Vol.2019-SLP-126 No.10
2019/2/28

国際会議Interspeech2018報告

高木 信二[1]　安藤 厚志[2]　越智 景子[3]　沢田 慶[4]　塩田 さやか[5]　鈴木 雅之[6]　玉森 聡[7]　俵 直弘[8]
福田 隆[6]　増村 亮[2]

概要：2018 年 9 月 2 日から 9 月 6 日にかけ，ハイデラバード・インドで Interspeech2018 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．本項では，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

2018 年 9 月 2 日から 9 月 6 日にかけ，ハイデラバード・インドで Interspeech2018 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．1,668 件の投稿があり，749 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，音声強調・音源分離，音声認識，話者認識，感情認識，音声合成，医療・支援技術を中心として報告する．

2. 音声強調・音源分離

音声強調・音源分離では，音源分離が 4 セッション，音声強調が 2 セッションの計 6 セッション，57 件の発表が行われた．ここではモノラルおよびマルチチャネル音声強調について，深層学習に基づく手法を中心に研究発表論文を紹介する．

モノラル信号を対象とした音声強調法，特に音源分離の分野においては，Ideal binary mask や Ideal relation mask 等の時間周波数マスクを CNN や LSTM などにより推定する手法が多く提案されている．DNN で時間周波数マスクを推定する場合，各周波数成分がどの音源に属するかが定まらない label permutation の問題が発生する．この問題を解決するために Permutation invariant training (PIT) や deep clustering などがよく用いられるが，今回もこれら手法を拡張した研究が注目を集めていた．例えば，[1] では Deep clustering の拡張である Chimera++ network を音源分離に適用している．原型となる Deep clustering では時間周波数軸上で，同一話者発話の周波数成分は同士は近くに，それ以外の周波数成分同士は遠くに射影される非線形空間を DNN により学習し，この空間上で k-means クラスタリングを行うことでマスクを作成する．一方，Chimera++では，Deep clustering により得られた空間上で phase-sensitive mask (PSM) を推定することで，特徴量抽出とマスク推定を同時に行う．本研究では Chinema++ で推定したマスクにより得られた分離信号の振幅スペクトルに対し，混合音声の位相を初期値として，時間領域への変換と位相修正を繰り返し行うことで分離音声の位相を推定するネットワークを構築し，Chinema++と共に end-to-end で最適化することで高い品質の分離音声を生成できることが示されている．

[1] でも言及されているように，これまでの多くの研究では強調音声の振幅スペクトルの推定が関心の対象で，位相に対しては混合音声のものをそのまま使う手法が中心であったのに対し，近年では強調音声の位相も同時に推定する研究が注目を集めている．しかし，位相スペクトルは，周期性や隣接するフレームにおける無矛盾性などの制約を満たさなければならないため，振幅スペクトルに比べて推定が困難であることが知られている．そこで，[2] では位相スペクトルを直接推定する代わりに，位相を複数の範囲に量子化し，各位相クラスへの割当を推定する識別問題として位相を推定する PhaseNet が提案された．PhaseNet により分離された信号は SDR と音声認識の単語誤り率で評価され，8 程度の量子化数で十分な品質が得られることが示されている．

更に，位相推定の問題を回避するため，時間周波数領域ではなく時間領域で強調音声の波形推定を行う DNN が複

1 国立情報学研究所
2 日本電信電話株式会社
3 東京工科大学
4 マイクロソフトディベロップメント株式会社
5 首都大学東京
6 日本アイ・ビー・エム株式会社
7 愛知工業大学
8 早稲田大学

数されている．例えば，[3] では 1 次元畳み込み LSTM を用いて，残響環境下音声からクリーン音声へのマッピングを波形レベルで推定する手法が提案されている．他にも，[4] では短時間（逆）フーリエ変換が行列で表記できることを用いて，波形レベルでの復元誤差と時間周波数領域上でのマスク推定に用いる手法が提案されてる．提案法は非負値行列分解に基づくモノラル音声強調法と比較され，同程度の明瞭性を維持しつつ，音質の改善に成功している．

また，Interspeech2017 で提案された敵対的生成学習 (generative adversarial training; GAT) に基づく音声強調法を皮切りに，従来の DNN に基づく音声強調の正則化項として，新たな目的関数を導入する手法が近年多く提案されており，今回もいくつか発表があった．例えば [5] では，混合音声からクリーン音声へのマッピングを推定すると同時に，クリーン音声から混合音声へのマッピングを同時に行う cycle-consistency を目的関数として導入することで，単純に混合音声からクリーン音声への一方向のマッピングを学習する場合よりも，低い単語誤り率が達成できる強調音声を生成できることが示されている．他にも LSTM に基づく Ideal ratio mask ベースの手法に対して GAT を適用する手法 [6] や，畳み込み LSTM に基づくスペクトルマッピングベースの手法に GAT を適用することで，残響除去においても GAT が有効であることを示した研究 [7] などがあった．

マルチチャネル音声強調・分離の研究では，従来の MVDR やウィナービームフォーマと DNN によるマスク推定を組み合わせた枠組みが多く提案されていた．例えば，チャネルごとに単チャネル LSTM による Phase-sensitive mask (PSM) を推定し，得られた強調音声で MVDR beamformer を構築する手法 [8] や，構築した MVDR beamformer の出力と各チャネルの時間差情報を用いて，マルチチャネル LSTM による PSM をさらに推定する手法などが提案されている [9]．さらに大規模な手法として，[10] で提案された手法では，マルチチャネル信号の中からランダムに選択された 2 つのマイクロフォンペアを用いて 2 チャンネル LSTM により PSM 推定を行い multi-channel Wiener filter (MCWF) を目標信号ごとに構築する．各マイクロフォンペアで構築した MCWF により強調した信号は，各チャネルの振幅スペクトルと共に後段のマルチチャネル LSTM に入力され，最終的な PSM が構築される．このとき上記枠組みは一つのネットワークで表現でき誤差逆伝搬法により end-to-end で学習可能で．チャネルごとに PSM を推定し，beamformer を構築する手法や，マルチチャネル Deep clustering に基づく手法よりも高い SDR が得られることが示されている（俵）．

