Vol.2017-SLP-117 No.3
2017/7/27

国際会議ICASSP2017報告

浅見 太一[1]　大谷 大和[2]　岡本 拓磨[3]　小川 哲司[4]　落合 翼[5]　亀岡 弘和[1]　駒谷 和範[6]　高木 信二[7]
高道 慎之介[8]　俵 直弘[4]　南條 浩輝[9]　橋本 佳[10]　福田 隆[11]　増村 亮[1]　松田 繁樹[5]　李 晃伸[10]
渡部 晋治[12]

概要：2017 年 3 月 5 日から 9 日にかけ，合衆国・ニューオリンズで IEEE 主催の ICASSP が開催された．ICASSP は音声言語情報処理の分野におけるトップカンファレンス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，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

2017 年 3 月 5 日から 9 日にかけ，合衆国・ニューオリンズで IEEE 主催の ICASSP2017 (The 42nd IEEE International Conference on Acoustics, Speech and Signal Processing) が開催された．ICASSP は Interspeech と並んで音声言語情報処理分野におけるトップカンファレンスと位置づけられており，前者のほうが信号処理寄りで技術色の濃い会議となっている．通常論文の投稿数は 2,518 件あり，うち 1,220 件が受理された（受理率 48.5%）．本稿では音声情報処理に関する分野に注目し，関連度の高い以下の 3 トラックを中心として，ICASSP2017 における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[*1]．（李）

- Audio and Acoustic Signal Processing (AASP)
- Speech Processing (SP)
- Human Language Technology (HLT)

2. 音声分析・音響信号処理

音声分析は 2 つのポスターセッション (SP-P1: peech Production, Coding and Reconstruction, SP-P12: Speech Analysis) があり，音響信号処理は 2 つのオーラルセッションと 3 つのポスターセッション (AASP-L2: Non-Negative Audio Modeling, AASP-L6: Acoustic Array Processing I) および 3 つのポスターセッション (AASP-P4: Applications and Theory of Acoustic Processing, AASP-P6: Microphone Arrays, Array Signal Processing and Source Localization, Noise, AASP-P8: Echo, Feedback and Reverberation Reduction) により構成されていた．これらの分野も他と同様，深層学習に基づく手法が多く見られた．以下では，いくつかの発表について報告する．

2.1 音声分析

以下 2 つは，音声分析からの報告であり，両方ともに，位相スペクトルを考慮したモデル化が共通点である．

STRAIGHT 等のボコーダは声道スペクトルを高精細に分析，抽出するのに対して，声帯振動を高精細に分析，合成する方式として Glottal inverse filtering に基づく glottal vocoder がある．前者と比較して，後者は励振源の位相を忠実にモデル化する方式であり，STRAIGHT よりも高音質な音声合成を実現している．声道閉門区間に応じた時間重み窓を適用した線形予測により高精度な声帯振動を推定できるが，Airaksinen らは聴覚特性も考慮するために，さらに周波数ワーピングも加えた方式を提案した．DNN 音声合成聴取実験の結果，男性音では周波数ワーピングの効果が見られたが，女性音声では効果が見られず，今後の課題が残る [1].

深層学習ベースでの単一マイクロホン収録における残響除去はこれまでにもいくつか研究があるが，多くは振幅スペクトルのみを考慮した方式であり，位相の回復が課題であった．提案法では，人間の知覚メカニズムに基づき，時間-周波数の 2 次元複素スペクトルをさらにもう一回フー

[1] 日本電信電話株式会社
[2] 株式会社エーアイ
[3] 情報通信研究機構
[4] 早稲田大学
[5] 同志社大学
[6] 大阪大学
[7] 国立情報学研究所
[8] 東京大学
[9] 京都大学
[10] 名古屋工業大学
[11] 日本 IBM 株式会社
[12] Mitsubishi Electric Research Laboratories
[*1] 著者は 50 音順.

ⓒ 2017 Information Processing Society of Japan

リエ変換を施し，そのうち時間方向の変調である「rate 領域」を回復させるネットワークを構築する．これは，時間領域では畳み込みとなる残響成分が「rate 領域」では，近似的に掛け算となることに基づく．提案法により，既存手法よりも高精度な残響除去を実現している [2]．(岡本)

2.2 音響信号処理

AASP-L2: Non-Negative Audio Modeling セッションでは，例年同様，非負値行列因子分解 (Non-negative Matrix Factorization; NMF) の新応用や新拡張に関する研究が多く発表されていた．Task-Driven NMF (TD-NMF) は，NMF により推定される非負係数ベクトル（各基底のアクティベーション値を格納した非負ベクトル）を特徴量として各種識別タスクを解決することを目的とした手法で，NMF による基底学習（すなわち特徴量器の学習）と識別器の学習を多段的に行うのではなく，識別スコアが最適となるように識別器のパラメータと NMF の基底を同時学習する点が特徴である．これは，特徴抽出器と識別器や回帰分析器を一体の NN として扱う深層学習と通底する考え方といえる．本セッションでは，TD-NMF を音響イベント検出と話者識別へ適用する発表があった [3], [4].

