Vol.2018-MUS-118 No.14
Vol.2018-SLP-120 No.14
2018/2/21

国際会議Interspeech2017報告

高木 信二[1]　倉田 岳人[2]　郡山 知樹[3]　塩田 さやか[4]　鈴木 雅之[2]　玉森 聡[5]　俵 直弘[6]　中鹿 亘[7]
福田 隆[2]　増村 亮[8]　森勢 将雅[9]　山岸 順一[1]　山本 克彦[10]

概要：2017 年 8 月 20 日から 8 月 24 日にかけ，ストックホルム・スウェーデンで Interspeech2017 が開催された．Interspeech 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付けられており，今後の本分野の動向に大きく影響を与えている．本稿では，本会議における研究動向，注目すべき発表について報告する．

1. はじめに

2017 年 8 月 20 日から 8 月 24 日にかけ，ストックホルム・スウェーデンで Interspeech2017 が開催された．Interspeech 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付けられている．1,711 件の投稿があり (内 1,582 件がレビューされ)，799 件が受理された．本稿では，本会議における研究動向，注目すべき発表について，音声認識，話者認識，音声明瞭性・音声分析，音声合成を中心として報告する．

2. 音声認識

2.1 電話会話音声認識

電話会話音声認識を対象として，音声認識システムと人間の音声認識能力の比較に関して報告が行われていた．英語のベンチマーク評価データの Switchboard (SWB) と CallHome (CH) を対象としているが，SWB は他人同士の既知の話題についての会話，CH は知人，家族同士の未知の話題についての会話という異なる特徴がある．人間の認識誤り率は，[1] では，SWB と CH について 5.9% と 11.3% と報告されているが，[2] では各々 5.1% と 6.8% と報告されている．両者では人間による書き起こしパイプラインが異なり，「人間の音声認識能力」の定義自身から活発な議論が行われていた．また，[2], [3] では，音声認識率システムによるこの時点での最高性能として，SWB で 5.5%，CH で 10.3% という認識率が報告されている．音響モデルとして LSTM，ResNet を利用したハイブリッドシステムによって N-best 仮説を出力し，LSTM および CNN を利用した言語モデルによるリスコアリングによってこれらの認識率は達成されている．電話会話音声認識の性能は，SWB に限定した場合，人間の音声認識能力とほぼ同等となっており，共通のデータセットによる産学での競争は，音声認識システムの性能向上に寄与していると言える．(倉田)

2.2 End-to-end 音声認識

音響モデル・発音モデル・言語モデルを別々に扱う hybrid 型の音声認識に対し，それら単一のモデルで扱う end-to-end 型の音声認識にも様々な発表があった．音声認識率の面では，SWB/CH 等のタスクでは hybrid 型が，CSJ 等のタスクでは end-to-end 型がそれぞれ state-of-the-art の性能を達成している．開発の面では，end-to-end 型にはシステム全体の構築に必要な作業が少ないという利点が，hybrid 型にはドメイン適応などの調整が行いやすいという利点がある．今後も暫くの間は，hybrid 型と end-to-end 型の両方の研究・開発が行われることが予想される．

[4] では，end-to-end 型のモデルである grapheme-CTC，RNN-Transducer，Attention-based encoder-decoder を，様々な評価タスクで比較している．学習データサイズは約 12,500 時間である．結果，言語モデルを併用せず単独で最も性能が高いのは Attention-based encoder-decoder であった．また，hybrid 型音声認識システムと比較すると，数字認識タスクでは end-to-end 型が，ディクテーションタスクでは性能は同程度，ボイスサーチタスクでは hybrid 型の方が，良い性能を示している．これは，発音モデリングと言語モデリングが簡単なタスクでは end-to-end 型が

1 国立情報学研究所
2 日本 IBM 株式会社
3 東京工業大学
4 首都大学東京
5 名古屋大学
6 早稲田大学
7 電気通信大学
8 日本電信電話株式会社
9 山梨大学
10 和歌山大学

優れており，難しいタスクでは hybrid 型が優れていることを示唆している．[5] は，言語モデルを併用しないと高い性能が得られない文字単位の CTC に対し，言語モデルを併用しなくても高い性能が得られる，単語を直接出力する word-CTC の研究を行っている．Word-CTC は YouTube 125,000 時間のような超大量データを用いた場合には高い性能を示すことが既に知られている [6]．[5] では，文字などのより細かい単位の CTC と，word embeddig を用いて pre-training を行うことで，英語電話会話音声 2000 時間を学習データとするの SWB/CH でも高い性能が得られるようになっている．ただし，state-of-the-art の hybrid 型の音声認識率と比べると，まだ大きな差がある．[7] では，CTC と attention-based encoder-decoder を併用してデコーディングを行う手法，さらにそれを LSTM LM でリスコアリングする手法を提案している．この手法を用いることで，日本語講義音声 581 時間を学習データとする CSJ と，中国語電話会話音声 167 時間を学習データとする MTS で，state-of-the-art の性能を達成している．(鈴木)

2.3 Knowledge distillation

近年，knowledge distillation の枠組みを音声認識に利用する研究が急加速している．今回の Interspeech 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 [8], [9], [10]．これは教師となるニューラルネットワークに学習データを入力し，そこから生成される事後確率分布を生徒となるニューラルネットワークのためのソフトラベルとして利用する方法であり，教師ネットワークの識別能力を生徒ネットワークへ転移させることを目指している．典型的には，教師と生徒ネットワーク間の出力分布の KL ダイバージェンスを最小化するように学習を行うことで，教師から生徒ネットワークへの “知識蒸留” が実現され，教師側には認識精度重視の大規模なネットワーク，生徒側にはデコード速度を重視したコンパクトな構成のネットワークが採用されることが多い．

