Vol.2015-SLP-107 No.3
2015/7/16

国際会議ICASSP2015参加報告

岡本 拓磨[1]　小川 哲司[2]　落合 翼[3]　柏木 陽佑[4]　亀岡 弘和[5,4]　木下 慶介[5]　郡山 知樹[6]
齋藤 大輔[4]　篠崎 隆宏[6]　高木 信二[7]　滝口 哲也[8]　太刀岡 勇気[9]　俵 直弘[2]　橋本 佳[10]
藤本 雅清[5]　松田 繁樹[3]　三村 正人[11]　吉岡 拓也[5]　渡部 晋治[12]

概要：2015年4月19日から24日にかけオーストラリア・ブリスベンで開催されたIEEE主催のICASSP2015に参加した．ICASSPは音声情報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，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

2015年4月19日から24日にかけオーストラリア・ブリスベンで開催されたIEEE主催のICASSP2015に参加した．ICASSPは音響，音声および信号処理分野でトップレベルの会議である．通常論文の投稿数は2322件あり，採択数は1207件（受理率52%）であった．本稿では音声情報処理に関する分野に注目し，ICASSP2015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. 音声強調（音源分離，雑音除去，残響除去）

本節では，1ch（モノラル）音声強調およびマルチチャネル音声強調の研究動向について概説する．なお，音声認識の前処理を前提とした音声強調方法に関する研究動向については次節にて紹介する．

1ch信号を対象とした音声強調，特に音源分離アプローチとしては，非負値行列因子分解（Non-negative Matrix Factorization; NMF）に基づく方法が依然として関心を集めている．特に最近は，クリーンな音声信号から基底スペクトルを事前学習しておき，そのサブセットを基に観測スペクトログラムに対しNMFを適用する半教師ありアプローチが多く提案されている（例えば[1]）．また，数年前のOzerovらによるNMFの多チャンネル拡張の登場を皮切りに，最近は多チャンネルNMFの新しいバージョンの提案や改良の検討も増えてきている（[2], [3], [4]等）．また，NMFに基づく音源分離のアプローチは，振幅スペクトルないしパワースペクトルが加法的であるという必ずしも正しくない仮定に基づいている．この問題点に鑑み，加法性が成り立つ複素スペクトログラムの領域でNMF-likeな信号分解を実現するアプローチが亀岡らによって以前提案されているが，このアプローチと同様の方向性で各構成音の位相を考慮した新たな改良アプローチ([3], [5]等）が検討されている点も注目に値する．一方，これらブラインド音声強調処理とは一線を画す枠組みとして，ユーザが音声強調・音源分離・多重音解析などに有用な何かしらのヒントをシステムに与えることで各種処理を補助する仕組が提案されており，これらは総称してInformed source separationと呼ばれる．この枠組は数年前から流行しており，今年のICASSPでもNMF等に基づく方法等，いくつかの発表があった（[6]等）．

また，1chブラインド音声強調の分野での比較的新しい潮流として，深層学習を用いた方法，Denoising autoencoder（DA）の拡張に関する研究発表が多く見られた．例えば，その一端を紹介すると，[7]では，DAを用いてクリーン音声の時間領域信号を直接回復する方法が提案されていた．通常のDAでは，観測の特徴量からクリーンな特徴量もしくは音声強調用マスク（Ideal Binary Mask等）を推定するが，[7]では音声強調用マスクを推定し，観測信号に乗算，強調結果を観測信号の位相と掛け合わせ，逆フーリエ変換を行い時間領域信号を得る，という部分までがネットワーク構造の中に組み入れられ，それらの処理ステップをも逆誤差伝搬法で最適化する方法が提案されていた．また，[8]では，Bidirectional Long-Short-Term Memory (BLSTM)

1 情報通信研究機構
2 早稲田大学
3 同志社大学
4 東京大学
5 日本電信電話株式会社
6 東京工業大学
7 国立情報学研究所
8 神戸大学
9 三菱電機株式会社
10 名古屋工業大学
11 京都大学
12 Mitsubishi Electric Research Laboratories

を用いた，観測信号に基づくマスク推定の技術について，いくつかの拡張方法が検討されていた．具体的には，(a) 推定されたマスクの値が位相情報を考慮した場合に最適となるようマスク設計方法に新たな基準を与えたこと，(b) BLSTM の入力に観測信号だけではなく，補助情報として観測信号を音声認識して得られた結果である HMM 状態事後確率や，それに関連する情報を入力していたこと，が新しく，いずれの工夫も性能改善に寄与していた．

また，DNN などに代表されるいわゆる Black box なアプローチと NMF などに代表されるモデルベースのアプローチの欠点を互いの利点で補い合うことを目的とした 1ch 音源分離手法として Deep NMF[9] という方法が提案されていた．この方法では，NMF の乗法更新アルゴリズムの各更新ステップは unfolding（展開）され DNN のような深層ネットワークの形に成形され，そのネットワークは誤差逆伝搬法を用いて最適化される．現段階では教師あり音源分離タスクにおいて DNN の性能には及んでいないものの，このような識別モデルと生成モデルのハイブリッドアプローチの動向は今後の注目に値する．