上記セッションと AASP-P7: Source Separation & Denoising セッションでは，NMF を複素スペクトル領域や時間領域に拡張する新モデルがいくつか提案されていた．NMF の複素スペクトル領域や時間領域への拡張モデルは，位相スペクトルを潜在変数として扱うタイプ（Itakura-Saito (IS) NMF, Cauchy NMF，Lévy NMF など）とモデルパラメータ（確定変数）として扱うタイプ（複素 NMF，Time-domain Spectrogram Factorization (TSF)）に大別される．前者のタイプでは，α 安定分布に従う音源モデルを用いた多チャンネル音源分離の手法 [5]，後者のタイプでは一般化 KL ダイバージェンス規準を用いた複素 NMF[6]，TSF の高速アルゴリズム [7] が新たに提案されていた．また，関連する研究で，多チャンネル観測信号から各音源の振幅スペクトルと混合行列が既知の下でそれぞれの位相スペクトルを最適推定する方法が提案されていた [8]．この研究では，多チャンネル Wiener フィルタにおいて位相スペクトルを潜在変数と扱うことに対する問題点が提起され，より高い精度で信号分離を行うためには位相スペクトルを確定変数として直接推定することの重要性が主張されている．

AASP-L3: Deep Learning for Source Separation and Enhancement I セッションでは，基底の非負制約を外した NMF が ReLU を中間層の活性化関数とした Autoencoder (AE) と見なせる点に着目し，それを多層化した Non-negative Autoencoder (NAE) なる方法が提案されていた [9]．教師あり音源分離タスクにおいて有効性が示されている．（亀岡）

3. 音源分離・音声強調

音源分離・音声強調における技術的なトレンドとしては，深層ニューラルネットワーク (deep neural network; DNN) によるスペクトルの補正および時間周波数マスク（フィルタ係数）の推定に大別される．前者の主な狙いは，denoising auto-encoder (DAE) を用いて歪を含むスペクトルから歪を含まないクリーンなスペクトルを推定することである．後者では，歪を含むスペクトルから時間周波数マスク（もしくは目的信号の存在確率）を推定する．加えて，DNN による時間周波数マスキングと最小分散無歪応答 (minimum variance distortionless response; MVDR) ビームフォーマとの統合が盛んに検討されており，CHiME Challenge において優秀な成績を収めるなど，音声認識のための音声強調処理における近年の state-of-the-art を達成している．MVDR ビームフォーマのような線形処理では信号処理歪の発生が少なく，DNN 音響モデルとの親和性が高いことが主な理由と考えられる．

DAE を用いたスペクトルの補正としては，文献 [10], [11] などが挙げられる．文献 [10] では，DAE による音声強調によって生じるスペクトル平坦化の低減を目的として，mixture density netowrk によるクリーン音声の分布推定とモデルベース音声強調を統合する方式が提案されている．DAE によるスペクトルの歪の抑圧は，スペクトルからスペクトルの 1-to-1 のマッピングを前提としているが，この仮定は一般的に正しくない．実際，歪を含むスペクトルに対して，クリーンな信号のスペクトルと歪のスペクトルの組み合わせは複数あり得る．そこで，歪を含むスペクトルから歪を含まないスペクトルの分布を推定し，その後段で分布ベースの音声強調を行っている．同様に，DAE を用いた音声強調におけるスペクトル平滑化の低減のために，マルチストリーム音声強調方式が提案されている [11]．ここでは，特定の雑音の低減に特化した DAE を複数構築しておき，入力に適した DAE を選択的に用いることで，スペクトルの平坦化を抑圧しつつ高精度な音声強調を実現することが狙いである．しかし，システムの未知雑音に対する頑健性の担保が課題であろう．

DNN を用いた時間周波数マスクの推定およびフィルタ係数の推定としては，文献 [12], [13] などが挙げられる．文献 [12] では，時間周波数マスク推定の高精度化を目的として，stacking DNN を用いて過去数フレームのマスクを入力として現フレームのマスクを推定している．また，文献 [13] では，適応的ビームフォーマによる非定常信号の高精度な取り扱いを目的として，LSTM に基づくビームフォーマ推定，音声認識結果のビームフォーマ推定への利用，音声認識との結合学習を提案している．

MVDR ビームフォーマに基づく方式では，ビームフォーマの推定に必要となる音源の位置ベクトルの推定に DNN

ⓒ 2017 Information Processing Society of Japan

による時間周波数マスキングを活用している．文献 [14] では，DNN に基づくマスク推定と MVDR ビームフォーマの反復により，MVDR ビームフォーマに基づく音声強調の高精度化を実現している．（小川）

音声分離手法における問題点である permutation 問題（分離された音源が元のどの音源に対応するかに任意性がある）を扱う手法として，permutation 不変な学習方法が提案された [15]．本手法では，Deep network によって出力される複数 (S) の分離音源と元の複数の音源の組み合わせ（全部で $S!$ 通り）のなかから，最もコストの小さくなる組み合わせを探索し，その特定の組み合わせによるコストを用いて network を学習する手法である．WSJ コーパスから抽出した 2 音源を異なる SNR ごとに混ぜ合わせて混合音声を作成し，音言分離実験を行い，高い分離性能を実現した (SDR 10dB)．同様の効果をデンマーク語音声データで作成した混合音でも示している．（渡部）

4. 音声認識

4.1 フレームワーク

深層学習の発展により，音響モデル・発音辞書・言語モデルといった従来のモデル分類をまたがる研究が多く見られた．それらの代表例として，上記の複数のモデルによる処理を deep network によって一括的に扱う End-to-End 音声認識・音声処理に多くの興味深い研究が見られた（例えば SP-L2: End to End Speech Processing など）．これらのアプローチを実現する手法である Connectionist Temporal Classification (CTC) や Sequence to sequence の拡張もしくは応用研究が多く見られた．

