報処理技術の応用に関する，ISCA の Special Interest Group による INTERSPEECH のサテライトワークショップである．2007，2009，2010，2011, 2013 年に開催され, 今回で 6 回目となる．今回は, 32 件の論文と 6 件のデモが採択され, INTERSPEECH 直前の 9 月 4，5 日に Leipzig で開催された．研究発表の日程はこれまでの 3 日間から初めて 2 日間に短縮されたが，73 名という参加者数も発表件数も前回と同程度であった．今回は，口頭発表 5 セッション（Automatic Assessment，Assessment and Practice，Grammar，Pronunciation Analysis, Text），ポスター発表 2 セッション（From Pronunciation to Conversation，From Phones to Serious Games），および，デモセッションで構成された．

音声言語情報処理技術の進歩に伴い，計算機支援型言語学習 CALL（Computer-Assisted Language Learning）に関する研究も発展を遂げてきている．近年の新たな傾向の一つとして，合成音声の品質の向上により，学習者が母語話者ほど実音声と合成音声の自然性の違いに敏感ではないことを利用して，学習対象言語の音声の手本として，従来

の母語話者による発話を録音した実音声の代わりに，合成音声を利用する試みが増加していることが挙げられる．以下では，これに関する発表について紹介する．

[32] では，アイルランド語リスニング CALL システムを用いて，このシステムで利用される合成音声の許容度の評価に加えて，ユーザの合成音声一般に対する事前の態度による評価への影響と，音声以外の要素による評価への影響の調査が行われた．UNESCO が危機言語に分類するアイルランド語は，アイルランドの第一公用語であるが，多くの国民にとって義務教育での必修に過ぎず，日常的な利用は少ない．近年その復興政策を政府が実施しているが，ダブリン大学ではその CALL システムとして，合成音声を用いた仮想現実の学習環境を備えた Fáilte go TCD が開発されている．合成音声は，ダブリン大学で開発されたアイルランド語 TTS システム ABAIR を利用して作成された．被験者 252 人による明瞭さ，品質，魅力についての評価では，各々 64.7，72.2，62.3 %が中立あるいは肯定的であると答えた．実験の結果，合成音声一般に対して事前に好印象を持っていたユーザほどこの CALL システムでの合成音声の許容度が高く，事前の態度が評価に強く影響していることが示された．また，キャラクタの動きや方言の使われ方等音声以外の種々の要素も評価に影響していることが示され，CALL システムでの合成音声の許容度はシステム全体の質次第のようであると考察されている．母語話者の少ない危機言語にとって，音声合成技術は教育資源の拡充という重要な役割を担う．より良い学習環境を提供するために，音声合成をはじめとする関連技術のさらなる発展が期待されている． （中村）

7. 音声合成・声質変換

音声合成ならびに声質変換の関連セッションはオーラル 5 つ，ポスター 3 つで構成されていた．5 つのオーラルセッションは，DNN，統計的パラメトリック音声合成，韻律モデリング，声質変換ならびに音声合成の評価に関するもので，それぞれこの分野における主要な話題を取り扱っているといえる．以下では著者らの注目する発表を紹介する．

[33] では，DNN 音声合成における学習と合成の間にある本質的な不一致の問題に取り組んでいる．従来のモデル学習では，フレームごとに出力静的・動的特徴量と学習データの誤差が最小になるようにモデルパラメータの更新が行われている．しかし，合成では出力静的・動的特徴量系列からパラメータ生成アルゴリズムにより生成された静的特徴量系列を用いているため，モデルの学習基準が合成処理に対して一致していないと考えられる．これを解決するために，出力静的・動的特徴量から生成された静的特徴量系列と学習データの静的特徴量系列との誤差 (Sequence generation error: SGE) が最小となるようなモデル学習を提案している．主観評価において提案法により改善がみられたと報告している．

[34] では，文献 [33] と同様に，学習と合成間の不一致に着目した学習法を提案している．基本的には SGE と同じく静的特徴量系列の誤差最小化基準による学習 (Minimum trajectory error: MTE) を行うが，勾配の求め方が SGE とやや異なっている．さらにこの文献では，コンテキスト情報をよりよく捉えるために入力コンテキストに関するボトルネック特徴量を導入している．評価実験では，MTE の導入より音質が向上し，さらにボトルネック特徴量によって音質が改善したことを示している．

[35] では，DNN ベースの音声認識で提案されている 3 つの話者適応手法を DNN 音声合成に導入し，その性能を評価している．この文献では話者適応手法として，入力層に i-vector と gender-code，隠れ層に learning hidden unit contributions, 出力層に feature transform を適用している．また，出力層での feature transform として声質変換で用いられている混合正規分布モデルによる特徴量変換法を採用している．評価では，適応文数を 10 文および 100 文とした場合の HMM 音声合成の話者適応を自然性および話者類似性の観点で評価している．評価結果では，すべての場合において DNN 音声合成の性能が上回っている．