3. 音声認識

3.1 オンライン End-to-end 音声認識

End-to-end（E2E）音声認識の研究はここ数年で急加速したが，最近はオンライン音声認識での可動性を意識した研究例が増えており，今回の会議でもいくつか関連発表があった．ここではオンライン処理に関するものを中心に研究発表論文を紹介する．

双方向（bidirectional）LSTM は様々な条件で高い認識性能をもたらす可能性が示唆され，E2E 音声認識の枠組みでも多くの研究が行われている．高性能を実現するためには，パラメータチューニングやネットワークの初期化戦略などが重要であり，特にこの性質は単方向（unidirectional）LSTM のようなオンライン向けネットワークの学習で顕著であった．効果的な方法の一つは，triphone モデルなどの既存のネットワークを初期値として学習を始める方法であるが，この戦略は発音辞書が不要であるという E2E システムの利点を損なってしまう．文献 [11] では，知識蒸留（knowledge distillation）学習の枠組みを利用して，十分に学習された高精度なオフライン向け bidirectional LSTM から unidirectional LSTM に知識を転移する学習方法を提案している．この方法では，教師ネットワークである bidirectional LSTM と生徒ネットワーク側である unidirectional LSTM のどちらについても発音辞書を用いずに学習が進められるため，E2E 音声認識学習の利点を損なわずに unidirectional LSTM の性能を改善することができる．実験ではカリキュラムラーニングとラベルスムージングも併用した比較を行っており，ランダムな初期化に基づく簡便な方法と比較して 19%の改善を達成したことを報告している．

発音辞書や（理想的には）言語モデルを必要としない E2E 音声認識システムは，HMM とのハイブリッドシステムに比べて低リソースで実現できるため，多くの計算機資源を割くことができない組み込み型のシステムで大きな効果を発揮する．しかし，モデルのさらなるコンパクト化にはどうしても認識精度が犠牲になってしまう問題があった．文献 [12] では行列分解処理や知識蒸留学習，ネットワークパラメータの削減処理などを検討し，また，各種手法の組み合わせとパラメータサイズを考慮して最も効果的な方法を模索している．各手法の単独利用でも有意な性能改善を示しているが，それぞれを組み合わせることによって性能をさらに引き上げることができると結論づけている．

他方，文献 [13] では E2E システムにおける事前学習の一つを提案している．具体的には，マックスプーリング処理の窓長を変えながら layerwise に初期化を進め，段階的に LSTM 層を追加することによって，最終的な認識性能の改善につなげている．また，ネットワーク学習が収束す

るまでの学習時間についても言及がある．著者らは 1000 時間の LibriSpeech タスクにおいて，dev-clean で 3.54%，test-clean で 3.82%の最高水準の性能を実現したことを報告している．文献 [13] では，直接的にオンライン音声認識での実験等は行っていないが，オンラインを対象とした E2E 音声認識にも効果が期待できる方法として紹介させて頂いた．

3.2 音声認識モデルの適応

テキストデータのみを使った音声認識モデルの適応は、実用的に非常に重要なタスクである。言語モデルを明示的に利用する場合には、ドメイン適応先のテキストデータを用いて言語モデルのみを適応することで、比較的簡単にドメイン適応を実現することができる。しかし、音声認識モデルとして E2E モデルを用いる場合、テキストデータを使ったモデル適応が存在しないという欠点があった。Interspeech 2018 では、この問題を解決するための方法が、いくつか発表された。文献 [14] では、予め学習した言語モデルを組み込んだ E2E モデル を構築することで、適応時に、言語モデルに相当する部分のみをテキストデータで適応すること可能にする手法を提案している。文献 [15] では、text-to-speech モデルを使ってドメイン適応先の音声を合成することで、適応を可能にする手法を提案している。

他にも、音声認識のモデル適応には数多くの興味深い発表があった。文献 [16] では、ニューラルネットを使った言語モデルの適応において、精度を改善する正則化手法を提案している。文献 [17] では、補助的な入力を受け付けるモデルを構築することで、E2E モデルの適応する手法を提案している。文献 [18] では、因子分析されたレイヤーを用いた LSTM 音響モデルのドメイン適応が効率的に行えることを示している。文献 [19] では、batch normalization を含むモデルの、再学習による適応の性能を向上させる手法を提案している。

3.3 音声の埋め込みベクトル化と応用

自然言語処理や動画像処理の分野と同様に，音声分野においてもデータを固定長のベクトル表現に埋め込む検討が増えてきている．特に Interspeech2018 では，大量のラベルなし音声データからベクトル埋め込みの機構を獲得する検討やその応用が注目を集めた．

文献 [20] では，音声データを固定長のベクトル表現に埋め込む方法として，自然言語処理分野における Continuous Bag-of-Words や Skip-gram と類似した考え方に基づく手法を提案している．この研究では，音声中の単語境界を既知として，ある単語に対応するセグメントの音声から前後のセグメントの音声を予測するモデルを学習することにより，音声の意味的な成分を固定長のベクトル表現に埋め込む仕組みを獲得している．類似表現を検索する実験において，音声から学習したベクトル表現は，テキストから学習したものよりも高い検索性能を示しており，音声に含まれるリッチな情報を考慮した埋め込みが獲得できることが示唆されている．

文献 [21] でも類似した手法が提案されているが，文献 [20] と異なる点は，単語境界の情報を与えず学習を行う点が異なっている．この研究では，学習時に 0.5 秒の固定長セグメントごとに音声を分割し，2 つの音声セグメントが隣接する 2 セグメントか，そうでないかを予測するモデルを学習することにより，音声を固定長のベクトル表現に埋め込む機構を獲得している．実験では，音声認識の追加特徴量として有効であるかが評価されており，未知のドメインの音声を認識する際に有効であることが示されている．