Li らは knowledge distillation の枠組みを利用した音響ドメイン適応法を提案した [8]．通常，高精度な音響適応処理には，発話内容を表す書き起こしデータが必要となるが，生徒側ネットワークの学習が教師ネットワークから生成されるソフトラベルのみを用いて進められるという性質を利用して，書き起こしデータ不要の適応処理の可能性を，クリーン音声の音響モデルから noisy 音声の音響モデルへの適応および，大人音声モデルから子供音声モデルへの適応という形の実験にて実証した．具体的にはソースドメイン（教師）とターゲットドメイン（生徒側）のパラレルデータを用意し，両者の出力の KL ダイバージェンスが最小となるように学習を進めることで，音響ドメインが適応された生徒ネットワークを得る．

上記 Li らの方法を含めて，従来の knowledge distillation は，教師と生徒ネットワークの出力層の構成が同じ，すなわち音素決定木が同一であるということを暗黙に仮定している．しかし，音素決定木は音響特性の違いを表す大きな要素の一つであるため，音素決定木が同一であるという仮定は，knowledge distillation の能力や，生徒側のネットワーク構成のフレキシビリティを制限してしまうことになる．この問題に対して Wong らは，複数の教師ネットワーク，および生徒ネットワーク間で異なる音素決定木を持つネットワークに対する knowledge distillation を提案した [9]．生徒側に存在しない音素状態クラスについて，教師の音素状態クラスから事後確率を推定する，すなわち生徒・教師間の論理音素クラスに対する事後確率分布の KL ダイバージェンスを最小化することによって生徒ネットワークの学習を進めている．(福田)

2.4 言語モデル

今回の Interspeech における言語モデルのセッションでは，ニューラル言語モデルのドメイン適応に関する研究発表が散見された．言語モデルの学習データとして，書き言葉調のテキストデータを集めることは容易であるが，話し言葉調のテキストデータを大量に集めることは困難である．そこで，言語モデルにおけるドメイン適応では，書き言葉調のテキストデータから学習したモデルをバックグラウンドモデルとして，話し言葉調のテキストデータで適応化することで，高精度なモデル化を目指している．

Deena らは，話し言葉におけるドメイン特有の補助特徴量を用いた RNN 言語モデルを構築する際に，書き言葉調のテキストデータを活かすための学習手法を提案している [11]．この場合，書き言葉調のテキストデータからは補助特徴量を抽出できないことが課題であるが，テキストから補助特徴量を予測するモデルを別途学習しておくことで，書き言葉調のテキストデータからバックグラウンドモデルを構築して適応化を行っている．Ma らは，DNN 言語モデルと LSTM 言語モデルのドメイン適応のために，いくつかのモデルベースの適応手法の比較を行っており，バックグラウンドモデルの一部に線形層を設けて学習しておき，その層のみファインチューニングする手法が有望であることを報告している [12]．また Singh らは，LSTM 言語モデルにバックオフ n-gram 構造に近似するという前提のもと，Marginal 適応の導入を提案している [13]．Marginal 適応は，これまでトピックモデルに基づく言語モデル適応等のために頻繁に利用されてきたものであるが，ニューラル言語モデルにおいても有効である点は非常に興味深い．これらのドメイン適応の検討では，現状大幅な改善効果は得られていないものの，ニューラル言語モデルの実用的な利用に向けて，今後さらに重要な研究領域となるだろう．

end-to-end 型の音声認識とニューラル言語モデルの組み合わせについても検討が進んできている．end-to-end 型の

音声認識は，入力の音声信号に条件付けられた言語モデルとして表すことができ，言語モデルを間接的に内包していることが知られている．これに対して Hori らは，end-to-end 型の音声認識と RNN 言語モデルの組み合わせる方法ついて興味深い報告をしている [7]．組み合わせる方法としては，対数確率レベルで線形補間する方法と，ソフトマックス関数による確率値への変換前にベクトルレベルで統合する方法の 2 つを検討しており，後者が特に有効であることを報告している．注目すべきは，RNN 言語モデルの学習データと end-to-end 型の音声認識の学習データが同一であっても，組み合わせにより改善効果を得ている点である．すなわち，言語モデル単体でのモデル化は，end-to-end 型の音声認識で間接的にモデル化する場合とは異なる特性を持っていると考えられ，言語モデル単体でのモデル化が今後も重要な技術であることを示唆している．(増村)

3. 話者認識

話者認識では，2 つのスペシャルセッションと 4 つのオーラルセッション，3 つのポスターセッションがあり計 74 件の発表があった．

オーラルセッションの 1 つは話者認識のコンペティションである The 2016 NIST Speaker Recognition Evaluation (SRE16) に関する報告で 6 件の発表があった．[14] ではコンペティションの説明や提出されたシステムの傾向等が詳しく解析されている．英語音声が主な対象だった SRE12 に対し，SRE16 では対象言語の変更（広東語、タガログ語）や収録環境の増加、およびテスト発話長の変動の増加等の変更があった．以上を踏まえ [14] では全体の傾向として，これら音響的ミスマッチが増加した影響により，いずれのシステムも例年に比べ性能が低かったことを指摘している．また，今回は与えられた学習セット以外のデータも利用可能とするタスクが用意されたが，いずれのシステムでもデータ追加による効果は限定的か，逆に性能を低下させるものであった．報告ではその原因として言語や収録環境のミスマッチやデータ量の不足を指摘しており，これら環境の違いを吸収するための更なる研究が必要であると結んでいる．具体的なアプローチは，いずれの報告も従来の i-vector に基づくシステムを主体とするもので，特徴量の種類や i-vector の導出法の組み合わせを変えることで複数の相補的なシステムを構築し，統合することで性能改善を目指すというアプローチであった．特に， Nuance とトリノ大学による手法 [15] では，特徴量として MFCC または PLP を，モデル手法として DNN または GMM に基づく i-vector を，スコア算出に pairwise SVM を採用し，これら要素技術の組み合わせを変え構築した 7 システムを統合することで SRE16 における最高性能を達成している．