文献 [16] は，著者らの先行研究である事後確率最大化基準に基づく CTC 学習の再定式化を，ベイズリスク最小化学習に応用した研究である．提案法は，確率統計的音声認識の主要基準である事後確率最大化にもとづいた場合，CTC 定式化には従来の言語モデル及び subword 系列の事後確率に加えて，subword 系列の事前確率及び単語-subword 遷移確率による補正因子が必要であることを理論的に指摘し，それらの効果を実験で示した研究である．先行研究ではこの効果をデコーディング時のみにおいて示していたが，[16] はそれをベイズリスク最小化学習に応用した場合にも，補正項の効果が重要であることを実験的に示した．条件によっては効果が小さい時もあるものの，全ての条件において従来の CTC に基づくベイズリスク最小化学習を上回っており，その理論的な指摘が妥当であることを十分に示した非常に価値のある研究といえる．（渡部）

Rao らは，マルチタスク学習の枠組みで階層的な LSTM grapheme CTC ネットワークを構築する方法を検討している [17]．この方法では層数の多い bidirectional LSTM(BLSTM) を用い，ネットワークのほぼ中間にあたる隠れ層（実験では第 5 層）を phoneme CTC ターゲットと接続して分類を行うタスクと，ネットワークの最終隠れ層と grapheme CTC ターゲットを接続して分類を行うマルチタスクによってネットワーク全体を統合的に学習している．さらに，中間の層と接続している phoneme CTC をアクセント（US, British, Australian, India English）に応じて並列化することによって，最終目標の grapheme CTC の精度がより高くなるような学習方法を提案している．著者らは，数千時間以上の大規模な学習コーパスを用いて，単純な構成の grapheme CTC が精度上 phoneme CTC に及ばないことを示したのと同時に，提案法であるアクセントに基づく複数の phoneme CTC を活用した方法が，grapheme CTC の性能を大幅に改善できることを実証した．（福田）

他方，knowledge distillation の枠組みを音声認識に利用する研究も急加速している．ICASSP でもこの傾向は顕著であり，複数の関連研究発表があった．これは教師ネットワークから生成されるソフトラベルをターゲットモデルの学習に用いる方法であり，教師ネットワークの能力・特性をターゲットモデルに反映させるという意味で，student-teacher training とも称される．典型的には，教師とターゲットモデル間の出力分布の KL ダイバージェンスを最小化するように学習を行う．この枠組みにおいて，Cui らは高い識別能力を持つ LSTM ネットワークを教師側として用い，種別が異なる単純な構造の DNN をターゲットとして学習することを試みた [18]．著者らはフレーム単位で事後確率の高い Top 50 状態のみをソフトラベルとして用い，音響的に識別の難しいフレームに対するソフトラベルがターゲットモデルの改善に大きく寄与することを見出した．また，LSTM や VGG などの様々なネットワークに由来するマルチリンガル特徴量で DNN モデル群を構築し，それらを教師モデルとして活用することも提案している．（福田）

4.2 耐雑音

Ravanelli らは，音声認識システムにおける「音声強調部」と「音声認識部」を協調させるための，モデル構造，並びに，その学習方法を提案している [19]．従来のシステム構造では，システムの処理の流れは，音声強調部から音声認識部への一方向であり，音声強調部に対して音声認識部の情報が利用されることはなかった．本研究では，音声強調部と音声認識部を，相互に階層的に積み上げたモデル構造を提案している．より具体的には，第一階層では，音声強調部と音声認識部は独立に処理を行う．一方で，上位階層では，音声強調部は音声認識部の出力を，音声認識部は音声強調部の出力を利用した上で処理を行う．こうした枠組みを多階層に積み上げることで，音声強調部と音声認識部のより協調した挙動を促している．評価実験を通して，一方向の情報のやり取りを行っていた従来手法と比較して，双方向の情報のやり取りを行う提案手法は，より高

ⓒ 2017 Information Processing Society of Japan

い認識性能を獲得することが確認されている．(落合)

従来より音声認識性能を改善する一般的な方法として、クリーン音声データに雑音重畳や残響畳み込みを行った学習データを用いてモデルを推定する、マルチコンディション学習手法が広く利用されている。しかしながら、マルチコンディション学習で用いられる学習データの SNR の範囲によって、大きく認識性能が変動する。SNR の分布の決定は最終的な認識性能への影響が非常に大きい為、慎重に行わなくてはならない。Sivasankaran らは、音声認識に用いられる DNN の学習データを、SNR 毎にサブセットを準備し、開発セットに対する認識性能を基準に用いて、各サブセット重み調整する事により、評価セットに対する音声認識性能を改善する手法 [20] を提案した。アブストラクトでも述べられているが、評価セットと自動チューニングされた学習セットの SNR の分布がお互いに異なっている事が報告されている。（松田）

5. 話者認識・話者識別

話者認識・話者識別については，オーラル 1 セッション (SP-L5: Speaker Diarization and Recognition)，ポスター 2 セッション (SP-P7: Speaker Verification, SP-P8: Speaker Recognition) で合計 25 件の発表が行われた．

話者照合や話者ダイアライゼーションでは，発話に含まれる話者性を i-vector によって表現し，2 つの i-vector の同一性をベクトル間距離尺度や Probabilistic linear discriminant analysis (PLDA) で判定する枠組みが一般的である．そのため，i-vector や PLDA の算出方法を改良する試みが複数発表された．

Cumani らは従来の Joint factor analysis (JFA) と i-vector 双方の利点を持つ新たな表現として e-vector と呼ばれる新たな枠組みを提案している [21]．e-vector では JFA のように話者変動のみを考慮した話者部分空間を構築し，これを初期値として i-vector の total variability 空間を再学習することで話者間変動と話者内変動を同時に考慮した部分空間を構築できることが示されている．このモデルを PLDA と組み合わせることで，従来の senote-based i-vector+PLDA に比べ EER を相対的に 10%程度改善している．