文献 [22] では，系列識別モデルに基づくベクトル埋め込み手であるの Audio2Vec を利用して，音声と音素列の組を全く用いずに音素認識をモデル化する方法を提案している．この研究では，最初に大量の音声データを Audio2Vec を利用してベクトル系列に変換し，そのベクトル空間をクラスタリングすることで，音素と類似した意味を持つ離散ラベル表現を獲得しておく．次に，そのラベル表現と音素シンボルを紐づけるための変換モデルをモデル化することにより音素認識を実現する．ラベル表現と音素を紐づけための変換モデルの学習にもポイントがあり，音声と音素列の組は全く用いずに，Generative Adversarial Network に基づき，本当に存在する音素系列かそうでないかを見分けるネットワークを欺けるように学習することで，変換モデルを獲得している．実験では，35%程度の音素識別率のモデリングが可能であることを示しており，ラベルなし音声データを利用する方法として興味深い検討と言える．

4. 話者認識

話者認識では，話者照合 4 セッション，話者ダイアライゼーション 2 セッション，詐称者検出 1 セッションの計 7 セッション，69 件の発表が行われた．

近年の技術トレンドは，可変長の発話から固定長の話者表現を抽出する話者表現抽出を高度化するための取組みである．特に，深層学習に基づく話者表現抽出は因子分析モデルに基づく話者表現 (i-vector) と同等以上の発表件数に増加しており，非常に多くの注目を集めた．

文献 [23] では，現在の深層学習に基づく話者表現抽出のデファクトである x-vector に self attention を導入し，発話の一部区間に話者性が強く表れる場合において高精度に話者表現抽出を行う手法が提案された．このとき，互いが異なるよう制約を加えた複数の注機機構を導入することで照合誤りがさらに減少することが示されている．話者表現抽出に self attention を導入するアプローチは文献 [24] でも報告されている．

文献 [25] では，angular softmax を損失関数に導入し，

超球面空間に話者表現を埋め込む手法が提案された．本手法は超球面空間において 2 つの話者表現ベクトルがなす角度を遠ざける/近づけるように話者表現抽出を学習するため，話者表現の類似度評価にコサイン類似度を用いる場合に適している．実験では，一般的な softmax による損失関数に比べて約 20%と大幅な誤り削減を達成し，また i-vector/PLDA と比べても約 25%の誤り削減を示した．また同文献では，話者内変動の低減のため，発話ごとの話者表現とそれらの話者平均とのユークリッド距離を損失関数に用いる center loss が用いられた．center loss は顔認証において提案された手法であるが，話者照合でも有効性を示し [26]，また発話長の変動に対する頑健性向上が可能となることが明らかとなった [27].

文献 [28] では，Triplet に基づく話者表現抽出モデルの学習において，話者表現を条件に用いる Conditional Generative Adversarial Network (CGAN) を連結し同時最適化する手法を提案した．このとき，CGAN 部は Discriminator loss だけでなく話者分類を行う softmax loss を用いてマルチタスク学習を行うことで，自然発話に近づけつつ話者性も表現されている音声が生成されることを狙っている．本手法は triplet loss のみに基づく話者表現抽出に比べて誤り削減率 30%以上という非常に高い有効性を示しており，話者表現抽出において生成モデルの観点を加えることで話者表現の表現力向上や頑健性向上に繋がることが示唆された．

2015 年から始まった話者照合に対するなりすまし攻撃検出のコンペティションである ASVspoof の流れを汲んで Interspeech2018 でもなりすまし検出に関する論文が多く発表された．とくに 2017 年に開催された ASVspoof2017 では登録話者の声を録音再生する論理アクセス攻撃に焦点を当てられていたが，最も精度の高かった Lightly-CNN (LCNN) [29] に基づく手法においても EER が 6.73%と十分な精度ではなかったため，再生攻撃検出法に関する論文が多くを占めていた．技術的な傾向としては，どのような特徴量を用いるかを議論するものがほとんどであり，識別モデルとしては依然として GMM を学習し，対数尤度で評価するものとなっていた．その中でも，LCNN より低い EER を得た文献について紹介する．

文献 [30] では，再生音声と実発話の違いがスペクトログラムの低周波帯域と高周波帯域にそれぞれ現れることに着目し，低周波数帯域の分解能が高い Mel filterbank slope (MFS) および高周波数帯域の特徴も平等に取れる Linear filterbank slope (LFS) 両方の特徴量を抽出することを提案している．2 つの特徴量が必要となるのは，ASVspoof で用いたなりすまし収録機器の影響で，高スペックなマイクと低スペックなマイクでスペクトログラム特徴の出方が異なることに起因している．しかしながら，テスト時に収録機器の情報はわからないため，特徴量としては MFS および LFS，Constant Q cepstrum coefficient (CQCC)，MFCC の 4 種類を用い，盗聴量ごとに GMM を学習した後，最適な特徴量を選択する Dicision-level feature switching (DLFS) 仕組みを用いて特徴量選択を行っている．この手法において選択する特徴量が MFCC, LFS, MFS とした際の EER が 6.23%となっており，これまで発表された ASVspoof2017 の EER としては最も低くなっている．

文献 [31] も文献 [30] と同様に特徴量の抽出に着目した手法となっている．こちらは各発話の Moduration spectrum から静的および動的な特徴量である Modulation static energy cepstral coefficient (MSE-CC) と Modulation centroid frequency cepstral coefficient (MCF-CC) を抽出している．MSE-CC と MCF-CC は発話全体から得られる特徴量であるため，さらに，Short term cepstral coefficients (STCC) という Moduration spectrum からフレーム単位で抽出される特徴量も使用している．こちらも特徴量ごとに GMM を学習し，その組み合わせで評価を行っている．最終的にから MCF-CC, MSE-CC, STCC すべてを使うことで LCNN よりも低い EER である 6.32%を達成している．