マルチチャネル音声強調の研究では，空間情報を積極的に利用し強調を行うビームフォーマ (GSC，MVDR ビームフォーマ，マルチチャネル Wiener Filter 等) を用いた技術の研究が以前に引き続き推し進められていた．特に，残響除去の研究は盛り上がりを見せており，残響除去技術単体の研究 ([4]) だけでなく，残響除去と雑音除去やエコーキャンセラを統合的な枠組みで扱う研究が散見された．また，マルチチャネル音声強調のためのビームフォーマ設計のために必要となる，目的音の到来方向 (Direction-of-arrival:DOA) 推定や残響時間推定などの研究も多く見られた．（木下，亀岡）

3. 耐残響・雑音

耐残響・雑音音声認識分野では深層学習，すなわち DNN の強みを活かし，通常の音響特徴に周囲の状況・環境に関する情報（マルチチャネル入力を含む）を補助特徴量として付加する研究が数多く見られた．また，音声強調や音声分離など，これまで信号処理的な手法が主流であったフロントエンド処理においても深層学習に基づく方法が多数見られるようになった．

まず，補助特徴量に関する文献を紹介する．文献 [10] は，DNN 音響モデルの入力およびターゲットに複数の情報を用いた残響下音声認識手法を提案している．入力としては，通常の音響特徴量以外に各部屋のインパルス応答の推定値を利用する．また，音素状態とともにクリーン音声特徴量をターゲットとして用いた音素識別と音声強調のマルチタスク学習を行っている．評価実験ではいずれの手法も認識精度の向上に寄与することが示されているが，認識時に補助特徴量を与える必要がない点でマルチタスク学習の方が有利である．雑音下音声認識では，音声強調による歪みを考慮するため，不確定性（uncertainty）に基づくデコーディング技術がよく使われる．特徴量から推定した不確定性は過小評価されることが多く，リスケーリングによる補正が有効である．文献 [11] では，不確定性を HMM 状態毎に推定することに加えて，相互情報量最大化基準に基づく識別学習を行うことで，性能を向上させた．しかしこの技術は GMM においては理論的な補正が可能であるが，DNN ではそのような補正方法が知られておらず，単純に入力特徴量に推定された雑音を補助特徴量として加えることが多い．文献 [12] では，音源の推定到来方向のばらつきを推定し，不確定性を表す特徴量として利用している．このような特徴量を用いて入力特徴量の重みづけを行うことで性能を向上させた．

次に，フロントエンド処理を紹介する．深層学習に基づく耐残響・雑音音声認識のフロントエンドとして，Denoising autoencoder（DA）が広く利用されつつある．しかし DA は，観測特徴量からクリーン音声特徴量への変換を直接学習するため，学習データに含まれない雑音環境では性能が劣化する．そこで文献 [13] では，Bayesian Feature Enhancement（BFE）を用いて観測特徴量よりクリーン音声特徴量を抽出し，得られたクリーン音声特徴量をターゲットとして DA を未知の雑音環境に適応させる方法を提案した．実験の結果，BFE 単体の場合に比べて残響下音声認識の性能が改善することを示した．信号処理的アプローチによるフロントエンド処理としては，Non-negative Matrix Factorization（NMF）を用いた手法が数多く発表された．文献 [14] では，クリーン音声と雑音の基底のペアである coupled dictionary を用いた NMF 音声強調について述べられている．coupled dictionary は事前にメルフィルタバンク（Mel），振幅（パワー）スペクトル（DFT），変調スペクトル（MS）領域のそれぞれで学習し，NMF による分解時にも，それぞれの領域でアクティベーション系列を推定する．音声の再構成時には各領域での音声基底と音声アクティベーション系列の組み合わせを評価し，MS 音声基底と DFT 音声アクティベーション系列の組み合わせが最良であることを実験的に示した．

音声特徴量と口唇特徴量を統合し，頑健性を確立するマルチモーダル音声認識においても深層学習が利用されている．文献 [15] では，音声特徴と口唇特徴に対して各々独立に DNN を構成し，最終層で特徴統合を行っている．また，単純にノード結合を行うのではなく，二つの特徴の相関を考慮した「bilinear term」を導入している．（藤本，三村，太刀岡，滝口）

4. モデル適応

音響モデルの話者（ないし環境）適応化については，1 件のオーラルセッションに加えて，音響モデルのポスター

セッションにも話者適応化に関する発表が含まれていた．多くの研究は，今日標準的な音響モデルである DNN-HMM を話者適応，ないし適応学習をすることを目的としていた．以下，DNN-HMM 音響モデルを対象とした話者適応化に関連する発表を，話者毎に推定するパラメータの違いに着目して 1) 音響モデル適応型，2) 補間係数適応型，3) 補助コード適応型，4) 活性化関数適応型，の 4 つに分類して述べる．