特徴抽出に関する重要な新規技術としては，Snyder らによる Time delay DNN (TDNN) に基づく end-to-end の発話表現抽出法 [16] があった．従来の i-vector における DNN の利用は，DNN-HMM 音響モデルから得られる音素状態 (senone) の事後確率を i-vector 算出時の統計量に利用するものか，DNN-HMM 音響モデルのボトルネック層の出力から直接 i-vector を構築するアプローチが主であった．DNN により発話表現を end-to-end で得る手法も試みられているが，いずれの手法もネットワーク構築時に数万人規模の非常に多くの話者数が必要であるという問題があった．対して，本研究では TDNN と statistic pooling を組み合わせることで，6000 話者程度の学習データでも従来の i-vector よりも高い性能を達成する発話表現が得られることを示した，

近年，注目を浴びている Anti-spoofing に関連する報告として 3 つのオーラルセッションがあり，そのうちの 2 セッションがなりすまし攻撃に対するコンペティションである ASVspoof 2017 Challenge に関する報告であった．ASVspoof 2017 Challenge は 2015 年に開催された ASVspoof 2015 Challenge の第二弾となっている．前回の ASVspoof2015 においては，なりすまし攻撃が様々な手法による音声合成や声質変換を使った合成音声がシステムに直接入力されることを想定していた．一方，ASVspoof2017 ではなりすまし攻撃として，収録された実発話をスピーカー等で再生し，再収録したものが用意された．その際の収録機器・再生機器のバリエーションおよび環境も様々なものが用意されており，前回よりも難易度が高い設定となっていた．参加チームの結果や使用した特徴量・識別器などの大まかなまとめに関しては ASVspoof 2017 Challenge の HP [17] および [18] に記載されている．傾向としては前回と同様, GMM-UBM ベースの識別器を用いるものが主流であったが GMM-UBM と合わせて CNN や RNN などのニューラルネットを使った手法とスコア統合することで高い識別率を得たところもあった．前回は多くの参加チームが MFCC に加えて位相などを使った様々な種類の特徴量抽出を行うことで高い識別性能を得ていたが今回は有用な特徴量だけを選択して使用するという傾向に変わっていた．実際，上位 10 チームの使用した特徴量の種類は平均して 3 程度であった．特に, ロシアの ITMO 大学と STC-innovations による手法 [19] において，使用された特徴量は i-vector 用の LPCC とニューラルネットのための FFT のみだった．識別器には Max-Feature-Map (MFM) という活性化関数を畳込み層で用いることで CNN の構造削減を行った Light CNN， i-vector を用いた SVM, CNN と RNN による end-to-end 識別という 3 種類のシステムをスコア統合することで ASVspoof 2017 における最高性能を達成していた．また，ベースラインとして公開されている Constant Q cepstral coefficients (CQCC) を特徴量として用いたものも多く，2 番目に高い性能を得た [20] においても用いられている．こちらの手法も使用した特徴量は

CQCC, MFCC, PLP と少なく，代わりに決定木を用いた勾配ブースティング法 （GBDT） やランダムフォレストなど弱識別器によるアンサンブル学習を組み合わせる手法となっていた．再収録音声による初めてのコンペティションであったためそれほど突出した技術があったわけではないが今後は更に深層学習を取り入れる手法が多く出てくると予想される．（俵・塩田）

4. 音声明瞭性，音声分析

4.1 音声明瞭性

意味を持つ音声（単語や文章）の明瞭性を定量的に表した音声了解度 (speech intelligibility) は，雑音の有無や歪みの程度を評価する音声品質 (speech quality) と同様に，音声の評価項目として重要な要素である．聴取実験型の主観評価は，被験者が呈示音声を正しく聴き取れた正答率を了解度とするため信頼性が高いが，実験実施に時間的・金銭的な制約が生じる．そこで，音声了解度予測指標 (speech intelligibility prediction metric) や客観了解度尺度 (objective intelligibility measure) と呼ばれるような，入力音声を信号処理ベースで分析して内部指標を算出し，非線形関数 を用いて了解度に変換する手法が古くから提案されている．これらの手法を用いることで，音声強調処理を適用した強調音声の了解度を客観評価の結果として見積もることができる．しかし，様々な非線形処理を施した強調音声の主観評価結果を総じて高い精度で予測できる指標はまだ確立されていない．

近年提案された音声了解度予測指標の中で，デファクト標準になりつつある指標が short-time objective intelligibility (STOI) measure である．STOI は，時間周波数 (T-F) 領域で非線形処理を施した音声の了解度を精度良く予測できることに加えて，非常にシンプルなアルゴリズム ([21] の論文紹介で解説) で設計されている．以上の理由から，Interspeech 2017 においても，多くの発表で音声強調処理手法の客観評価に STOI が使用されていた．

しかし， STOI による客観評価の結果も主観評価の結果と必ずしも対応しないことが報告されている．[22] の研究では，deep neural network (DNN) ベースの音声強調処理の客観評価を STOI で行った上で，その後実施した主観評価の結果と比較している．この研究で使用されていた音声強調処理手法では，DNN の入力特徴量として短時間フーリエ変換で得た 11 フレーム分の対数パワースペクトルを使用し，出力特徴量として 1 フレーム分の強調音声の対数パワースペクトルを得るシンプルなものである．結果として，雑音音声の信号対雑音比 (SNR) が -14 dB ～ 4 dB の範囲のとき，強調音声を評価した STOI の数値は強調前 (雑音音声) よりも上昇した．これは，STOI による予測結果では強調処理によって音声了解度が改善されたと解釈できる．しかし，主観評価の結果では，強調音声の了解度が強調前よりも低下したことが，全ての被験者の結果において明らかになった．つまり，強調処理によって音声の明瞭性は劣化し， STOI の予測結果とは真逆の結果になったことを示している．また，音声品質の改善に用いられた DNN 出力における系列内変動 (global variance) の補償処理は，音声了解度の改善には効果がないことも述べられている．このような結果は，音声強調分野ではよくある事例であるが，客観評価だけではなく主観評価の結果も報告している本研究は印象に残った．