文献 [32] では，フィルタバンクや特徴量抽出を工夫するのではなく，適切な特徴量を抽出するための Deep siamese ネットワークを用いて embedding による特徴量抽出を行うことを提案しいている．Siamese ネットワーク自体は既に署名検証や顔認証など様々な認証システムとして用いられているネットワーク構造で，ペアで入力されたデータが同じクラスに所属するかどうかを類似度から判定する枠組みとなっている．しかしながら，なりすまし検出に同じネットワークを利用できなかったため類似性を測る損失関数をさらに別のネットワークとして用意することで embedding された特徴量抽出を行っている．最終的な識別器としては GMM を使用しており，EER は 6.40%と LCNN よりも低くなっている．

5. 感情認識

感情認識は，6 セッション計 49 件の発表が行われた．このうち 2 セッションは INTERSPEECH 恒例のスペシャルセッションである Computational Paralinguistics ChallengE (ComParE) であり，本年のタスクは障がい者音声感情分類，自己評価感情 (快-不快) 分類，泣き声分類，心音分類であった．本節ではこれらのスペシャルセッションを除いた，一般的な感情認識に関する発表について報告する．

感情認識では昨年に引き続き，ニューラルネットワークを用いた技術が多数みられた．特に，信号波形やスペクトログラムを入力特徴とする手法は MFCC や基本周波数などのヒューリスティックな特徴量を用いる手法に比べて軒並み高い精度を示しており，感情認識でも raw data を用いる手法が流行する兆しを見せた．

文献 [33] では MFCC, パワースペクトル，信号波形の三

通りの入力特徴を比較評価し，信号波形が最も感情分類に効果的であることを示した．さらに，モデル構造の違いやデータ拡張の有無，学習データ発話長と感情認識精度との関係性を実験的に調査し，TDNN-LSTM-attention 構造が最も精度が高い点，話速とパワーのデータ拡張が効果的である点，学習データを可変長でなく固定長とすることで精度が向上する点などが報告された．感情認識の精度向上に向けて多くの示唆を与えており，非常に参考となる文献の一つと言える．

文献 [34] は音素系列を言語情報として利用する感情認識を提案した．感情認識に言語情報を用いるアプローチは文献 [35] も含め多数の従来研究が存在するが，単語でなく音素単位の情報でも感情分類精度を大きく改善することを示したという点で興味深いと言える．ただし現行の手法では書き起こしから音素系列を得ているため，音素認識結果を用いた際の有効性評価は今後の課題である．

感情認識モデル学習における教師ラベル付きデータの不足は感情認識における普遍的課題の一つであるが，この課題に対する発表もいくつかなされた．文献 [36] は Generative Adversarial Network (GAN) によるデータ拡張によって学習データ量を疑似的に増加させる手法を提案し，同一コーパス・クロスコーパス評価の両方で性能が改善することを示した．文献 [37] は過学習防止のための正則化に向け，入力データの再構成を教師なし補助タスクに追加する Ladder Network を利用し，感情回帰精度を向上させた．さらに上記手法はいずれも半教師あり学習に拡張可能であり，今後は教師ラベルなしデータを活用した感情認識の検討が進むと考えられる．

6. 音声合成

音声合成に関するセッションは計 6 セッションあり，57 件の発表が行われた．

6.1 発話スタイル

Text-to-speech (TTS) システムの品質は，DNN の導入により劇的に向上しており，日々の生活で合成された音声を耳にすることも多くなってきた．しかし，人間のような多様な発話スタイルを自在にコントロールして合成することは，まだまだ難しいタスクである．そのため，多様な発話スタイルを合成することを目指した TTS システムの研究・開発が行われており，Interspeech2018 でもいくつもの手法が提案された．

文献 [38] では，DNN に基づく音素継続長モデルと音響モデルの入力として，言語特徴量に加え低次元 (4 次元) の感情をコントロールするベクトルを用いてる．感情コントロールベクトルの各次元は，それぞれ喜び・悲しみ・怒り・平静を表している．音声データから感情を予測するために設計された extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) を特徴量とした感情予測モデルにより感情を予測し，感情コントロールベクトルとして利用している．

文献 [39] では，感情や韻律を効率よくモデル化する EMPHASIS が提案された．EMPHASIS では，CBHG (1-D convolution bank + highway network + bidirectional GRU) を音素継続長モデルと音響モデルとして用いている．CBHG の入力は音素に関する言語特徴量と感情・韻律に関する言語特徴量としており，それぞれ個別の 1 次元畳み込みフィルタを介し後段のネットワークにて統合がなされている．感情・韻律の特徴量は，音素に関する特徴量と比べ弱い特徴量であるが，個別の畳み込みフィルタを利用することで，感情・韻律の特徴量が考慮されにくくなる問題を緩和している．さらに，出力層においても音響特徴量の種類ごとに個別の bidirectional GRU 層を持つ構造により，合成音声の音質を改善している．

文献 [40] では，自己回帰モデルを導入した seq2seq モデル (VoiceLoop) を VAE に組み込んだ VAE-Loop が提案された．VAE により，学習データの潜在表現を潜在変数を用いてモデル化することができる．そして，VAE の潜在変数を VoiceLoop の入力とする構造により，潜在表現を考慮した音声合成を実現している．実験により，潜在変数の値を変更することで，話者性や発話スタイルがコントロールできることが示されている．

6.2 WaveNet ボコーダ

Interspeech2017 において，高品質な音声波形生成のための WaveNet ボコーダが提案された [41]．WaveNet ボコーダは音声波形をランダムサンプリングにより直接生成する自己回帰型のニューラルネットであり，音響特徴量を補助入力に取ることで波形生成をコントロールすることができる．本会議では引き続き WaveNet ボコーダを音声合成・音声変換タスクに活用する研究が発表された．複数話者への対応も検討されており，研究の方向性として自然である．特に，複数話者のデータから話者非依存な WaveNet を学習しておき，少量の単独話者データで適応させるケースが見られた [42], [43], [44].