PLDA スコアを改良する方法として，登録発話が複数得られる環境下での精度改善を試みる手法が提案された．Madikeri らの手法では登録発話の i-vector 間の共分散情報を用いて PLDA モデルの共分散行列のスケールを変換することで，話者内分散をスコアレベルで考慮できる新たなスコア算出法を導入した [22]．提案手法をスコア平均法と組み合わせることで，従来の統合手法に対し EER を相対的に 29%程度改善することが示されている．

また，テキスト依存型話者認識に i-vector を適用する研究も近年盛んに研究されており，今回の ICASSP においても新たな手法が提案された．本分野でもテキスト独立型話者認識と同様に Senone DNN-based i-vector+PLDA に基づくモデルベース手法が依然主流であるが，Dey らは，新たに online i-vector と動的計画法を組み合わせたテンプレートベース手法を提案しその有効性を示した [23]．本手法では登録・照合発話に対し 200ms 毎に i-vector を算出することで各発話を i-vector 系列に変換する．これら i-vector 系列に対し動的計画法を適用することで Senone-based i-vector に基づくモデル型話者認識システムに比べ，EER を相対的に 75%程度改善することが示された．（俵）

更に，今回の ICASSP では，DNN を用いて，より識別的な（同一性判定が容易な）話者表現を得る方法が複数発表された．いずれも，「同一話者の話者表現の距離が小さくなり，異なる話者の話者表現の距離が大きくなる空間を得る」という考え方に基づいている．

Lee らは，Discriminative autoencoder (DCAE) によって i-vector を識別的な空間に射影する手法を提案した [24]．DCAE は，出力層で i-vector の再構成誤差を小さくする通常の Autoencoder の損失関数に加え，中間層においてノイズ成分を分離した上で同一話者内の距離を小さく，かつ異なる話者間の距離を大きくするように設計された目的関数によって学習される．これにより，i-vector の持つ話者情報を保持したまま識別性能が向上したベクトルが中間層の出力として得られる．DCAE の中間層出力を用いてコサイン類似度で話者同一性を判定する方法により，従来の i-vector + PLDA と比較して EER を相対的に 36%削減している．

Bredin は，LSTM と Average pooling によって発話の音響特徴量系列を固定長のベクトル（話者表現）に変換するネットワーク（TristouNet）を提案している [25]．TristouNet は，学習データからサンプリングした同一話者の 2 発話と，異なる話者の 1 発話を変換したとき，同一話者の 2 つのベクトルの距離が，異なる話者のベクトルとの距離よりも小さくなるように設計された Triplet loss 関数を最小化するように学習される．TristouNet により，2 秒程度の短い発話での話者交代検出において，BIC や Gaussian divergence に基づく従来法から精度を大きく改善できることが報告されている．

Garcia-Romero らは，発話の音響特徴量系列を固定長の話者表現に変換するネットワークと，2 つの話者表現の同一性スコアを計算する関数を同時に学習する枠組みを提案した [26]．話者表現を得るネットワークとして CNN，同一性スコアを計算する関数として 2 つのベクトルの線形変換とシグモイド関数が用いられている．最終的なスコアが同一話者の場合に高く，異なる話者の場合に低くなるように，全てのパラメータが同時に学習される．このモデルは，現在 state-of-the-art とされている Senone DNN-based i-vector + PLDA に比べ非常にシンプルで，パラメータ数

は約 1/50 であるにも関わらず，話者ダイアライゼーションにおいて同等の精度を達成することが確認された．（浅見）

6. 音声合成・声質変換

音声合成・声質変換に関するセッションは，オーラル 1 つ (SP-L4: Speech Synthesis)，ポスター 1 つ (SP-P10: Speech Synthesis and Voice Conversion) で構成され，2 セッション合わせて 16 件の講演があった．（高道）

6.1 短遅延音声合成

Wang らは単方向 Long short-term memory (LSTM) と畳込みニューラルネットワーク (CNN) を組み合わせた統計的パラメトリック音声合成手法について提案している [27]．本手法で採用されているネットワーク構造は，中間層が単方向 LSTM とアフィン変換層，出力が CNN で構成されている．時刻 t の出力音響特徴量は，LSTM からアフィン変換層を通じて得られた中間パラメータを，予め定義された先読み数 N に基づき，時刻 $t+N$ までの分を CNN によって畳み込むことで得られる．このネットワーク構造により，過去から未来までの音響的および言語的情報を考慮しつつ，短遅延で当該時刻の音響特徴量を出力することが可能となる．実験では，従来の LSTM を用いた DNN 音声合成と比較して有意に品質が改善することを示した．また出力を静的特徴量のみにすることで，動的特徴量を同時に出力する場合よりわずかに品質が向上することが確認された．（大谷）

6.2 ロンバード音声合成

Bollepalli らはロンバード効果について，LSTM に基づく音声合成において調査を行っている [28]．ロンバード音声の合成を目標とした 3 種の適応手法を試みており，それぞれ，ロンバード効果の有無を表現する one-hot ベクトル，Learning hidden unit contributions (LHUC)，ロンバード音声を用いた追加の fine-tuning を利用している．客観評価実験の結果では適応発話数が 10 と 500 のそれぞれ，追加の fine-tuning による適応が一番良い結果となった．また，HMM 音声合成と LSTM 音声合成を比較した主観評価結果において，LSTM を用いる有効性が示した．（高木）