上記のような STOI の予測結果と主観評価結果の矛盾を，バンド周波数に対応する重み付け関数 (band importance functions; BIFs) を用いて改善しようと検討した研究が [21] である．オリジナルの STOI は，1/3 オクターブバンドを通過し，正規化・クリッピング処理された狭帯域音声 (強調音声とクリーン音声) の振幅包絡間の相関係数を短時間フレーム毎に計算する．最終的に，バンドのチャンネル数 × 時間フレーム数だけ得られる相関係数を，均一の BIF (全ての係数が 1) で平均化することで，一つの値（内部指標）を得る．この研究では，オリジナルの STOI で行われる平均化処理の前に BIF を導入することで，バンド毎の重要度を調整した内部指標を得ることができる．さらに，この BIF は主観評価データを持つ複数のデータセット (異なる雑音条件と信号処理条件) を使用して決定される．しかし，上記の手続きを経て得られた各条件での BIFs は，同じ音声データセット内では関係性が見られるものの，全ての条件で良い予測結果を示すものは無かった．結論として，オリジナルの STOI で使用されている均一の BIF が最適であることが述べられている．

では，今後どのような客観評価指標が主流になるのだろうか．その候補の一つとして考えられるものが， DNN ベースの音声認識器を応用した手法 [23] である．提案法の音声認識システムは Kaldi speech recognition toolkit をベースに構築されており，DNN の入力に 23 チャンネルのメルフィルタバンクで分析された対数メルスペクトル (11 フレーム分) を使用し，出力として 3 つ組み音素 (トライフォン) が得られる．結果として，提案法は主観評価の結果を二種類の雑音 (時間的に定常・非定常) 条件において高精度で予測することができた．しかし，学習時に同じ条件の雑音音声データを使用していることから，予測結果に関しては驚きが少ない．この研究の興味深い点は，音声認識器が入力のどの部分に注目して音素を区別しているのかを，layer-wise relevance propagation (LRP) と呼ばれる手法を用いて分析しているところである．LRP は DNN の出力データを逆伝搬させることにより，各層の出力と重みを元に貢献度を求めていく手法である．この研究の場合は，音素 (triphone) の決定において，入力 (対数メルスペクトルの時系列データ) のどの部分がどれくらいの貢献度を示していたかを知ることができる．実際に，論文中の Fig. 3

において入力音声の対数メルスペクトログラムと LRP によって得られた貢献度の T-F マップを比較することができる．ただし，現状では，未学習のデータ（例えば新規の強調音声）が入力された場合の予測精度に関しては不明である．

このように，音声の明瞭性を客観的に評価することは，STOI のような単純な枠組みでは不可能であることが既に報告されている．そのために，人間の聴覚特性に基づいた予測指標 (例えば著者らの [24]) を設計することも一つの手であるが，深層学習分野で提案されている分析手法を活用することによって，指標が音声のどのような部分に注目して明瞭性を評価しているのかを理解することも重要である．加えて，今回の Interspeech 2017 の前日に開催された 1st International Workshop on Challenges in Hearing Assistive Technology (CHAT-2017) でも音声の明瞭性に関する研究成果が多数報告されていたので，公式サイト[*1]で配布されているプロシーディングスを参照されたい (山本)．

4.2 Speech analysis

音声処理において，もはや Deep neural network (DNN) は主流なツールといえる．従来のテキスト音声合成では，STRAIGHT や WORLD などのツールを用いて音声からパラメータを取り出して学習し，テキストからパラメータを出力して上記のツールで合成する仕組みを採用してきた．2016 年には，パラメータを使わずに波形を直接生成する WaveNet の発表があり，テキスト音声合成において音声パラメータは必須とは言い難い存在となった．また，パラメータ推定においても DNN ベースのものが提案されるようになりつつあることから，今後は音声分析技術そのものが不要になるかもしれない．そのような状況ではあるが，学習データが不要な信号処理によりパラメータ推定を行う研究は，学習データに非依存な結果が得られるメリットがあるため，現在も目的に応じた新たな手法が提案されている．本節では，「Speech and Harmonic Analysis」セッションから 2 つの論文を紹介する．

1 つは，音声波形からの基本周波数（F0）を推定する方法 adYANGsaf を提案した論文 [25] である．この方法は，YANGsaf という方法を改良したもので，他手法と比較して静かな環境での高い推定精度と，低 SNR 環境における頑健性を両立した優れた方法である．多くの F0 推定法はフレーム単位で F0 を求めるが，この方法では，1 回各フレームについて F0 を求めた後に，Adaptive Kalman Filter により処理することで，F0 の「軌跡」を求めることに重きを置いた方法である．Adaptive Kalman Filter におけるパラメータの最適化まで行っており，加えて， 1 回求めた結果を活用して 2, 3 回反復処理により結果を収束させることで，より信頼できる F0 軌跡を求めることを目指している．音声データベースを用いた評価により，雑音が無い場合での正確な推定と，雑音が存在する場合におけるミスの低減が両立できることを示している．反復処理が入るため実時間処理は不可能であるが，統計的音声合成などでは学習時の実時間処理は必須ではないため，精度を追求した本手法には高い価値があるといえる．

もう 1 つは，音声波形から声帯振動が生じた時刻を推定する（これを Epoch extraction と呼ぶ）方法を提案した論文 [26] である．こちらもベースとなる方法が存在し，それの性能を改善するアプローチである．核となるアイディアは，Zero frequency filtering (ZFF) と呼ばれる，0 Hz にピークがある特殊なフィルタと移動平均フィルタを組み合わせた処理である．本論文では，後者である移動平均フィルタを adaptive に設計し，主に怒りの感情音声と低 SNR 音声に対して頑健に動作することを示している．