文献 [42] で提案された「GlotNet」では，声門励起信号が WaveNet により直接生成され，それを線形自己回帰フィルタに通すことで音声波形が生成される．励起信号の生成過程は混合ロジスティック分布からのサンプリングによって実現されており，WaveNet は mixture density network として混合ロジスティック分布のパラメータを表現している．励起信号自体は話者依存性が低いため，十分な量のデータで GlotNet を事前に訓練しておけば，それをもとにして特定話者の少量のデータによる適応を効果的に実現することができる．音質および話者類似性に関する主観評価実験により，WaveNet ボコーダによる音声の直接生成と比

較して，GlotNet はそれと同等以上の性能を示す結果が得られている．

文献 [43] ではターゲット話者の少量の音声データを用いたロバストな音声変換モデルの構築を検討している．提案手法においては双方向 LSTM-RNN を特徴量変換モデルとして採用している [45]．変換元話者の音声から抽出された MFCC は，話者非依存の音声認識器を用いて言語特徴量（phonetic posteriorgrams (PPG)，音素事後確率情報）に変換される．続いて PPG が上記 RNN に入力され，ターゲット話者の音響特徴量（メルケプストラムと基本周波数）が出力される．PPG には話者非依存の言語的コンテキスト情報が含まれているため，話者非依存の WaveNet ボコーダ構築にも効果的に働くことが期待される．提案手法は，声質変換手法の性能を競うコンテスト VCC2018 において，自然性に関する 5 段階 MOS 値と話者類似性に関する preference score の総合で第 1 位を達成している．特に MOS 値は 4.13 であり，顕著な有効性が示されている．

文献 [44] では，非負値行列因子分解に基づく変換手法が提案されている．話者ペアのパラレルデータから，事例ベースの「辞書」（スペクトルからなる行列）があらかじめ作成される．変換時にはソース話者のスペクトログラムを NMF により分解し，アクティベーション行列を推定したのち，それをターゲット話者の辞書と掛け合わせることにより変換スペクトル特徴量系列が得られる．ソース話者の辞書には上記の PPG が連結されており，推定されるアクティベーション行列の話者非依存性が高められることが期待される．実験では，メルケプストラム距離，音質・話者性に関する preference score，および音質に関する 5 段階 MOS 値を評価している．それらの結果によれば，いずれの評価尺度についても提案手法にポジティブな結果が得られている．

文献 [46] では，WaveNet ボコーダによって生成される音声の品質劣化を軽減する手法が提案された．WaveNet ボコーダによる音声生成では，時として値が急激にジャンプするサンプルが出現し，ノイズが混入する問題があった．この問題に対処するため，本研究では WaveNet の出力分布関数を修正し，線形予測係数に基づく自己回帰過程の尤度関数を制約として課すことを提案している（ペナルティ項の導入）．極端な振幅のジャンプは尤度ペナルティ項により抑制される．ただし線形予測係数は，音響特徴量を一旦 WORLD ボコーダに通すことで得られる音声セグメントから抽出される．この音声セグメントはまた，WaveNet による生成波形の破綻検出に利用され，破綻の程度が一定以上になった場合には当該時刻の音声サンプルが再生成されるしくみである．声質変換に関する主観評価実験により，提案手法の導入前後で話者類似性を低下させずに音質が向上することが示された．

6.3 その他

文献 [47] では, progressive deep neural networks(PDNN) による統計的パラメトリック音声合成で用いられる音響特徴量 (U/V，F0，LSP) のモデル化を提案している．従来，U/V，F0，LSP のような複数の音響特徴量を用いる場合，全ての音響特徴量を連結し DNN 音響モデルの学習が行われる．しかし，音響徴量間の次元数が違うことなどから全特徴量に対して適切な音響モデル構築が困難である．本論文で用いられる PDNN は，1 つ目のタスク用の DNN を学習しパラメータを固定し，次に学習済みの DNN と隠れ層を接続した 2 つ目のタスク用の DNN を学習するというように，サブ DNN の学習を順々に行う．各サブ DNN は異なる最適化手順を用いて学習できると共に，隠れ層が接続されていることから，これまで学習されてきたサブ DNN の情報を利用可能となる．主観評価実験において，PDNN により合成音声の品質が向上したことが示されている．

文献 [48] では，合成された音声において不気味な谷現象が観測されるかどうかについて調査している．人間の音声と様々な年代 (1974 年から 2018 年の間) の 12 個の音声合成システムから合成された音声を likeability('Please listen to the voice and judge the level of Likability. i.e., How much do you like the voice speaking?') と human likeness('Please listen to the voice and judge the level of human likeness. i.e., How close to human would you rate the voice speaking?') という観点でリスニングテストを行い評価している．likeabikity と human likeness の評価値から計算された相関係数より，これらに高い線形の関係性があることが読み取れ，本実験では音声合成における不気味の谷現象には否定的な結果となっている．

7. 医療・支援技術

今回の Interspeech2018 では, 医療および支援技術に関係する研究発表が盛んに行われ, スペシャルセッション “Integrating Speech Science and Technology for Clinical Applications”, など 6 つの関連セッションとその他のテーマのセッションで合計 66 件の発表があった．学習と福祉のための音声・言語処理 (香港大 H. Meng) の演題があり, また,Perspective Talk において, 近年の Deep Learning 技術を使った人間の聴覚・音声処理に関わる技術が取り上げられた (コロンビア大 N. Mesgarani)．そこでは, 脳波により脳の活動からコンピュータ動かすインターフェースである Brain Compute Interface (BCI) への応用で発話障害者の意思伝達の支援が可能となることが期待される，脳の聴覚野の活動を測った信号から音声へとデコーディングする研究や, 人のカクテルパーティー効果のように補聴器上で特定の音声を選択的に強調することを目指して, 被験者が注意を向けた対象音声を脳波を使って追跡する研究が紹介された．

音声信号処理の医療・福祉応用には多岐のものが含まれ, 大きくは

(1) 発声や構音に関わる音声・発話・構音障害, または他の疾患・障害の影響で音声が変化したり音声に特徴が現れるものについて, 音声を使って検出・評価する技術
(2) 構音障害を伴う音声を入力とした音声認識技術
(3) 補聴器・人工内耳の改良・評価のための技術
(4) 音声分析を用いた音声リハビリ
(5) 障害のある人への音声信号処理技術を使った支援機器・インタフェース