6.3 特徴量抽出

Hu らは，what-where auto-encoder を用いたスペクトル特徴量抽出法を提案している [29]．畳み込み構造を持つ auto-encoder による教師なし特徴量抽出は，広く利用される技術である．しかしながら，周波数方向に窓をもつ max-pooling をスペクトルパラメータに適用する場合，その周波数帯域における最大値（例えばフォルマントの強さ．what に相当．）は抽出されるが，どの周波数から抽出されたか（例えばフォルマント周波数．where に相当．）の情報は失われる．そのため，特徴量抽出における過剰平滑化が生じると，著者らは主張している．本論文では，STRAIGHT によるスペクトル包絡に対して，what と where の情報を保存する what-where auto-encoder による特徴量抽出を行い，合成音声を用いた評価で有効性を確認した．（高道）

6.4 音声特徴量の依存関係のモデル化

Li らは，ピッチパラメータとスペクトルパラメータ間の依存性を考慮した構造的出力層を用いた bidirectional LSTM (BLSTM)-RNN に基づく音声合成を提案している [30]．従来の DNN や LSTM-RNN に基づく音響モデルでは，ネットワークの内部でピッチパラメータとスペクトルパラメータの依存性がモデル化されることを期待していたが，本手法では，出力層に予測したピッチパラメータをスペクトルパラメータの予測に利用する構造を導入することで，ピッチパラメータとスペクトルパラメータ間の依存性をより明確に考慮する．また，これらのパラメータのコスト関数に重み係数を導入した．客観・主観評価実験の結果から，構造的出力層を用いた BLSTM-RNN は従来法から改善し，有効性を示した．（橋本）

7. HLT

HLT (human language technology) についてはオーラル 2 セッション，ポスター 3 セッションの合計 5 セッションがあった．Language Modeling (HLT-P1), Spoken Term Detection (HLT-P2), Spoken Language Understanding I & II (HLT-L2, HLT-P3), Keyword Search (HLT-L1) の 5 セッションである．本節では，HLT セッションの発表を中心として，言語モデル，音声検索語検出，音声対話システムについて述べる．（南條）

7.1 言語モデル

言語モデルについてはポスター 1 セッションのみであり，10 件の発表があった．以下では，いくつかの興味深い発表を取り上げて概説する．

ここ数年の言語モデル研究では，RNN を用いることで長距離のコンテキストを考慮できるようなモデル化の検討が広く検討されている．これまでは，単一話者を想定したモデル化が広く行われてきたが，Liu らは 2 話者の対話におけるインタラクションをコンテキストとして考慮可能な RNN 言語モデルを提案している [31]．Switchboard Dialog Act Corpus を用いた実験において，対話インタラクションを陽に考慮することでパープレキシティの改善が報告されている．今後このような検討は，さらに多人数話者に拡張して発展していくと考えられる．

単語よりも短い単位を言語モデルで扱う検討も広がっている．Hwang らは文字レベルの RNN 言語モデルにおいて単語単位のコンテキストを直接考慮する方法を提案して

ⓒ 2017 Information Processing Society of Japan

いる [32]．文字レベル言語モデルの欠点は，単語単位を考慮できない点であるが，この研究では単語境界を示す文字（スペース）とその他の文字を分けて扱うことで，単語単位のコンテキストを文字レベルの言語モデルに伝搬する機構を導入している．実験では，文字を扱う CTC ベースの End-to-End 音声認識に導入することで，単語誤り率の改善を報告している．また，Irie らは単語埋め込みを Unicode から組み上げる RNN 言語モデルを検討している [33]．組み上げる際には，近年有効とされている CNN とマックスプーリングを用いる方法を用いており，少資源言語における実験において有効であることを報告している．

言語モデルに関連する研究として，句読点付与（エクスクラメーション，クエスチョンを含む）がある．Klejch らは，音声認識結果から Sequence-to-Sequence の方式で句読点付与を行う際に，言語特徴のみでなく音響特徴も利用する方法を提案している [34]．この研究では，句読点付与が単語境界で発生することを考慮して，音響特徴を単語レベルで組み上げる機構を導入している．実験では，言語特徴と音響特徴を組み合わせることで，精緻に句読点付与が可能であることを報告している．言語特徴と音声特徴を組み合わせる観点において、非常に有用な方式であると考えられる．（増村）

7.2 音声検索語検出

Spoken Term Detection（STD）は，日本語では「音声検索語検出」と訳され，音声中で検索対象の語句がそのまま出現する位置や発話区間を特定する処理を指す．ICASSP2017 では，1 つのポスターセッション（HLT-P2）が設けられ 10 件の発表があった（別途 Keyword Search というオーラルセッション（HLT-L1）での 6 件の発表も含めると 16 件．ただし，この節では HLT-L1 は扱わない）．それらのうちのいくつかの発表を取り上げる．

日本でも STD の研究は盛んであり，国立情報学研究所による情報アクセス技術の評価のためのワークショップの NTCIR-9[35], NTCIR-10[36], NTCIR-11[37], NTCIR-12[38] において，SpokenDoc, SpokenQuery&Doc のサブタスクとして STD が設定され，評価基盤が整備されている．日本国内ではこの基盤データを使った研究が盛んであるが，HLT-P2 STD セッションでは残念ながら使われている発表はなかった．ICASSP2017 では，IARPA BABEL を使った発表が目立っており，Multilingual システム，特に low-resource 言語をターゲットとしている研究が目につく．STD ではなく “KWS (keyword search)” と表記されていることがあるが，本節では KWS にも「検索語検出」の訳をあてている．

音声検索語検出は基本的に音声認識の精度に影響を受けるため，音声認識の精度向上が重要である．そのような研究として，RNN 言語モデルを用いた研究 [39][40][41] が見られた．文献 [39], [40] では，RNN 言語モデルから新たなテキストを生成して学習に用いている．文献 [40] では，さらに機械翻訳も使って low-resource 言語の言語モデルを強化している．これらの研究では，音声認識精度と検索語検出精度の改善が報告されている．