どちらの論文で紹介した方法も，信号処理の理論の複雑さとしては基礎的な範囲にとどまっている．むしろ，近年提案された実音声の分析に優れた方法は，アイディアの源泉が歴史的に古いところにあることも多く，温故知新の流れがあるように感じる．本節で紹介した方法も，ベースとなるアイディアに，提案された当時の計算機能力では実現できなかったリッチな処理（反復処理や並列処理）を加えることで性能が飛躍的に改善する事例と解釈できる．奇抜な理論ではなく，一見古い理論の中に高性能を達成する金脈があると考えると，学習データを使わないシンプルな信号処理による音声分析技術は，まだまだ廃れないと感じるところである（森勢）．

4.3 Speech production

次に、Electro-Magnetic Articulograph (EMA) 等により取得された、舌・唇等の調音器官が発話中にどのように動いているかという情報 “調音運動” に関する研究を報告する．調音運動を取得するには特殊なセンサーもしくは装置が必要であることから、通常の音声波形ほどの大規模データを収集することが容易ではなく、これまで Deep neural network (DNN) の恩恵を大いに受けてきたとは言い難い状況であった．

しかし、Interspeech2017 では、この調音運動データと DNN を組み合わせる研究が幾つか見受けられた．まず、ここでは、その中の二つを紹介する．1 つは、音声波形から調音運動を推定する調音推定というタスクにおいて、不特定話者 DNN が学習可能か、学習された不特定話者 DNN は他の言語でも利用可能か検証した論文 [27] である．22 名のイギリス英語話者、21 名のオランダ語話者の調音運動データを利用し、MFCC から調音運動を推定する 3 層の DNN を言語別に学習したところ、相関係数 $r = 0.53$ でこれらの話者の調音運動が予測可能になったとのことであ

[*1] http://spandh.dcs.shef.ac.uk/chat2017/

る。また、イギリス英語話者のデータから学習した DNN を元に、オランダ語話者の調音運動を推定する、もしくはその逆のパターンを評価したところ、相関係数は予想通り多少劣化したが、それでも $r = 0.43$ の精度を示したとのことである。これまで言語非依存の調音推定実験はあまり研究されていないことから、この結果は興味深い内容である。2 つ目の論文 [28] も、不特定話者に対して DNN を利用し調音推定する実験結果を報告している．こちらの論文では不特定 DNN をよりよく学習する Multi-task learning や正則化手法の有効性を主に評価している。

調音運動はパーキンソン病などによりもたらされる構音障害などを音声から判別する際にも重要である。論文 [29] では、パーキンソン病患者か健常者かを音声から判定する CNN を提案している。パーキンソン病患者は有声から無声へ、もしくは無声から有声へ遷移する際に発音の障害が起きやすいという。そこで、この有声・無声クラスが切り替わる領域の短時間フーリエ変換もしくは連続ウェーブレット変換された音声信号を入力とし、convolution と max-pooling を繰り返し行い、パーキンソン病患者か健常者か区別する CNN を構築した。50 名のコロンビア人のパーキンソン病患者、88 名のドイツ人のパーキンソン病患者、20 人のチェコ人のパーキンソン病患者で検証を行ったところ、それぞれの言語において、約 8 割程度の精度で正しくパーキンソン病患者と健常者とを識別できたとの報告があった。また、同じ実験を、読み上げ音声、自発音声、言語聴覚士などが利用する/pa-ta-ka/という音声で比較した結果も報告されている。それによると、興味深いことに、上記 CNN ではどのタイプの音声でも同じ精度の識別ができたとのことである。（山岸）

5. 音声合成

音声変換と音声合成に関するセッションは，オーラルが 5 つ，ポスター 2 つで構成されており，うち音声変換に関するセッションはオーラルとポスターでそれぞれ 1 つであった．各オーラルセッションでは，深層学習に基づく最新手法や統計的パラメトリック音声合成，韻律表現，音声合成の評価に関する話題が取り扱われた．以下，著者らの注目する発表を紹介する．

5.1 音声合成の表現性向上：ポーズ推定

スマートスピーカーなど音声合成に普及に伴い，文単位ではなく段落単位などの長い音声の合成への需要が高まっている．長い音声を合成する際の課題の一つに適切なポーズ位置の挿入がある．文献 [30] において，著者らはポーズの発生する呼気段落境界をテキストから予測する手法を，オーディオブック音声を用いて評価した．提案手法では句読点や形態素，word2vec，係り受けなどの情報を入力とした LSTM-RNN でモデル化している．実験では，決定木による手法に比べ自然なポーズ予測が行えることを示した．

文献 [31] では句境界のテキストからの自動予測手法が提案された．著者らは，形態素解析や係り受け解析などで必要となる，人手によるアノテーションを最小化することを目的としている．そこでこの研究では，先述の文献 [30] のような形態素や係り受け情報の代わりに，言語コーパスを用いて学習した「コンマ予測器」の中間表現を，RNN の入力とする手法を提案している．また，提案法では句境界の有無だけでなく，ポーズの無音区間の長さを補助タスクとして学習する．主観評価では，提案法によってフレージング（ポーズ，リズム，流れ）のスコアが上昇することを示した．

5.2 声質変換

文献 [32] ではコンピュータ・ヴィジョンの分野で近年提案された画像のスタイル変換から着想を得て，Siamese Autoencoder (SAE) を用いた声質変換・話者識別手法が提案された ．SAE は，隠れ層をスタイル (話者) ユニット $\boldsymbol{h}_s$ とコンテンツ (音韻) ユニット $\boldsymbol{h}_c$ の 2 つに分割した Autoencoder (AE) を複数用意して，各 AE ごとにパラメータ共有させたネットワーク群である．SAE の目的関数は $L = L_r + L_c + L_s$ で定義され，AE の再構築エラー L_r，コンテンツエラー（異なる AE の $\boldsymbol{h}_c$ の非類似度など）L_c，及びスタイルエラー（異なる AE の $\boldsymbol{h}_s$ の類似度など）L_s を最小化するように学習される．2 話者で学習させた SAE の性能は従来の NN と同程度であったが，音響特徴量から自動的にスタイル・コンテンツといった，意味のある情報を抽出できたということが重要であると言える．