1) 音響モデル適応アプローチでは，重み行列等の DNN パラメータそのものが話者毎に推定される．DNN を構成するパラメータ数は膨大なため，過学習を防ぐための正則化が重要である．[16] では，パラメータ空間における係留点から大きく逸脱しないための正則化項を用いている．この方法では，線形変換層を用いて話者適応学習（すなわち，音響モデル学習時に正準モデルを最適化すること）を行うことで，正則化項の係留点自体が自動的に最適化されることを可能にしている．この種のアプローチの発想は，GMM-HMM 音響モデルにおける MLLR や MAP 適応に近いといえる．

2) 補間係数適応アプローチでは，音響モデルを複数のニューラルネットの線形和として分解表現し，適応時には，各基底ネットワークに乗算される補完係数のみを推定する．これによって，非常に少ないパラメータで高速な適応を実現している．各基底ネットワークは，音響モデル学習時に適応学習される．今回の ICASSP では，この考え方に基づくアプローチが複数の研究機関から提案された [17], [18], [19]．これらの手法は GMM-HMM 音響モデルにおける cluster adaptive training (CAT) や gating network に類似した発想に基づいている．

3) 補助コード適応アプローチでは，DNN の入力層や隠れ層に対して，話者ないし環境情報を表現する補助コードを入力として与えることで，当該層のバイアスを話者・環境毎に間接的に推定する．話者適応化を対象とした従来研究では，話者認識で用いられる i-vector が補助コードとして使用されていた．これに対して [20], [21] では，話者認識を行う DNN のボトルネック特徴量に基づいた補助コードの推定方法を提案するなど，話者適応化により効果的な補助コードを模索している．補助コードは多くの場合 DNN の学習に依存せず外部から与えられているの対して，speaker code 法と呼ばれる [22] の一連の先行研究では，補助コード自体を DNN とともに学習する．今回発表された [22] では，特異値分解を用いて識別的な補助コード及びそれに結合する重み行列を効果的に初期化する方法が提案された．なお，2) 補完係数適応アプローチと 3) 補助コード適応アプローチは異なるコンセプトの基で提案されたものであるが，3) は 2) において，バイアスのみを補間するのと等価であることが容易に分かる．

4) 活性化関数適応アプローチでは，ニューラルネットを構成する各ユニットの非線形活性化関数に係るパラメータを話者毎に推定する [23], [24]．[23] では，活性化関数としてシグモイド関数を仮定し，そのゲインとバイアスを話者毎に最適化している．学習時には標準シグモイド関数を用いて DNN を学習し，適応時にのみ各ユニットのゲインとバイアスを話者毎に推定する．この方法は活性化関数の直前に対角な線形変換層を挿入しているのと等価であるため，実際にはシグモイド関数以外の活性化関数にも用いることができる．[24] では，プーリング構造をもつネットワークを仮定して，プーリング関数のパラメータを話者毎に推定する方法が検討された．プーリング処理では通常，いくつかのユニット出力から最大値や平均値が求められるが，この研究ではプーリング関数をガウシアンカーネルで記述し，そのパラメータを話者毎に変えることで話者適応化を実現している．活性化関数適応アプローチでは，適応時に推定するパラメータ数は音響モデルのユニット数のオーダーであるため，適応データが比較的少量である場合にも効果をもつ．

以上みてきた方法は，DNN-HMM 音響モデルの話者適応化を目的としていたが，[25] では GMM-HMM 音響モデルに対して，話者適応のための非線形変換を DNN を用いて推定する方法が提案された．この方法では，CMLLR を用いて話者変換 DNN を初期化した後，最尤推定によってネットワークのパラメータを学習することで，CMLLR よりも高精度な話者適応を実現している． （吉岡，落合）

5. 高性能音声認識・その他

ニューラルネットワーク音響モデル自体の性能改善として活性化関数に関する研究が提案されている．Dropuout や DropConnect ではデータ数の増大による計算量の増加が問題となる．これを回避するため，ニューラルネットワークの活性化関数の入力，もしくは出力にガウスノイズを加える手法が提案された [26]．入力側のガウスは Dropout の近似となり，出力側のガウスは DropConnect と似た働きを行うことができる．また，RNN の LSTM ユニットと Maxout 関数を組み合わせる研究も提案された [27]．こちらは，LSTM ユニットの入力部における活性化関数として Maxout 関数を導入するという単純なアプローチながらも，従来の LSTM と比較して高い性能を得ることを示している．

聴覚や信号処理の分野において過去に研究された知見と，ニューラルネットワーク内に（大規模データを用いた学習によって自動的に）獲得された処理との関連性について議論する研究がいくつか発表されている．特に，畳み込み層を持つニューラルネットワークに音声波形を直接入力し，HMM の状態出力確率を計算する手法 [28], [29] において，ガンマトーンフィルタやバンドパスフィルタに類似した特徴量が学習によって自動的に獲得されている事が示