音声検索語検出では，検出時の照合単位も重要である．典型的には，音素や音節といった言語知識を利用した単位（トップダウン）が利用されるが，データから構築するボトムアップのアプローチもある．文献 [42] では Zero-resource 言語での検出を，文献 [43] では音声認識辞書に登録されていない未知語の検出を頑健に行うために，音素や音節の役目をする照合単位の自動獲得を目指している．文献 [42] では，高速化のための並列計算も提案している．

文献 [44] では，種々の音響，言語モデルからなる多様な音声認識結果，検索結果を組み合わせる効果を示している．近年提案された効果的な音声認識，検索語検出の技術を組み合わせたものであり，現状の技術と精度を確認するにはよいであろう．

文献 [45] では，言語間での検出精度の比較を行っている．検索対象語の長さ，他の語との紛らわしさなどと検出精度の関係を調べている．言語依存の特徴は示されなかったが，さらなる分析が期待される．日本語のデータも示されており，他の言語と比較して難しいように思われる．これは，日本語の音声認識精度が低いのもあるが，日本語の音素数が少なく音素列が近い単語が多いことも一因のように思われる．日本語での検索語検出精度は他の言語と比べて低く，改善する余地が多い．日本語に適した方法を研究し，日本語以外の類似する性質を持つ言語に応用できることを示していくのも一つの案といえる．

また，文献 [45] ではネイティブの人間同士で音素ラベリングが一致しないようなものに関しては，検出精度が低いことが示されている．当たり前のように思えるが，人間が検出できない音声を検出できれば非常に役に立つと思われるので，こういった方向の研究も考えてもよいであろう．（南條）

7.3 音声対話システム

ICASSP での音声対話システムの発表は比較的少数であり，今年は対話システムに関する独立したセッションはなかった．一方で，初日に，Yun-Nung Chen, Asli Celikyilmaz, Dilek Hakkani-Tür により T2: Deep Learning for Dialogue Systems というチュートリアルが行われた．目的指向型の音声対話システムの各モジュール（音声言語理解，対話状態追跡，対話戦略，自然言語文生成）を，従来の機械学習に基づく手法から深層学習に置き換える話が主であった．これ以外にも，上記の各モジュールの入出力を組み合わせて End-to-End 問題として捉える（ユーザ発話からシステム応答への seq2seq モデルを含む）試みの紹介

や，システム評価の方法論や最近の動向についても言及があった．このチュートリアルの発表者を著者に含む発表も音声言語理解のセッションにおいて行われた [46]．資料は Web で入手できる（2017 年 6 月現在）[*2]．

音声対話システムでの言語理解 (NLU) において，未知の入力が現れた際の問題に取り組んだ研究もあった [47]．ここでは未知の入力として personalized concepts を挙げている．これは例えば，ユーザが具体的な内容を言わずに，場所や時間を “near my workplace” や “around brunch time” と言う場合に相当する．まず入力文中の未知部分の同定のために，LSTM ベースのパーザの結果に加えて，パーズ結果の確信度や文の係り受け構造を利用することの有効性を示している．そのうえで，同定された未知のスロット値 XXX に対して，”Can you define XXX?” と尋ねてその具体的な値を取得し，それを当初の未知語部分に代入して再パーズすれば，より高い性能で言語理解が可能となるとしている．“Alice’s birthday” を personalized concepts とした場合の対話例を以下に示す．

User: I need a reservation on Alice’s birthday at Evvia.
System: Can you define “Alice’s Birthday”?
User: Alice’s birthday is March 9.
System: Reserve(restaurant=Evvia,date=3/9)

システムとの対話におけるユーザの満足度などの予測を，音響的特徴を使って向上させる試みもなされた [48]．ユーザが探していた情報を得られたと感じた度合や，システムが自分を理解していたかどうかといった主観的な尺度を対話終了後にユーザに尋ね，それらを 3 値程度に離散化したものを目的変数として予測する．さらにはターン数などの客観的指標も予測対象としている．ベースラインの特徴セットとして，ある従来研究で使われていた，対話状態に関する特徴（ユーザの対話行為やシステムの行為，対話状態のエントロピーなど）が採用されている．これに加えて音響的特徴として，音声波形の RMS (root mean square) 値やピッチから得た情報を使うことで，予測性能が向上したとしている．例えば音響的特徴によりターン数の予測性能が向上しているが，これは対話が長くなるとユーザの声が苛立つことに相当するとしている．ベースラインの特徴が少なく必ずしも十分とは思えないが，音響的特徴の活用法のひとつを示す研究である．（駒谷）