といったものに分けられる．(1) の対象の疾患・障害としては,Interspeech2018 では, 声帯の病変などから起こる各種の音声障害, 神経疾患による麻痺による構音障害や, 口蓋裂による開鼻声などの構音障害, 韻律や声質に特徴が表れるうつ病や双極性障害といった気分障害, 認知症, 自閉スペクトラム症などが取り上げられていた．自動評価には, 検査を簡便することや定量化, スクリーニング検査を大規模に行いやすくすること, 人間が気づきにくい特徴を捉えることなどにより, 早期発見・早期治療を可能にするという利点がある．

音声の特徴に応じてさまざまな手法が取られるが, うつ病の自動判別では音声だけではなく発話内容のテキスト情報も影響を受けるという観点から, 言語情報と音響特徴量それぞれを入力とした LSTM を用いた方法 [49] や, 声質の関わる特徴量を用いた MFCC の i-Vector を用いた手法 [50] がある．i-Vector を用いた手法には, 声質の判別のための研究 [51] やパーキンソン病を対象として歩き方・手書きとともに音響特徴量の i-Vector を用いた研究 [52] があった．声帯の障害や筋萎縮性側索硬化症 (ALS) などの神経疾患による声質の変化については, スペクトログラムを入力に用いる Convolutional Neural Network (CNN) による判別を用いた発表があった [53] [54].

この分野の特徴として, プライバシーの問題で公開が難しい場合や, 疾患によっては対象者が少ないためコーパスが小規模であったり, 重症度やその他の特徴のばらつきが大きいことがあるため, 精度の高低は単純に比較できない場合もあることに注意する必要がある．それに対して, 利用を簡便にしたり, 録音データの収集を広く行えるようにするための工夫として, 携帯端末のアプリ上での録音や, 人間が対話によって検査・問診する代わりにバーチャルのエージェントが被検者に質問して答えてもらう方法を取った研究もあった．

文献 [55] は, 音声障害のリスクがある小児を, 単語音声からスクリーニング検査で発見する研究である．165 名の児童がスマートフォンアプリ上に表示した物や動物の名前 29 単語を呼称したデータを用いた．MFCC とその Δ,Δ^2 特徴量の GMM を学習して i-Vector に変換し, リスク・非リスクのクラス分類のために L2-Regularized Logistic Regression (L2LR) を用いてグループレベルの話者照合を拡張した臨床グループ照合を行う．そのために, 各単語について, リスクあり・非リスクの各クラスのモデルに対して尤度から計算したスコアの総和を取り, スコアが高いモデルに分類した．

文献 [51] は自然言語処理と音声認識を組み合わせた認知症診断のための研究である．認知症は, 初期には脳画像や問診による診断が難しい．会話に集中しながら言いたい内容を明確に表現する能力が認知症により影響を受けるとことから, 患者の発話内容の言語情報に特徴が表れると考えられるため, 文献 [51] では, を w2vec に代表される単語のベクトル表現化技術を使って発話内容から初期段階で使える検出手法の開発を目指している．言語情報を抽出するために,Kaldi(TDDN-LSTM レシピ) により音声認識を行った．発話した単語から重要でないものと複数回出現したものを取り除いた後,GloVe によりベクトル表現を得る．各単語のベクトルにより (平均値, 分散) からなる 2 次元ベクトルを計算し, 幅 80 単語の sliding window ごとに学習データから算出した各クラスの (平均値, 分散) のベクトルとの距離を計算する．その距離の値の系列から,CNN と LSTM を組み合わせた識別器により識別を行う．絵についての説明の録音 (DementiaBank), 心理士との会話 (Hallam), バーチャルアシスタントとの会話 (IVA) の 3 つのコーパスを分析したところ, 音声認識の Word Error Rate は 26%–45%程度と高かったが, 提案手法を用いた認知症か否かの識別実験では, 人手での書き起こしを使った場合と同程度か上回る識別率が得られ，音声認識の誤認識に頑健であることが示された．

参考文献

[1] Zhong-Qiu Wang, Jonathan Le Roux, DeLiang Wang, and John R. Hershey: End-to-End Speech Seperation with Unfolded Iterative Phase Reconstruction, *Proc. Interspeech*, pp. 2708–2712 (2018).
[2] Naoya Takahashi, Nabarun Goswami Purvi Agrawal, and Yuki Mitsufuji: PhaseNet: Discretized Phase Modeling with Deep Neural Networks for Audio Source Seperation, *Proc. Interspeech*, pp. 2713–2727 (2018).
[3] Nima Mesgarani Yi Luo: Real-time Single-channel Dereverberation and Separation with Time-domain Audio Separation Network, *Proc. Interspeech*, pp. 342–345 (2018).
[4] Yuxuan Wang and DeLiang Wang: A deep neural network for time-domain signal reconstruction, *Proc. Interspeech*, pp. 4390–4394 (2018).
[5] Yifan Gong Biing-Hwang (Fred) Juang Zhong Meng, Jinyu Li: Cycle-Consistent Speech Enhancement, *Proc. Interspeech*, pp. 1165–1169 (2018).
[6] Yanmin Qian Dan Su Dong Yu Lianwu Chen, Meng Yu: Permutation Invariant Training of Generative Adversarial Network for Monaural Speech Separation, *Proc. Interspeech*, pp. 302–306 (2018).
[7] Shuang Xu Bo Xu Chenxing Li, Tieqiang Wang: Single-channel Speech Dereverberation via Generative Adver-