文献 [33] では Variational Autoencoder (VAE) と Wasserstein 距離関数に基づく Generative Adversarial Networks (W-GAN) を組み合わせ，高品質かつパラレルデータ不要な声質変換手法が提案された．従来から提案されていた，VAE を用いた非パラレル声質変換の評価関数に，W-GAN の評価関数を加え，VAE と W-GAN のネットワークを同時に最適化させている．また論文では，通常の GAN よりも，W-GAN の方が声質変換に適していることが示されている（W-GAN の主問題の計算にはパラレルコーパスを，W-GAN の双対問題には非パラレルコーパスを利用することが適している）．従来の VAE による非パラレル声質変換手法に対し，提案手法では，同性間の変換，異性間の変換ともに，MOS による主観評価基準で有意に改善することが報告されている．

5.3 WaveNet

WaveNet とは，波形を直接的に予測・生成するための，強力な自己回帰型畳み込みニューラルネットの一つである．テキスト音声合成 (Text-to-Speech; TTS) の波形生成器として用いられ，非常に高品質な音声の生成が可能で

あることが定量的に示されたことにより，近年多くの研究者の注目を集めている．今回の会議では “WaveNet and Novel Paradigms” というオーラルセッションが企画され，WaveNet の注目度の高さが伺える．

文献 [34] では話者依存型 WaveNet ボコーダが提案された．既存のボコーダを WaveNet に置き換えることで，音声生成過程のモデル化に伴う音質の劣化を避け，高品質かつ柔軟なボコーダの実現を目的とした試みである．CMU-ARCTIC データベースを用いた実験では，基本周波数とメルケプストラムを WaveNet の補助特徴量とし，メルケプストラムボコーダと比較して Signal-to-Noise Ratio が大きく改善されることが示された．これにより，WaveNet ボコーダは励振源の情報をより適切に捉え，位相情報を復元することができたと報告されている．この実験結果は，続く自然性に関する 5 段階 MOS 試験での WaveNet ボコーダの高評価値を裏付けるものと言える．本報告と関連して，文献 [35] では話者依存型 WaveNet ボコーダを GMM 声質変換における波形生成モジュールとして用いる手法が提案された．WaveNet ボコーダによる高品質な波形生成により，話者性に関して変換精度が向上した実験結果が報告されている．

文献 [36] では深層学習に基づく歌声合成のための新しいネットワーク構造が提案された．入力がスペクトログラムと歌詞に対応した言語特徴量，出力がフレーム単位のボコーダの音響特徴量である点を除き，ネットワークの構成要素は WaveNet と共通している．高速化のためにネットワークの規模はオリジナルの WaveNet よりも小さく抑えられている．さらに独自の工夫として，各音響特徴量ごとにネットワークを構築しており，スペクトル特徴量と基本周波数に関しては mixture density output を採用している点が挙げられる．特に後者により，出力分布の形状を効率よく柔軟に表現できることから，結果的にパラメータ数の増加を抑えることに貢献している．客観・主観評価実験により，従来の隠れマルコフモデルに基づく歌声合成および波形接続型の歌声合成手法と比較して，提案手法が高品質な歌声を合成できることが示された．

5.4 音声波形の直接生成

統計的パラメトリック型 TTS では，その波形生成部において，ボコーダの適用による合成音声の品質劣化が問題となっていた．文献 [37] では，source-filter モデルや harmonics plus noise モデルを仮定せず，直接音声の FFT スペクトルを表現することを検討している．ここでは複素数である FFT スペクトルをそのまま用いるのでなく，Fundamental Frequency，Magnitude Spectrum，Real Spectrum, Imaginary Spectrum の実数値として扱うことを提案している．提案法により，ボコーダ音に含まれるバジー感のない音声合成が実現されている．

5.5 言語特徴量の埋め込み

深層学習に基づく TTS を実現するためには，離散的な言語特徴量 (end-to-end 音声合成の場合はテキスト情報) を連続的な特徴量へと変換し（連続特徴量空間への埋め込み），その時間解像度を入力特徴量とマッチさせる必要がある．そこで文献 [38] では，階層的な encoder-decoder モデルを用いた言語特徴量の埋め込み技術が提案された．提案モデルでは，語・シラブル・音素に対応する各言語特徴量は個別の encoder モデルを通って変換されるが，各 encoder モデルが時間解像度の順に階層的に接続されており，階層的な言語特徴量の関係性を考慮しながら，モデルの中で時間解像度のマッチングを図る点がユニークである．decoder モデルの出力はフレーム単位でのボコーダの音響特徴量である．主観評価実験の結果より，LSTM-RNN に基づく統計的パラメトリック型 TTS と比較して高品質な音声を合成できることが示された．

5.6 音素継続長のモデル化

TTS では，音素継続長のモデル化も品質に関わる重要な要素である．文献 [39] では，音素継続長をガウス分布ではなく離散分布を用いモデル化することを検討している．WaveNet 等の音声波形のモデル化する手法においては，また，音声波形をそのまま用いず量子化し高い性能を示していることから，音素継続長のモデル化においても離散分布を用いることを検討することは興味深い．LSTM-RNN に基づく継続長モデルの構築が行われた．よりロバストな音素継続長のモデル化が実現されている．