された．更に，複数マイクで観測されたマルチチャネルの音声波形を入力する事により，遅延和アレーによる雑音抑圧処理が獲得されている事が示された．

また，識別的アプローチであるディープネットワークと生成的アプローチである混合ガウス分布モデルを統合する試みとして，通常のニューラルネットワークで用いられる各層の役割を混合ガウス分布にもどつく入出力変換に置き換える手法が提案された [30], [31]．文献 [30] は，全ガウス分布の共分散行列を単一の共分散行列で共有するという条件の下，混合ガウス分布モデルと等価な対数線形モデルがパラメータの変換により解析的に求まる性質を利用して，混合ガウス分布モデルのパラメータを従来のディープネットワークに組み込む手法を提案した．文献 [31] では，直接混合ガウス分布モデルをひとつの層として表現し，それらを積み重ねた深層ネットワークを実現した．特に後者の手法は，理論的には識別器準のみならず尤度基準での学習も可能であり，生成モデルの利点である教師無し学習や過学習の緩和効果等もディープネットワークに組み入れることも将来的には可能である．

学習法方に注目した研究としては低資源言語を対象とした音声ドキュメント検索のセッションが立てられた他，数は多くないものの一般の音声認識を対象とした発表も行なわれた．それらの中でマルチタスク学習に関する発表としては，ニューラルネットワークの学習時のターゲットとしてコンテキスト依存音素の HMM 状態に加えてコンテキスト非依存音素の HMM 状態を用いる方法の提案が挙げられる．これは同じ中心音素を持つ音素状態を関連付けながら学習することを目的としたものであり，単語誤り率の削減に有効であることが示されている [32]．また，音声特徴量の教師なし学習法としてディープニューラルネットを用いて拡張した正準相関分析を応用した手法の提案があり，話者非依存音素認識において効果のあることが報告されている [33]．（柏木，松田，渡部，篠崎）

6. 話者および言語認識

話者認識に関する発表は，現在の主流である i-vector/PLDA システムをベースとして発話長やチャネルのミスマッチに対して頑健な方式の提案や，i-vector/PLDA システムに代わる新たな方式の提案，話者ダイアリゼーションに関する報告が中心であった．また，DNN を話者認識に利用する試みについても少ないながら報告があった．短い発話に対して頑健な方式提案として，i-vector の算出に必要な Baum-welch 統計量の計算に確率的な発話区間検出の情報を利用する手法が提案されている [34]．従来の手法では発話区間以外のフレームは Total variability 行列の学習や i-vector の推定に使用しなかったのに対し，提案手法では各フレームを発話／非発話の事後確率で重み付けして用いることで，リソースが限定されている条件下でも頑健な統計量の算出を実現した．NIST SRE2012 および DARPA RATS タスクにおいて，チャネルミスマッチ（microphone/telephone）条件下において，従来の発話区間検出を用いた手法よりも高い性能が得られることを示した．従来の i-vector/PLDA 型話者照合システムに関する興味深い解析として，照合器としての性能を定量的に評価する試みがある [35]．この研究では，各話者の i-vector 空間をターゲット話者と詐称話者についてそれぞれガウス分布でモデル化し，これらモデル間の Kullback-Leibler divergence を評価することで，システムが 5 秒および 40 秒以上の照合音声に対しそれぞれ 127 bit，182 bit のパスワード相当の強度を持つことを明らかにしている．また，i-vector/PLDA システムの高精度化として，音響特徴量に焦点を当てた研究も報告されている．話者照合の音響特徴量としては，MFCC に Δ や $\Delta\Delta$ パラメータを付与することで時間コンテキストを扱うのが一般的である．それに対し，時間・周波数スペクトルに対して直接 2 次元 DCT を施すことで時間コンテキストを扱う試みがある [36]．ベクトル化した 2 次元 DCT 係数行列に対して主成分分析を施し音響特徴量として用いることで，MFCC と Δ パラメータに基づく従来の方式の性能を平均 20%改善できることを報告している．ここでは，UBM/i-vector フレームワークを用いて評価が行われているが，DNN/i-vector フレームワークにおいても同様の改善が見られる点についても言及されている．話者ダイアリゼーションに関する報告としては，システムを構成する主要ステージの一つであるリセグメンテーションの改善により，現状の最高性能を達成したという報告がなされている [37]．ここでは，因子分析に基づく部分空間上で変分ベイズモデルを用いてダイアリゼーションを行う方式 (いわゆる VB diarization) に対して，HMM を導入して話者遷移を制限することの有効性を明らかにしている．