sarial Training, *Proc. Interspeech*, pp. 1309–1313 (2018).
[8] Risheng Xia Junfeng Li Yonghong Yan Lu Yin, Ziteng Wang: Multi-talker Speech Separation Based on Permutation Invariant Training and Beamforming, *Proc. Interspeech*, pp. 851–855 (2018).
[9] DeLiang Wang Zhong-Qiu Wang: Integrating Spectral and Spatial Features for Multi-Channel Speaker Separation, *Proc. Interspeech*, pp. 2718–2722 (2018).
[10] DeLiang Wang Zhong-Qiu Wang: All-Neural Multi-Channel Speech Enhancement, *Proc. Interspeech*, pp. 3234–3238 (2018).
[11] Suyoun Kim, Michael Seltzer, Jinyu Li, and Rui Zhao: Improved Training for Online End-to-end Speech Recognition Systems, *Proc. Interspeech*, pp. 2913–2917 (2018).
[12] Ruoming Pang, Tara Sainath, Rohit Prabhavalkar, Suyog Gupta, Yonghui Wu, Shuyuan Zhang, and Chung-Cheng Chiu: Compression of End-to-End Models, *Proc. Interspeech*, pp. 27–31 (2018).
[13] Albert Zeyer, Kazuki Irie, Ralf Schlüter, and Hermann Ney: Improved Training of End-to-end Attention Models for Speech Recognition, *Proc. Interspeech*, pp. 7–11 (2018).
[14] Anuroop Sriram, Heewoo Jun, Sanjeev Satheesh, and Adam Coates: Cold Fusion: Training Seq2Seq Models Together with Language Models, *Proc. Interspeech 2018*, pp. 387–391 (online), DOI: 10.21437/Interspeech.2018-1392 (2018).
[15] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura: Machine Speech Chain with One-shot Speaker Adaptation, *Proc. Interspeech 2018*, pp. 887–891 (online), DOI: 10.21437/Interspeech.2018-1558 (2018).
[16] Jesús Andrés-Ferrer, Nathan Bodenstab, and Paul Vozila: Efficient Language Model Adaptation with Noise Contrastive Estimation and Kullback-Leibler Regularization, *Proc. Interspeech 2018*, pp. 3368–3372 (online), DOI: 10.21437/Interspeech.2018-1345 (2018).
[17] Marc Delcroix, Shinji Watanabe, Atsunori Ogawa, Shigeki Karita, and Tomohiro Nakatani: Auxiliary Feature Based Adaptation of End-to-end ASR Systems, *Proc. Interspeech 2018*, pp. 2444–2448 (online), DOI: 10.21437/Interspeech.2018-1438 (2018).
[18] Khe Chai Sim, Arun Narayanan, Ananya Misra, Anshuman Tripathi, Golan Pundak, Tara Sainath, Parisa Haghani, Bo Li, and Michiel Bacchiani: Domain Adaptation Using Factorized Hidden Layer for Robust Automatic Speech Recognition, *Proc. Interspeech 2018*, pp. 892–896 (online), DOI: 10.21437/Interspeech.2018-2246 (2018).
[19] Masayuki Suzuki, Tohru Nagano, Gakuto Kurata, and Samuel Thomas: Inference-Invariant Transformation of Batch Normalization for Domain Adaptation of Acoustic Models, *Proc. Interspeech 2018*, pp. 2893–2897 (online), DOI: 10.21437/Interspeech.2018-1563 (2018).
[20] Yu-An Chung and James Glass: Speech2Vec: A Sequence-to-Sequence Framework for Learning Word Embeddings from Speech, *Proc. Interspeech*, pp. 811–815 (2018).
[21] Benjamin Milde and Chris Biemann: Unspeech: Unsupervised Speech Context Embeddings, *Proc. Interspeech*, pp. 2693–2697 (2018).
[22] Da-Rong Liu, Kuan-Yu Chen, Hung yi Lee, and Lin shan Lee: Completely Unsupervised Phoneme Recognition by Adversarially Learning Mapping Relationships from Audio Embeddings, *Proc. Interspeech*, pp. 3748–3752 (2018).
[23] Yingke Zhu, Tom Ko, David Snyder amd Brian Mak, and Daniel Povey: Self-attentive Speaker Embeddings for Text-Independent Speaker Verification, *Proc. Interspeech*, pp. 3573–3577 (2018).
[24] Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda: Attentive Statistics Pooling for Deep Speaker Embedding, *Proc. Interspeech*, pp. 2252–2256 (2018).
[25] Zili Huang, Shuai Wang, and Kai Yu: Angular Softmax for Short-Duration Text-independent Speaker Verification, *Proc. Interspeech*, pp. 3623–3627 (2018).
[26] Sarthak Yadav and Atul Rai: Learning Discriminative Features for Speaker Identification and Verification, *Proc. Interspeech*, pp. 2237–2241 (2018).
[27] Na Li, Deyi Tuo, Dan Su, Zhifeng Li, and Dong Yu: Deep Discriminative Embeddings for Duration Robust Speaker Verification, *Proc. Interspeech*, pp. 2262–2266 (2018).
[28] Wenhao Ding and Liang He: MTGAN: Speaker Verification through Multitasking Triplet Generative Adversarial Networks, *Proc. Interspeech*, pp. 3633–3637 (2018).
[29] E. Malykh A. Kozlov O. Kudashev G. Lavrentyeva, S. Novoselov and V. Shchemelinin: Audio replay attack detection with deep learning frameworks, *Proc. Interspeech*, pp. 82—86 (2017).
[30] Hema a. Murthy Saranya M.S.: Decision-level feature switching as a paradigm for replay attack detection, *Proc. Interspeech*, pp. 686–690 (2018).
[31] Chamith Wijenayake Eliathamby Ambikairajah Gajan Suthokumar, Vidhyasaharan Sethu: Modulation Dynamic Features for the Detection of Replay Attacks, *Proc. Interspeech*, pp. 691–695 (2018).
[32] Eliathamby Ambikairajah Kaavya Sriskandaraja, Vidhyasaharan Sethu: Deep Siamese Architecture Based Replay Detection for Secure Voice Biometric, *Proc. Interspeech*, pp. 671–675 (2018).
[33] Mousmita Sarma, Pegah Ghahremani, Daniel Povey, Nagendra Kumar Goel, Kandarpa Kumar Sarma, and Najim Dehak: Emotion Identification from Raw Speech Signals Using DNNs, *Proc. Interspeech*, pp. 3097–3101 (2018).
[34] Promod Yenigalla, Abhay Kumar, Suraj Tripathi, Chirag Singh, Sibsambhu Kar, and Jithendra Vepa: Speech Emotion Recognition Using Spectrogram & Phoneme Embedding, *Proc. Interspeech*, pp. 3688–3692 (2018).
[35] Jaejin Cho, Raghavendra Pappagari, Purva Kulkarni, Jesús Villalba, Yishay Carmiel, and Najim Dehak: Deep neural networks for emotion recognition combining audio and transcripts, *Proc. Interspeech*, pp. 247–251 (2018).
[36] Saurabh Sahu, Rahul Gupta, and Carol Espy-Wilson: On Enhancing Speech Emotion Recognition using Generative Adversarial Networks, *Proc. Interspeech*, pp. 3693–3697 (2018).
[37] Srinivas Parthasarathy and Carlos Busso: Ladder Networks for Emotion Recognition: Using Unsupervised Auxiliary Tasks to Improve Predictions of Emotional Attributes, *Proc. Interspeech*, pp. 3698–3702 (2018).
[38] Zack Hodari, Oliver Watts, Srikanth Ronanki, and Simon Kin: Learning Interpretable Control Dimensions for Speech Synthesis by Using External Data, *Interspeech 2018*, pp. 32–36 (2018).
[39] Hao Li, Yongguo Kang, and Zhenyu Wang: EMPHASIS: An Emotional Phoneme-based Acoustic Model for Speech Synthesis System, *Interspeech 2018*, pp. 3077–3081 (2018).
[40] Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo: Ex-