5.7 End-to-end 音声合成

深層学習に基づく end-to-end 音声合成とは，大量の音声データとそれに対応する書き起こし文の対のみから，TTS を実現するニューラルネットワークを構築する技術である．Wang らは，encoder-decoder モデルと attention に基づく end-to-end 音声合成モデルを提案した [40]．テキスト系列は encoder により連続特徴量系列へと変換され，文脈情報を考慮するための attention も合わせて計算される．そして attention と連続特徴量系列から，decoder モデルが音声のスペクトログラム系列を出力する．最終的に，Griffin-Lim アルゴリズムによって位相成分を復元することで合成音声を得る．アメリカ英語の評価セットを用いて 5 段階 MOS 試験により自然性を評価した結果，提案手法の評価値は 3.82 であり，LSTM-RNN に基づく統計的パラメトリック型 TTS の評価値 3.62 を上回った．一方で，波形接続型 TTS の評価値 4.09 を下回り，音質の改善には課題が残った．今後の発展が期待される技術である（高木，中鹿，郡山，玉森）．

参考文献

[1] Stolcke, A. and Droppo, J.: Comparing Human and Machine Errors in Conversational Speech Transcription, *Proc. Interspeech 2017*, pp. 137–141 (online), DOI: 10.21437/Interspeech.2017-1544 (2017).
[2] Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., Cui, X., Ramabhadran, B., Picheny, M., Lim, L.-L., Roomi, B. and Hall, P.: English Conversational Telephone Speech Recognition by Humans and Machines, *Proc. Interspeech 2017*, pp. 132–136 (online), DOI: 10.21437/Interspeech.2017-405 (2017).
[3] Kurata, G., Sethy, A., Ramabhadran, B. and Saon, G.: Empirical Exploration of Novel Architectures and Objectives for Language Models, *Proc. Interspeech 2017*, pp. 279–283 (online), DOI: 10.21437/Interspeech.2017-723 (2017).
[4] Prabhavalkar, R., Rao, K., Sainath, T. N., Li, B., Johnson, L. and Jaitly, N.: A comparison of sequence-to-sequence models for speech recognition, *Proc. Interspeech*, pp. 939–943 (2017).
[5] Audhkhasi, K., Ramabhadran, B., Saon, G., Picheny, M. and Nahamoo, D.: Direct Acoustics-to-Word Models for English Conversational Speech Recognition, *Proc. Interspeech*, pp. 959–963 (2017).
[6] Soltau, H., Liao, H. and Sak, H.: Neural speech recognizer: Acoustic-to-word LSTM model for large vocabulary speech recognition, *arXiv preprint arXiv:1610.09975* (2016).
[7] Hori, T., Watanabe, S., Zhang, Y. and Chan, W.: Advances in Joint CTC-Attention based End-to-End Speech Recognition with a Deep CNN Encoder and RNN-LM, *Proc. Interspeech*, pp. 949–953 (2017).
[8] Li, J., Seltzer, M., Wang, X., Zhao, R. and Gong, Y.: Large-Scale Domain Adaptation via Teacher-Student Learning, *Proc. Interspeech*, pp. 2386–2390 (2017).
[9] Wong, J. and Gales, M.: Student-teacher training with diverse decision tree ensembles, *Proc. Interspeech*, pp. 117–121 (2017).
[10] Fukuda, T., Suzuki, M., Kurata, G., Cui, J., Thomas, S. and Ramabhadran, B.: Efficient knowledge distillation from an ensemble of teachers, *Proc. Interspeech*, pp. 3697–3701 (2017).
[11] Deena, S., Ng, R. W. M., Madhyastha, P., Specia, L. and Hain, T.: Semi-supervised Adaptation of RNNLMs by Fine-tuning with Domain-specific Auxiliary Features, *Proc. Interspeech*, pp. 2715–2719 (2017).
[12] Ma, M., Nirschl, M., Biadsy, F. and Kumar, S.: Approaches for Neural-Network Language Model Adaptation, *Proc. Interspeech*, pp. 259–263 (2017).
[13] Singh, M., Oualil, Y. and Klakow, D.: Approximated and domain-adapted LSTM language models for first-pass decoding in speech recognition, *Proc. Interspeech*, pp. 2720–2724 (2017).
[14] Sadjadi, S. O., Kheyrkhah, T., Tong, A., Greenberg, C., Reynolds, D., Singer, E., Mason, L. and Hernandez-Cordero, J.: The 2016 NIST Speaker Recognition Evaluation, *Proc. Interspeech 2017*, pp. 1353–1357 (2017).
[15] Colibro, D., Vair, C., Dalmasso, E., Farrell, K., Karvitsky, G., Cumani, S. and Laface, P.: Nuance - Politecnico di Torino's 2016 NIST Speaker Recognition Evaluation System, *Proc. Interspeech 2017*, pp. 1338–1342 (2017).
[16] Snyder, D., Garcia-Romero, D., Povey, D. and Khudanpur, S.: Deep Neural Network Embeddings for Text-Independent Speaker Verification, *Proc. Interspeech 2017*, pp. 999–1003 (2017).
[17] ASVspoof 2017 Challenge: http://www.asvspoof.org/.
[18] Kinnunen, T., Sahidullah, M., Delgado, H., Todisco, M., Evans, N. and and Kong Aik Lee, J. Y.: The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection, *Proc. Interspeech 2017*, pp. 2–6 (2017).
[19] Lavrentyeva, G., Novoselov, S., Malykh, E., Kozlov, A., Kudashev, O. and Shchemelinin, V.: Audio Replay Attack Detection with Deep Learning Frameworks, *Proc. Interspeech 2017*, pp. 82–86 (online), available from ⟨http://dx.doi.org/10.21437/Interspeech.2017-360⟩ (2017).
[20] Ji, Z., Li, Z.-Y., Li, P., An, M., Gao, S., Wu, D. and Zhao, F.: Ensemble Learning for Countermeasure of Audio Replay Spoofing Attack in ASVspoof2017, *Proc. Interspeech 2017*, pp. 87–91 (online), available from ⟨http://dx.doi.org/10.21437/Interspeech.2017-1246⟩ (2017).
[21] Andersen, A. H., de Haan, J. M., Tan, Z.-H. and Jensen, J.: On the Use of Band Importance Weighting in the Short-Time Objective Intelligibility Measure, *Proc. Interspeech 2017*, pp. 2963–2967 (online), DOI: 10.21437/Interspeech.2017-1043 (2017).
[22] Gelderblom, F. B., Tronstad, T. V. and Viggen, E. M.: Subjective Intelligibility of Deep Neural Network-Based Speech Enhancement, *Proc. Interspeech 2017*, pp. 1968–1972 (online), DOI: 10.21437/Interspeech.2017-1041 (2017).
[23] Spille, C. and Meyer, B. T.: Listening in the Dips: Comparing Relevant Features for Speech Recognition in Humans and Machines, *Proc. Interspeech 2017*, pp. 2968–2972 (online), DOI: 10.21437/Interspeech.2017-1168 (2017).
[24] Yamamoto, K., Irino, T., Matsui, T., Araki, S., Kinoshita, K. and Nakatani, T.: Predicting Speech Intelligibility Using a Gammachirp Envelope Distortion Index Based on the Signal-to-Distortion Ratio, *Proc. Interspeech 2017*, pp. 2949–2953 (online), DOI: 10.21437/Interspeech.2017-170 (2017).
[25] Hua, K.: Improving YANGsaf F0 Estimator with Adaptive Kalman Filter, *Proc. Interspeech 2017*, pp. 2301–2305 (2017).
[26] Gangamohan, P. and Yegnanarayana, B.: A Robust and Alternative Approach to Zero Frequency Filtering Method for Epoch Extraction, *Proc. Interspeech 2017*, pp. 2297–2300 (2017).
[27] Sivaraman, G., Espy-Wilson, C. and Wieling, M.: Analysis of acoustic-to-articulatory speech inversion across different accents and languages, *Proc. Interspeech 2017*, pp. 974–978 (2017).
[28] Badino, L., Franceschi, L., Arora, R., Donini, M. and Pontil, M.: A Speaker Adaptive DNN Training Approach for Speaker-Independent Acoustic Inversion, *Proc. Interspeech 2017*, pp. 984–988 (2017).
[29] Vásquez-Correa, J., Orozco-Arroyave, J. R. and Nöth, E.: Convolutional Neural Network to Model Articulation Impairments in Patients with Parkinson's Disease, *Proc. Interspeech 2017*, pp. 314–318 (online), DOI: 10.21437/Interspeech.2017-1078 (2017).
[30] Klimkov, V., Nadolski, A., Moinet, A., Putrycz, B., Barra-Chicote, R., Merritt, T. and Drugman, T.: Phrase Break Prediction for Long-Form Reading TTS: Exploiting Text Structure Information, *Proc. Interspeech 2017*, pp. 1064–1068 (online), DOI: 10.21437/Interspeech.2017-419 (2017).