言語識別に関する発表はオーラル 2 件があった．1 つ目は，Deep bottleneck network(DBN) を用いた方式であり，短い発話や方言において精度の改善を目的とした発表である [38]．従来法の入力層を音響特徴量，出力層を音素ラベルとした DNN の BN 特徴量を i-vector で表現する方法に対して，今回の方法は BN だけでなく出力層も用いている．出力層も音素ラベルではなく，Interspeech2014 で提案されている Senones を採用している．NIST LRE2011 のアラビア語方言および NIST LRE2009 の 23 言語を用いた実験において，従来法よりも精度の改善が確認されている．2 つ目は，i-vector 型ノイズロバスト言語識別のポスト処理における発表である．従来法では，linear discriminant analysis(LDA) を用いていたが，この方法は i-vector をガウス分布として仮定している，また，少ない言語数の場合効果が低いという問題があった．そこで，LDA の一般型である nearest neighbor discriminant analysis(NDA) を

導入し，DARPA RATS プログラムのノイズを含む音声を用いた実験により，従来法よりも精度のよい結果を得ている [39]．これらは共に i-vector を用いた方法であるが，ICASSP2014 から google によって提案された DNN 型言語識別器との比較が今後注目する所である． （俵，小川，岡本）

7. 音声合成・声質変換

音声合成に関するセッションはオーラルが 2 つ，ポスターが 2 つで構成され，オーラルセッションのうち 1 つがニューラルネットワークに関するものであった．セッション全体を通しても 36 件中 14 件がニューラルネットワーク/深層学習をキーワードに含んだものであり，音声合成分野においても高い注目が集まっていることが分かる．以下では著者らの注目する発表をいくつか紹介する．

声質変換においては，他分野と同様に時間依存性をモデル化した LSTM や RNN の使用を試みたものが複数発表された [40], [41]．また文献 [42] では，DNN に基づく声質変換において複数話者データの使用に関して検討されている．入力話者を複数用いた学習によって入力層を話者変動に対して汎化させることで，入力話者に依存しない声質変換の実現を目指している．また出力についても複数話者によるマルチタスク学習の枠組みの導入を検討している．

文献 [43] では，統計的パラメトリック音声合成において，Time-Frequency Trajectory Excitetion (TFTE) のモデリングを改善している．ピッチ依存 TFTE では特徴量の次元数が変化するため，統計的パラメトリック音声合成で扱うのが困難である．本文献では Predicted Average Block Coefficients (PABC) を用いることで，問題の解決を行っている．文献 [44] では，DNN 音声合成において波形の直接モデル化を提案している．従来，統計的パラメトリック音声合成では波形からの音響特徴量抽出と HMM や DNN を用いた音響モデリングを行う．本手法では音声信号 $\mathbf{x}$ を自己回帰型のガウス過程と仮定し，ケプストラム $\mathbf{c}$ を用いた確率密度関数 $P(\mathbf{x}|\mathbf{c})$ を定義し，特別な DNN の出力層を用いることで音響特徴量抽出，音響モデリングの統合を行っている．文献 [45] では，HMM 音声合成における音質劣化の原因について調査している．HMM 音声合成におけるモデリングの合成音声への影響は様々考えられるが，本文献では，時間領域におけるスムージング，スペクトルパラメータの分散，パラメータの平均化による影響に着目している．分析再合成を用いることでスムージング，分散の調整を行った音声サンプル，及び，HMM 音声合成と HMM 音声合成を模した音声サンプルを作成し，聴取実験を行っている．

文献 [46] ではニューラルネットワークによる単語の連続値ベクトル表現を，RNN に基づく音声合成の入力変数として使用する方法を提案した．語彙数 82000 程度の Broadcast news コーパスを用いた言語モデルでは，形態素や ToBI のラベルのない音声データに対し，自然性が向上することを示した．ニューラルネットワークに基づく音声合成は様々な手法が提案されているが，それらを話者適応に応用する手法は十分に検討されていなかった．文献 [47] では，DNN に基づく音声合成において出力層の重みパラメータのみを話者毎に学習し，隠れ層のパラメータを複数話者で共有する Multi-speaker DNN を提案した．Multi-speaker DNN の合成音声は話者依存の DNN に比べ自然性のスコアが高く，また，この枠組みにおいて隠れ層の共有パラメータを固定することで，DNN 音声合成に基づく話者適応手法が実現できることを示した．文献 [48] では，音声分析合成系における新たな合成器を提案した．STRAIGHT では FFT により計算コストが高くなる，また有声摩擦音の品質が悪いという問題があった．提案法では，非定常な変調正弦波の組合せで音声をモデル化している．携帯端末においても実時間合成の可能な混合励振源に基づく合成器と同程度の計算量で，STRAIGHT より自然性の高い音声を合成できることを示した．