[36] では，F0 パターンを連続ウェーブレット変換 (Continuous wavelet transform: CWT) によって分解してモデル化する手法について，異なるスケールの聴感的な影響を大規模な聴取実験によって明らかにしている．結果として，CWT によって分解されたスケールのうち，中間スケールがもっとも自然性評価に影響を与える一方，低い（変調）周波数に対応するスケールが HMM 音声合成によるフレーム単位のモデル化の結果に近くなるということが示された．

[37] は，音声合成研究における主観評価を定量的に分析することで，主観評価実験に必要な聴取者の数，実験設計について議論している．この文献では，INTERSPEECH2014 の音声合成に関連する研究発表で実施された聴取実験のうち，60 % は 20 人以下の主観評価実験に基づくものである一方，Blizzard Challenge 2013 の分析から，MOS テストにおける自然性評価を安定したものとするためには 30 人以上の聴取者が必要であるとの結果が示された．その他，聴取者がエキスパートかどうか，クラウドベースかどうかなど，様々な観点から主観評価実験への結果の影響を調査しており，音声合成研究における主観評価実験の設計指針について議論した興味深い内容である． （大谷, 齋藤）

参考文献

[1] Wang, Z.-Q. and Wang, D.: Joint training of speech separation, filterbank and acoustic model for robust automatic speech recognition, *Proc. Interspeech*, pp. 3264–3268 (2015).
[2] Chen, Z., Watanabe, S., Erdogan, H. and Hershey, J. R.: Speech enhancement and recognition using multi-task