[31] Rendel, A., Fernandez, R., Kons, Z., Rosenberg, A., Hoory, R. and Ramabhadran, B.: Weakly-Supervised Phrase Assignment from Text in a Speech-Synthesis System Using Noisy Labels, *Proc. Interspeech 2017*, pp. 759–763 (online), DOI: 10.21437/Interspeech.2017-487 (2017).
[32] Mohammadi, S. H. and Kain, A.: Siamese Autoencoders for Speech Style Extraction and Switching Applied to Voice Identification and Conversion, *Proc. Interspeech 2017*, pp. 1293–1297 (online), DOI: 10.21437/Interspeech.2017-1434 (2017).
[33] Hsu, C.-C., Hwang, H.-T., Wu, Y.-C., Tsao, Y. and Wang, H.-M.: Voice Conversion from Unaligned Corpora Using Variational Autoencoding Wasserstein Generative Adversarial Networks, *Proc. Interspeech 2017*, pp. 3364–3368 (online), DOI: 10.21437/Interspeech.2017-63 (2017).
[34] Tamamori, A., Hayashi, T., Kobayashi, K., Takeda, K. and Toda, T.: Speaker-Dependent WaveNet Vocoder, *Proc. Interspeech 2017*, pp. 1118–1122 (online), DOI: 10.21437/Interspeech.2017-314 (2017).
[35] Kobayashi, K., Hayashi, T., Tamamori, A. and Toda, T.: Statistical Voice Conversion with WaveNet-Based Waveform Generation, *Proc. Interspeech 2017*, pp. 1138–1142 (online), DOI: 10.21437/Interspeech.2017-986 (2017).
[36] Blaauw, M. and Bonada, J.: A Neural Parametric Singing Synthesizer, *Proc. Interspeech 2017*, pp. 4001–4005 (online), DOI: 10.21437/Interspeech.2017-1420 (2017).
[37] Espic, F., Botinhao, C. V. and King, S.: Direct Modelling of Magnitude and Phase Spectra for Statistical Parametric Speech Synthesis, *Proc. Interspeech 2017*, pp. 1383–1387 (online), DOI: 10.21437/Interspeech.2017-1647 (2017).
[38] Ronanki, S., Watts, O. and King, S.: A Hierarchical Encoder-Decoder Model for Statistical Parametric Speech Synthesis, *Proc. Interspeech 2017*, pp. 1133–1137 (online), DOI: 10.21437/Interspeech.2017-628 (2017).
[39] Chen, B., Bian, T. and Yu, K.: Discrete Duration Model for Speech Synthesis, *Proc. Interspeech 2017*, pp. 789–793 (online), DOI: 10.21437/Interspeech.2017-1144 (2017).
[40] Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R. and Saurous, R. A.: Tacotron: Towards End-to-End Speech Synthesis, *Proc. Interspeech 2017*, pp. 4006–4010 (online), DOI: 10.21437/Interspeech.2017-1452 (2017).