文献 [49] では，DNN 音声合成において，異なる 2 つの音響特徴量を用いてマルチタスク学習の枠組みで DNN の学習を行っている．また，DNN にボトルネック層を導入し，ボトルネック特徴量を入力として新たに DNN の学習を行っている．実験結果からマルチタスク学習の効果は大きくなかったが，ボトルネック特徴量を入力することで合成音声の自然性を改善することを示した．文献 [50] では，Mixture Density Network と同様に確率分布のパラメータを出力とする Real-valued Neural Autoregressive Density Estimator (RNADE) の学習アルゴリズムの改善が行われている．RNADE の学習アルゴリズムに，発話単位での音響特徴量系列の平滑化処理を組み込むことによって，音響特徴量の次元間の関係と時間方向の関係を考慮したモデル化が行われる．文献 [51] では， Long short-term memory recurrent neural network (LSTM-RNN) の出力層にリカレント構造を持たせることによって，低遅延なストリーミング音声合成を実現する手法を提案している．従来の DNN 音声合成においては，発話などの長い単位で平滑化処理を行うことで，滑らかな音響特徴量系列を出力していたが，出力層にリカレント構造を持つ LSTM-RNN を用いることで，発話単位の平滑化処理なしに滑らかな音響特徴量系列を出力可能にした． （齋藤，高木，郡山，橋本）

8. 著者へのアンケート

本 ICASSP 報告の著者を対象に，採択された論文の執筆方法や査読担当者による評価方法その他についてアンケートを行なった．結果を表 1 に示す．アンケートの集計では，論文の執筆や発表に関する設問は複数の共著者を重複して数えないようにしている．また査読担当者については

表 1 本 ICASSP 報告の著者へのアンケート．Type 1 の質問の回答は A:0-1，B:2-3，C:4-5，D:6-7，E:8 以上，のいずれか 1 つを選択．Type 2 の質問の回答は A:しない/いいえ，B:あまりしない/あまりそうでない，C:どちらとも言えない，D:ややする/ややそうである，E:する/はい，のいずれか 1 つを選択．いずれも各人数を集計．

. **Table 1** Questionnaire to the authors of this ICASSP report.

質問項目 (対象者：質問内容 (質問タイプ))	A	B	C	D	E
論文著者：初稿完成後の推敲回数 (Type1)	0	3	7	3	1
口頭発表者：口頭発表時のリハ回数 (Type1)	1	1	1	0	1
ポスター発表者：ポスター発表時のリハ回数 (Type1)	4	4	1	1	0
査読担当者：関連研究のサーベイの充実度は採否に影響するか (Type2)	0	0	2	3	5
査読担当者：趣旨やストーリーの明確さは採否に影響するか (Type2)	0	0	1	1	8
査読担当者：複数の評価タスクを用いるなど，実験内容の充実度は採否に影響するか (Type2)	0	0	0	7	3
査読担当者：英文のクオリティは実際問題として採否に影響するか (Type2)	0	2	0	5	3
査読担当者：個々の論文の評点時に，学会毎の採択率の違いを意識するか (Type2)	1	2	5	1	1
会議参加者：セッションチェアやチュートリアル等で日本の貢献は十分か (Type2)	0	9	8	1	0
海外経験者：海外と比べて日本の研究体制に不利な点はあるか (Type2)	0	0	4	0	4
海外経験者：海外と比べて日本の研究体制に有利な点はあるか (Type2)	0	0	4	1	1

2015 年の ICASSP に限定している．

この結果より，口頭またはポスター発表時の練習回数にはばらつきはあるものの，論文執筆時にはすべての著者が推敲を複数回行なっていることが分かる．また査読担当者の回答から，論文執筆においてはストーリーを明確にすることが特に重要と言える．さらにサーベイを充実させることや英文のクオリティを高めること等も，落とされない論文を書くために重要と考えられる．海外経験のある研究者への質問では，海外に比べて日本の研究体制に不利な点があるとする意見と，どちらとも言えないとする意見が同数であった．日本の研究体制で不利な点としては，研究機関を越えた人材交流が少ないことを複数の回答者が挙げている．人材交流を活性化すると，現状の待遇では日本から人材が流出してしまう心配があるとする意見もあった．また海外では大規模な競争的プロジェクトが行なわれ予算規模や人材の点で日本を大きく上回っているとする意見がある一方で，日本も音声音響信号処理の研究者の質や量は高いレベルにありその点はキープすべきとの回答もあった．音声合成は特に日本の貢献が大きい分野と言える．その他，研究テーマの選択に対する自由度は日本の方が比較的高いとする回答もあった．会議における情報収集のコツについての全員を対象とした質問では，多くの人と情報交換することが重要であるとの見解がほぼ一致した回答であった．