pressive Speech Synthesis via Modeling Expressions with Variational Autoencoder, *Interspeech 2018*, pp. 3067–3071 (2018).

[41] Akira Tamamori, Tomoki Hayashi, Kazuhiro Kobayashi, Kazuya Takeda, and Tomoki Toda: Speaker-Dependent WaveNet Vocoder, *Proc. Interspeech 2017*, pp. 1118–1122 (online), DOI: 10.21437/Interspeech.2017-314 (2017).

[42] Lauri Juvela, Vassilis Tsiaras, Bajibabu Bollepalli, Manu Airaksinen, Junichi Yamagishi, and Paavo Alku: Speaker-independent Raw Waveform Model for Glottal Excitation, *Proc. Interspeech 2018*, pp. 2012–2016 (online), DOI: 10.21437/Interspeech.2018-1635 (2018).

[43] Li-Juan Liu, Zhen-Hua Ling, Yuan Jiang, Ming Zhou, and Li-Rong Dai: WaveNet Vocoder with Limited Training Data for Voice Conversion, *Proc. Interspeech 2018*, pp. 1983–1987 (online), DOI: 10.21437/Interspeech.2018-1190 (2018).

[44] Berrak Sisman, Mingyang Zhang, and Haizhou Li: A Voice Conversion Framework with Tandem Feature Sparse Representation and Speaker-Adapted WaveNet Vocoder, *Proc. Interspeech 2018*, pp. 1978–1982 (online), DOI: 10.21437/Interspeech.2018-1131 (2018).

[45] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng: Phonetic posteriorgrams for many-to-one voice conversion without parallel data training, *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6 (online), DOI: 10.1109/ICME.2016.7552917 (2016).

[46] Yi-Chiao Wu, Kazuhiro Kobayashi, Tomoki Hayashi, Patrick Lumban Tobing, and Tomoki Toda: Collapsed Speech Segment Detection and Suppression for WaveNet Vocoder, *Proc. Interspeech 2018*, pp. 1988–1992 (online), DOI: 10.21437/Interspeech.2018-1210 (2018).

[47] Ruibo Fu, Jianhua Tao, Yibin Zheng, and Zhengqi Wene: Transfer Learning Based Progressive Neural Networks for Acoustic Modeling in Statistical Parametric Speech Synthesis, *Interspeech 2018*, pp. 907–911 (2018).

[48] Alice Baird, Emilia Parada-Cabaleiro, Simone Hantke, Felix Burkhardt, Nicholas Cummins, and Björn Schuller: The Perception and Analysis of the Likeability and Human Likeness of Synthesized Speech, *Interspeech 2018*, pp. 2863–2867 (2018).

[49] Ghassemi M. Al Hanai, T. and J. Glass: Detecting Depression with Audio/Text Sequence Modeling of Interviews, *Proc. Interspeech*, pp. 1716–1720 (2018).

[50] Guo J. Park S. J. Ravi V. Flint J. Afshan, A. and A. Alwan: Effectiveness of Voice Quality Features in Detecting Depression, *Proc. Interspeech*, pp. 1676–1680 (2018).

[51] Rudolph J. Dollaghan C. McGlothlin J. Campbell T. Kothalkar, P. and J. H. Hansen: Fusing Text-Dependent Word-Level i-Vector Models to Screen 'at Risk' Child Speech., *Proc. Interspeech*, pp. 36–78 (2018).

[52] Vásquez-Correa J. C. Orozco-Arroyave J. R. Garcia, N. and E. Nöth: Multimodal i-vectors to Detect and Evaluate Parkinson's Disease, *Proc. Interspeech*, pp. 2349–2353 (2018).

[53] Soraghan-J. Lowit A. Wu, H. and G. Di Caterina: A Deep Learning Method for Pathological Voice Detection Using Convolutional Deep Belief networks, *Proc. Interspeech*, pp. 446–450 (2018).

[54] Kim-M. Teplansky K. Green-J. R. Campbell T. F. Yunusova Y. Heitzman D. An, K. and J. Wang: Automatic Early Detection of Amyotrophic Lateral Sclerosis from Intelligible Speech Using Convolutional Neural Networks, *Proc. Interspeech*, pp. 1913–1917 (2018).

[55] Blackburn-D. Walker T. Venneri-A. Reuber M. Mirheidari, B. and H. Christensen: Detecting signs of dementia using word vector representations, *Proc. Interspeech*, pp. 1893–1897 (2018).