参考文献

[1] Airaksinen, M., Bollepalli, B., Pohjalainen, J. and Alku, P: Frequency-warped time-weighted linear prediction for glottal vocoding, *Proc. ICASSP*, Mar. 2017, pp. 5630–5634.
[2] Chen, T.-H., Huang, C. and Chi, T.-S.: Dereverberation based on bin-wise temporal variations of complex spectrogram, *Proc. ICASSP*, Mar. 2017, pp. 5635–5639.
[3] Bisot, V., Essid, S. and Richard, G.: Overlapping sound event detection with supervised nonnegative matrix factorization, *Proc. ICASSP*, Mar. 2017, pp. 31–35.
[4] Serizel, R., Bisotz, V., Essidz, S. and Richard, G.: Supervised group nonnegative matrix factorisation with similarity constraints and applications to speaker identification, *Proc. ICASSP*, Mar. 2017, pp. 36–40.
[5] Leglaive, S., Simsekli, U., Liutkus, A., Badeau, R. and Richard, G.: Alpha-stable multichannel audio source separation, *Proc. ICASSP*, Mar. 2017, pp. 576–580.
[6] Kameoka, H., Kagami, H. and Yukawa, M.: Complex NMF with the generalized Kullback-Leibler divergence, *Proc. ICASSP*, Mar. 2017, pp. 56–60.
[7] Kagami, H., Kameoka, H. and Yukawa, M.: A majorization-minimization algorithm with projected gradient updates for time-domain spectrogram factorization, *Proc. ICASSP*, Mar. 2017, pp. 561-565.
[8] Deleforge, A. and Traonmilin, Y.: Phase unmixing: Multichannel source separation with magnitude constraints, *Proc. ICASSP*, Mar. 2017, pp. 161–165.
[9] Smaragdis, P. and Venkataramani, S.: A neural network alternative to non-negative audio models, *Proc. ICASSP*, Mar. 2017, pp. 86–90.
[10] Kinoshita, K., Delcroix, M., Ogawa, A., Higuchi, T. and Nakatani, T.: Deep mixture density network for statistical model-based feature enhancement, *ICASSP*, Mar. 2017, pp. 251–256.
[11] Kim, M.: Collaborative deep learning for speech enhancement: A run-time model selection method using autoencoders, *Proc. ICASSP*, Mar. 2017, pp. 76–81.
[12] Wang, Z.-Q. and Wang, D.: Recurrent deep stacking networks for supervised speech separation, *Proc. ICASSP*, Mar. 2017, pp. 71–75.
[13] Meng, Z., Watanabe, S., Hershey, J. R. and Erdogan, H.: Deep long short-term memory adaptive beamforming networks for multichannel robust speech recognition, *Proc. ICASSP*, Mar. 2017, pp. 271–275.
[14] Zhang, X., Wang, Z.-Q. and Wang, D.: A speech enhancement algorithm by iterating single- and multi-microphone processing and its application to robust ASR, *Proc. ICASSP*, Mar. 2017, pp. 276–280.
[15] Yu, D., Kolbaek, M., Tan, Z.-H. and Jensen, J.: Permutation invariant training of deep models for speaker-independent multi-talker speech separation, *Proc. ICASSP*, Mar. 2017, pp. 241–245.
[16] Kanda, N., Lu, X. and Kawai, H.: Minimum Bayes risk training of CTC acoustic models in maximum a posteriori based decoding framework, *Proc. ICASSP*, Mar. 2017, pp. 4855–4859.
[17] Rao, K. and Sak, H.: Multi-accent Speech Recognition with Hierarchical Grapheme Based Models, *Proc. ICASSP*, Mar. 2017, pp. 4815–4819.
[18] Cui, J., Kingsbury, B., Ramabhadran, B., Saon, G., Sercu, T., Audhkhasi, K., Sethy, A., Nussbaum-Thom, M. and Rosenberg, A.: Knowledge Distillation Across Ensembles of Multilingual Models for Low-resource Languages, *Proc. ICASSP*, Mar. 2017, pp. 4825–4829.
[19] Ravanelli, M., Brakel, P., Omologo, M. and Bengio, Y.: A Network of Deep Neural Networks for Distant Speech Recognition, *Proc. ICASSP*, Mar. 2017, pp. 4880–4884.
[20] Sivasankaran, S., Vincent, E. and Illina, I.: Discriminative importance weighting of augmented training data for acoustic model training, *Proc. ICASSP*, Mar. 2017, pp. 4885–4889.