参考文献

[1] Barker, T., Virtanen, T., and Pontoppidan, N. H.: Low-latency sound-source-separation using non-negative matrix factorisation with coupled analysis and synthesis dictionaries, *Proc. ICASSP*, pp. 241–245 (2015).
[2] Kitamura, D., Ono, N., Sawada, H., Kameoka, H. and Saruwatari, H.: Efficient multichannel nonnegative matrix factorization exploiting rank-1 spatial model, *Proc. ICASSP*, pp. 276–280 (2015).
[3] Deleforge, A. and Kellermann, W.: Phase-optimized k-SVD for signal extraction from underdetermined multichannel sparse mixtures, *Proc. ICASSP*, pp. 355–359 (2015).
[4] Jukic, A., Mohammadiha, N., van Waterschoot, T., Gerkmann, T. and Doclo, S.: Multi-channel linear prediction-based speech dereverberation with low-rank power spectrogram approximation, *Proc. ICASSP*, pp. 96–100 (2015).
[5] Kameoka, H.: Multi-resolution signal decomposition with time-domain spectrogram factorization, *Proc. ICASSP*, pp. 86–90 (2015).
[6] de Andrade Scatolini, C., Richard, G. and Fuentes, B.: Multipitch estimation using a PLCA-based model: Impact of partial user annotation, *Proc. ICASSP*, pp. 186–190 (2015).
[7] Wang, Y. and Wang, D.: A deep neural network for time-domain signal reconstruction, *Proc. ICASSP*, pp. 4390–4394 (2015).
[8] Erdogan, H., Hershey, J. R., Watanabe, S. and Roux, J. L.: Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks, *Proc. ICASSP*, pp. 708–712 (2015).
[9] Roux, J. L. and Hershey, J. R.: Deep NMF for speech separation, *Proc. ICASSP*, pp. 66–70 (2015).
[10] Giri, R., Seltzer, M. L., Droppo, J. and Yu, D.: Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning, *Proc. ICASSP*, pp. 5014 – 5018 (2015).
[11] Tran, D. T., Vincent, E. and Jouvet, D.: Discriminative uncertainty estimation for noise robust ASR, *Proc. ICASSP*, pp. 5038 – 5042 (2015).
[12] Schwarz, A., Huemmer, C., Maas, R. and Kellermann, W.: Spatial diffuseness features for DNN-based speech recognition in noisy and reverberant environments, *Proc. ICASSP*, pp. 4380 – 4384 (2015).
[13] Heymann, J., Hab-Umbach, R., Golik, P. and Schlüter, R.: Unsupervised adaptation of a denoising autoencoder by Bayesian feature enhancement for reverberant ASR under mismatch conditions, *Proc. ICASSP*, pp. 5053 – 5057 (2015).
[14] Baby, D., Gemmeke, J. F., Virtanen, T. and Hamme, H. V.: Exemplar-based speech enhancement for deep neural network based automatic speech recognition,