learning of long short-term memory recurrent neural networks, *Proc. Interspeech*, pp. 3274–3278 (2015).
[3] Sainath, T. N., Weiss, R. J., Senior, A., Wilson, K. W. and Vinyals, O.: Learning the speech front-end with raw waveform CLDNNs, *Proc. Interspeech*, pp. 1–5 (2015).
[4] Pesán, J., Burget, L., Hermanský, H. and Veselý, K.: DNN derived filters for processing of modulation spectrum of speech, *Proc. Interspeech*, pp. 1908–1911 (2015).
[5] Nagamine, T., Seltzer, M. L. and Mesgarani, N.: Exploring how deep neural networks form phonemic categories, *Proc. Interspeech*, pp. 1912–1916 (2015).
[6] Chan, W., Ke, N. R. and Lane, I.: Transferring knowledge from a RNN to a DNN, *Proc. Interspeech*, pp. 106–111 (2015).
[7] Peddinti, V., Povey, D. and Khudanpur, S.: A time delay neural network architecture for efficient modeling of long temporal contexts, *Proc. Interspeech*, pp. 3214–3218 (2015).
[8] Tachioka, Y. and Watanabe, S.: Uncertainty training and decoding methods of deep neural networks based on stochastic representation of enhanced features, *Proc. Interspeech*, pp. 3541–3545 (2015).
[9] Abdelaziz, A., Watanabe, S., Hershey, J., Vincent, E. and Kolossa, D.: Uncertainty propagation through deep neural networks, *Proc. Interspeech*, pp. 3561–3566 (2015).
[10] Huemmer, C., Maas, R., Schwarz, A., Astudillo, R. and Kellermann, W.: Uncertainty decoding for DNN-HMM hybrid systems based on numerical sampling, *Proc. Interspeech*, pp. 3556–3560 (2015).
[11] Astudillo, R., Correia, J. and Trancoso, I.: Integration of DNN based speech enhancement and ASR, *Proc. Interspeech*, pp. 3576–3581 (2015).
[12] Peddinti, V., Chen, G., Povey, D. and Khudanpur, S.: Reverberation robust acoustic modeling using i-vectors with time delay neural networks, *Proc. Interspeech*, pp. 2440–2444 (2015).
[13] Lu, L. and Renals, S.: Feature-space speaker adaptation for probabilistic linear discriminant analysis acoustic models, *Proc. Interspeech*, pp. 2862–2866 (2015).
[14] Garimella, S., Mandal, A., Strom, N., Hoffmeister, B., Matsoukas, S. and Parthasarathi, S.: Robust i-vector based adaptation of DNN acoustic model for speech recognition, *Proc. Interspeech*, pp. 2877–2881 (2015).
[15] Arisoy, E. and Saraclar, M.: Multi-stream long short-term memory neural network language model, *Proc. Interspeech*, pp. 1413–1417 (2015).
[16] Hall, K., Cho, E., Allauzen, C., Beaufays, F., Coccaro, N., Nakajima, K., Riley, M., Roark, B., Rybach, D. and Zhang, L.: Composition-based on-the-fly rescoring for salient n-gram biasing, *Proc. Interspeech*, pp. 1418–1422 (2015).
[17] Aleksic, P., Ghodsi, M., Michaely, A., Allauzen, C., Hall, K., Roark, B., Rybach, D. and Moreno, P.: Bringing contextual information to Google speech recognition, *Proc. Interspeech*, pp. 468–472 (2015).
[18] Vasserman, L., Schogol, V. and Hall, K.: Sequence-based class tagging for robust transcription in ASR, *Proc. Interspeech*, pp. 473–477 (2015).
[19] Ravuri, S. and Stolcke, A.: Recurrent neural network and LSTM models for lexical utterance classification, *Proc. Interspeech*, pp. 135–139 (2015).
[20] Svec, J., Chýlek, A. and Smídl, L.: Hierarchical discriminative model for spoken language understanding based on convolutional neural network, *Proc. Interspeech*, pp. 1864–1868 (2015).
[21] Tam, Y.-C., Shi, Y., Chen, H. and Hwang, M.-Y.: RNN-based labeled data generation for spoken language understanding, *Proc. Interspeech*, pp. 125–129 (2015).
[22] Liu, C., Xu, P. and Sarikaya, R.: Deep contextual language understanding in spoken dialogue systems, *Proc. Interspeech*, pp. 120–124 (2015).
[23] Tafforeau, J., Artieres, T., Favre, B. and Bechet, F.: Adapting lexical representation and OOV handling from written to spoken language with word embedding, *Proc. Interspeech*, pp. 1408–1412 (2015).
[24] Patel, T. B. and Patil, H. A.: Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech, *Proc. Interspeech*, pp. 2062–2066 (2015).
[25] Xiao, X., Tian, X., Du, S., Xu, H., Chng, E. S. and Li, H.: Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for ASVspoof 2015 Challenge, *Proc. Interspeech*, pp. 2052–2056 (2015).
[26] Alam, M. J., Kenny, P., Bhattacharya, G. and Stafylakis, T.: Development of CRIM system for the Automatic Speaker Verification Spoofing and Countermeasures Challenge 2015, *Proc. Interspeech*, pp. 2072–2076 (2015).
[27] Alam, M. J., Kenny, P. and Stafylakis, T.: Combining amplitude and phase-based features for speaker verification with short duration utterances, *Proc. Interspeech*, pp. 249–253 (2015).
[28] Romero, D. G. and McCree, A.: Insights into deep neural networks for speaker recognition, *Proc. Interspeech*, pp. 1141–1145 (2015).
[29] Zheng, H., Zhang, S. and Liu, W.: Exploring robustness of DNN/RNN for extracting speaker Baum-Welch statistics in mismatched conditions, *Proc. Interspeech*, pp. 1161–1165 (2015).
[30] Chen, N., Qian, Y. and Yu, K.: Multi-task learning for text-dependent speaker verification, *Proc. Interspeech*, pp. 185–189 (2015).
[31] Chen, N., Qian, Y., Dinkel, H., Chen, B. and Yu, K.: Robust deep feature for spoofing detection – The SJTU system for ASVspoof, *Proc. Interspeech*, pp. 2097–2101 (2015).
[32] Chiaráin, N. N. and Chasaide, A. N.: Evaluating synthetic speech in an Irish CALL application: influences of predisposition and of the holistic environment, *Proc. SLaTE*, pp. 149–154 (2015).
[33] Fan, Y., Qian, Y., Soong, F. K. and He, L.: Sequence generation error (SGE) minimization based deep neural networks training for text-to-speech synthesis, *Proc. Interspeech*, pp. 864–868 (2015).
[34] Wu, Z. and King, S.: Minimum trajectory error training for deep neural networks, combined with stacked bottleneck features, *Proc. Interspeech*, pp. 309–313 (2015).
[35] Wu, Z., Swietojanski, P., Veaux, C., Renals, S. and King, S.: A study of speaker adaptation for DNN-based speech synthesis, *Proc. Interspeech*, pp. 879–883 (2015).
[36] Ribeiro, M., Yamagishi, J. and Clark, R.: A perceptual investigation of wavelet-based decomposition of $f0$ for text-to-speech synthesis, *Proc. Interspeech*, pp. 1586–1590 (2015).
[37] Wester, M., Valentini-Botinhao, C. and Henter, G.: Are we using enough listeners? No! An empirically-supported critique of Interspeech 2014 TTS evaluations, *Proc. Interspeech*, pp. 3476–3480 (2015).