[*2] https://sites.google.com/site/deeplearningdialogue/

ⓒ 2017 Information Processing Society of Japan

[21] Cumani, S. and Laface, P.: E-vectors: JFA and i-vectors revisited, *Proc. ICASSP*, Mar. 2017, pp. 5435–5439.
[22] Madikeri, S., Ferras, M., Motlicek, P. and Dey, S.: Intra-class covariance adaptation in PLDA back-ends for speaker verification, *Proc. ICASSP*, Mar. 2017, pp. 5365–5369.
[23] Dey, S., Motlicek, P., Madikeri, S. and Ferras, M.: Exploiting sequence information for text-dependent speaker verification, *Proc. ICASSP*, Mar. 2017, pp. 5370–5374.
[24] Lee, H.-S., Lu, Y.-D., Hsu, C.-C., Tsao, Y., Wang, H.-M. and Jeng, S.-K.: Discriminative Autoencoders for Speaker Verification, *Proc. ICASSP*, Mar. 2017, pp. 5375–5379.
[25] Bredin, H.: Tristounet: Triplet Loss for Speaker Turn Embedding, *Proc. ICASSP*, Mar. 2017, pp. 5430–5434.
[26] Garcia-Romero, D., Snyder, D., Sell, G., Povey, D. and McCree, A.: Speaker Diarization Using Deep Neural Network Embeddings, *Proc. ICASSP*, Mar. 2017, pp. 4930–4934.
[27] Wang, W. and Xu, B.: COMBINING UNIDIRECTIONAL LONG SHORT-TERM MEMORY WITH CONVOLUTIONAL OUTPUT LAYER FOR HIGH-PERFORMANCE SPEECH SYNTHESIS, *Proc. in ICASSP2017*, pp. 5500–5504.
[28] B. Bollepalli, M. Airaksinen, P. A.: LOMBARD SPEECH SYNTHESIS USING LONG SHORT-TERM MEMORY RECURRENT NEURAL NETWORKS, *Proc. ICASSP*, Mar. 2017, pp. 4915–1919.
[29] Hu, Y.-J., Ling, Z.-H. and Dai, L.-R.: EXTRACTING STRUCTURAL SPECTRAL FEATURES USING WHAT-WHERE AUTO-ENCODERS FOR STATISTICAL PARAMETRIC SPEECH SYNTHESIS, *Proc. ICASSP*, Mar. 2017, pp. 4915–1919.
[30] Li, R., Wu, Z., Liu, X., Meng, H. and Cai, L.: MULTI-TASK LEARNING OF STRUCTURED OUTPUT LAYER BIDIRECTIONAL LSTMS FOR SPEECH SYNTHESIS, *Proc. ICASSP*, Mar. 2017, pp. 5510–5514.
[31] Liu, B. and Lane, I.: Dialog Context Language Modeling with Recurrent Neural Networks, *Proc. ICASSP*, Mar. 2017, pp. 5715–5719.
[32] Hwang, K. and Sung, W.: Character-level Language Modeling with Hierarchical Recurrent Neural Networks, *Proc. ICASSP*, Mar. 2017, pp. 5720–5724.
[33] Irie, K., Golik, P., Schluter, R. and Ney, H.: Investigations on Byte-level Convolutional Neural Networks for Language Modeling in Low Resource Speech Recognition, *Proc. ICASSP*, Mar. 2017, pp. 5740–5744.
[34] Klejch, O., Bell, P. and Renals, S.: Sequence-to-Sequence Models for Punctuated Transcription combining Lexical and Acoustic Features, *Proc. ICASSP*, Mar. 2017, pp. 5700–5704.
[35] Akiba, T., Nishizaki, H., Aikawa, K., Kawahara, T. and Matsui, T.: Overview of the IR for Spoken Documents Task, *NTCIR-9 Workshop Meeting*, pp. 223–235 (2011).
[36] Akiba, T., Nishizki, H., Aikawa, K., Hu, X., Itoh, Y., Kawahara, T., Nakagawa, S., Nanjo, H. and Yamashita, Y.: Overview of the NTCIR-10 SpokenDoc-2 Task, *NTCIR-10 Workshop Meeting*, pp. 573–587 (2013).
[37] Akiba, T., Nishizaki, H., Nanjo, H. and Jones, G. J. F.: Overview of the NTCIR-11 SpokenQuery&Doc Task, *NTCIR-11 Workshop Meeting*, pp. 350–364 (2014).
[38] Akiba, T., Nishizaki, H., Nanjo, H. and Jones, G. J. F.: Overview of the NTCIR-12 SpokenQuery&Doc-2 Task, *NTCIR-12 Workshop Meeting*, pp. 167–179 (2016).
[39] Lileikyte, R., Fraga-Silva, T., Lamel, L., Gauvain, J. L., Laurent, A. and Huang, G.: Effective keyword search for low-resourced conversational speech, *Proc. ICASSP*, Mar. 2017, pp. 5785–5789.
[40] Huang, G., da Silva, T. F., Lamel, L., Gauvain, J. L., Gorin, A., Laurent, A., Lileikyte, R. and Messouadi, A.: An investigation into language model data augmentation for low-resourced STT and KWS, *Proc. ICASSP*, Mar. 2017, pp. 5790–5794.
[41] Chen, X., Ragni, A., Vasilakes, J., Liu, X., Knill, K. and Gales, M. J. F.: Recurrent neural network language models for keyword search, *Proc. ICASSP*, Mar. 2017, pp. 5775–5779.
[42] Oosterveld, B., Veale, R. and Scheutz, M.: A parallelized dynamic programming approach to zero resource spoken term discovery, *Proc. ICASSP*, Mar. 2017, pp. 5800–5804.
[43] van Heerden, C., Karakos, D., Narasimhan, K., Davel, M. and Schwartz, R.: Constructing sub-word units for spoken term detection, *Proc. ICASSP*, Mar. 2017, pp. 5780–5784.
[44] Alumae, T., Karakos, D., Hartmann, W., Hsiao, R., Zhang, L., Nguyen, L., Tsakalidis, S. and Schwartz, R.: The 2016 BBN Georgian telephone speech keyword spotting system, *Proc. ICASSP*, Mar. 2017, pp. 5755–5759.
[45] Hartmann, W., Karakos, D., Hsiao, R., Zhang, L., Alumae, T., Tsakalidis, S. and Schwartz, R.: Analysis of keyword spotting performance across IARPA babel languages, *Proc. ICASSP*, Mar. 2017, pp. 5765–5769.
[46] Yang, X., Chen, Y.-N., Hakkani-Tür, D., Crook, P., Li, X., Gao, J. and Deng, L.: End-to-End Joint Learning of Natural Language Understanding and Dialogue Manager, *Proc. ICASSP*, Mar. 2017, pp. 5690–5694.
[47] Jia, R., Heck, L., Hakkani-Tür, D. and Nikolov, G.: Learning Concepts through Conversations in Spoken Dialogues Systems, *Proc. ICASSP*, Mar. 2017, pp. 5725–5729.
[48] Papangelis, A., Kotti, M. and Stylianou, Y.: Predicting Dialogue Success, Naturalness, and Length with Acoustic Features, *Proc. ICASSP*, Mar. 2017, pp. 5010–5014.

ⓒ 2017 Information Processing Society of Japan

情報処理学会研究報告 Vol.2017-SLP-117 No.3

2017.7.28

国際会議 ICASSP2017 報告

正誤表

誤	正
< 7.2 節 6 ページ右側 上から 22 行目 > 「日本語のデータも示されており，他の言語と比較して難しいように思われる．これは，日本語の音声認識精度が低いのもあるが，日本語の音素数が少なく音素列が近い単語が多いことも一因のように思われる．日本語での検索語検出精度は他の言語と比べて低く，改善する余地が多い．」	「日本語のデータは示されていない．」