Proc. ICASSP, pp. 4485 – 4489 (2015).
[15] Mroueh, Y., Marcheret, E. and Goel, V.: Deep multimodal learning for audio-visual speech recognition, *Proc. ICASSP*, pp. 2130 – 2134 (2015).
[16] Ochiai, T., Matsuda, S., Watanabe, H., Lu, X., Hori, C. and Katagiri, S.: Speaker adaptive training for deep neural networks embedding linear transformation networks, *Proc. ICASSP.*, pp. 4605–4609 (2015).
[17] Tan, T., Qian, Y., Yin, M., Zhuang, Y. and Yu, K.: Cluster adaptive training for deep neural network, *Proc. ICASSP.*, pp. 4325–4329 (2015).
[18] Delcroix, M., Kinoshita, K., Hori, T. and Nakatani, T.: Context adaptive deep neural networks for fast acoustic model adaptation, *Proc. ICASSP.*, pp. 4535–4539 (2015).
[19] Wu, C. and Gales, M.: Multi-basis adaptive neural network for rapid adaptation in speech recognition, *Proc. ICASSP.*, pp. 4315–4319 (2015).
[20] Liu, Y., Karanasou, P. and Hain, T.: An investigation into speaker informed DNN front-end for LVCSR, *Proc. ICASSP.*, pp. 4300–4304 (2015).
[21] Huang, H. and Sim, K. C.: An investigation of augmenting speaker representations to improve speaker normalisation for DNN-based speech recognition, *Proc. ICASSP.*, pp. 4610–4613 (2015).
[22] Xue, S., Jiang, H., Dai, L. and Liu, Q.: Unsupervised speaker adaptation of deep neural network based on the combination of speaker codes and singular value decomposition for speech recognition, *Proc. ICASSP.*, pp. 4555–4559 (2015).
[23] Zhao, Y., Li, J., Xue, J. and Gong, Y.: Investigating online low-footprint speaker adaptation using generalized linear regression and click-through data, *Proc. ICASSP.*, pp. 4310–4314 (2015).
[24] Swietojanski, P. and Renals, S.: Differentiable pooling for unsupervised speaker adaptation, *Proc. ICASSP.*, pp. 4305–4309 (2015).
[25] Cui, X. and Goel, V.: Maximum likelihood nonlinear transformations based on deep neural networks, *Proc. ICASSP.*, pp. 4320–4324 (2015).
[26] Zhang, H., Miao, Y. and Metze, F.: Regularizing dnn acoustic models with gaussian stochastic neurons, *Proc. ICASSP*, pp. 4964 – 4968 (2015).
[27] Li, X. and Wu, X.: Improving ling short-term memory networks using maxout units for large vocabulary speech recognition, *Proc. ICASSP*, pp. 4600 – 4604 (2015).
[28] Hoshen, Y., Weiss, R. J. and Wilson, K. W.: Speech acoustic modeling from raw multichannel waveforms, *Proc. ICASSP*, pp. 4624 – 4628 (2015).
[29] Palaz, D., Magimai-Doss, M. and Collobert, R.: Convolutional neural networks-based continuous speech recognition using raw speech signal, *Proc. ICASSP*, pp. 4295 – 4299 (2015).
[30] Tüske, Z., Tahir, M. A., Schlüter, R. and Ney, H.: Integrating Gaussian mixtures into deep neural networks: softmax layer with hidden variables, *Proc. ICASSP*, pp. 4285 – 4289 (2015).
[31] Variani, E., McDermott, E. and Heigold, G.: A Gaussian mixture model layer jointly optimized with discriminative features within a deep neural network architecture, *Proc. ICASSP*, pp. 4270 – 4274 (2015).
[32] Bell, P. and Renals, S.: Regularization of context-dependent deep neural networks with context-independent multi-task training, *Proc. ICASSP* (2015).
[33] Wang, W., Arora, R., Livescu, K. and Bilmes, J.: Unsupervised learning of acoustic features via deep canonical correlation analysis, *Proc. ICASSP* (2015).
[34] McLaren, M. and Lei, Y.: Improved speaker recognition using DCT coefficients as features, *Proc. ICASSP*, pp. 4430 – 4434 (2015).
[35] Nautsch, A., Rathgeb, C., Saeidi, R. and Busch, C.: Entropy analysis of i-vector feature spaces in duration-sensitive speaker recognition, *Proc. ICASSP*, pp. 4674 – 4678 (2015).
[36] McLaren, M., Graciarena, M. and Lei, Y.: SOFTSAD: Integrated frame-based speech confidence for speaker recognition, *Proc. ICASSP*, pp. 4694 – 4698 (2015).
[37] Sell, G. and Garcia-Romero, D.: Diarization resegmentation in the factor analysis subspace, *Proc. ICASSP*, pp. 4794 – 4798 (2015).
[38] Song, Y., Cui, R., Hong, X., Mcloughlin, I., Shi, J. and Dai, L.: Improved language identification using deep bottleneck network, *Proc. ICASSP*, pp. 4200 – 4204 (2015).
[39] Sadjadi, S. O., Pelecanos, J. W. and Ganapathy, S.: Nearest neighbor discriminant analysis for language recognition, *Proc. ICASSP*, pp. 4205 – 4209 (2015).
[40] Sun, L., Kan, S., Li, K. and Meng, H.: Voice conversion using deep bidirectional long short-term memory based recurrent neural networks, *Proc. ICASSP*, pp. 4869–4873 (2015).
[41] Liu, P., Yu, Q. and Wu, Z.: A deep recurrent approach for acoustic-to-articulatory inversion, *Proc. ICASSP*, pp. 4450–4454 (2015).
[42] Liu, L. J., Ling, Z. H. and Dai, L. R.: Spectral conversion using deep neural networks trained with multi-source speakers, *Proc. ICASSP*, pp. 4849–4853 (2015).
[43] Song, E., Joo, Y. S. and Kang, H. G.: Improved time-frequency trajectory excitation modeling for a statistical parametric speech synthesis system, *Proc. ICASSP*, pp. 4950–4953 (2015).
[44] Tokuda, K. and Zen, H.: Directly modeling speech waveforms by neural networks for statistical parametric speech synthesis, *Proc. ICASSP*, pp. 4215–4219 (2015).
[45] Merritt, T., Latorre, J. and King, S.: Attributing modelling errors in HMM synthesis by stepping gradually from natural to modelled speech, *Proc. ICASSP*, pp. 4220–4224 (2015).
[46] Wang, P., Qian, Y., Soong, F. K. and He, L.: Word embedding for recurrent neural network based TTS synthesis, *Proc. ICASSP*, pp. 4879–4883 (2015).
[47] Fan, Y., Qian, Y. and Soong, F. K.: Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis, *Proc. ICASSP*, pp. 4475–4479 (2015).
[48] Agiomyrgiannakis, Y.: Vocaine the vocoder and applications in speech synthesis, *Proc. ICASSP*, pp. 4230–4234 (2015).
[49] Wu, Z., Valentini-Botinhao, C., Watts, O. and King, S.: Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis, *Proc. ICASSP*, pp. 4460–4464 (2015).
[50] Uria, B., Murray, I., Renals, S., Valentini-Botinhao, C. and Bridle, J.: Modelling acoustic feature dependencies with artificial neural networks: trajectory-RNADE, *Proc. ICASSP*, pp. 4465–4469 (2015).
[51] Zen, H. and Sak, H.: Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis, *Proc. ICASSP*, pp. 4470–4474 (2015).