Vol.2018-SLP-123 No.2
2018/7/26

国際会議ICASSP2018報告

秋田　祐哉[1]　安藤　厚志[2]　岡本　拓磨[3]　小川　厚徳[2]　神田　直之[4]　倉田　岳人[5]　郡山　知樹[6]
篠崎　隆宏[6]　高島　遼一[3]　太刀岡　勇気[7]　藤本　雅清[3]　増村　亮[2]

概要：2018 年 4 月 15 日から 20 日にかけて，カナダ・アルバータ州カルガリーにて IEEE 主催の国際会議 ICASSP2018 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，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

2018 年 4 月 15 日から 20 日にかけて，カナダ・アルバータ州カルガリーにて IEEE 主催の国際会議 ICASSP2018 (The 43rd 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．本年の論文の投稿数は 2,830 件あり，うち 1,406 件が採択された（採択率 49.7%）．本稿では音声言語情報処理に関する分野に注目して，ICASSP2018 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．（秋田）

2. 音声強調

音声強調のセッションでも多くが深層学習，及び Neural network（NN）に基づくアプローチを取り入れており，従来の物理モデルに基づくアプローチを圧倒していた．一方，一時期隆盛したスパース性に基づくアプローチはあまり見られなかった．また，特筆すべきは画像処理の分野から取り込まれた敵対的生成ネットワーク（GAN: Generative Adversarial Network）を用いた音声強調である．スペシャルセッション「GANs for speech enhancement」において多くの聴講者を集めており，その関心の高さが伺い知れた．従来の物理モデルに基づくアプローチ（音源位置と音響情報の同時利用）としては，音源位置推定（MCC-PHAT）／音源分離（BF）／話者同定（pitch）の複数特徴量を取り出し，Generalized labeled multi-Bernoulli でトラッキングとフィルタリングを統合的に行う手法 [1] が報告された．この手法は，トラッキング結果の明示的な可視化ができるため，原因分析がやりやすい．一方で，従来法と NN に基づく方法のハイブリッド手法も提案されている．例えばいくつかの固定のビームを設計しておき，それを選択するネットワークを Permutation invariant training（PIT）で学習する手法 [2] では，ロバスト性と計算の簡易性が利点として挙げられる．

音声強調における NN の学習法に関しても提案があった．例えば，音声合成で注目を集めている WaveNet を音声強調に応用した手法が提案された．この方法では，過去の情報だけでなく未来の情報も用いて Causal convolution を行うことにより，高品質な強調音声を得ることに成功した [3]．また，モデル学習を改善する Progressive learning（FF の途中に中間ターゲットを置きながら，少しずつ目標に近づける）に基づき，SNR を基準とした中間ターゲットを導入することで性能改善が得られることが報告された [4]．この方法では，中間ターゲットでコンテキスト拡張ができないことが問題であったが，LSTM を用いることで解決を試みた．（藤本，太刀岡）

3. 音声認識

3.1 フレームワーク

近年，音響モデル，言語モデル，発音辞書などの構成要素を一つのニューラルネットワークで表現した End-to-End(E2E) 型の音声認識の研究が活発化しており ICASSP2018 でも多くの発表が行われた．

E2E 型音声認識の枠組みは，認識対象がサブワード (音

[1] 京都大学
[2] 日本電信電話株式会社
[3] 情報通信研究機構
[4] 株式会社日立製作所
[5] 日本 IBM 株式会社
[6] 東京工業大学
[7] デンソーアイティーラボラトリ
*1 著者は 50 音順である．

素や文字) 単位か，単語単位かで大別できる．直接単語単位で認識を行う Acoustics-to-word (A2W) モデルは，辞書や複雑なデコーダが不要な反面，大量の学習データが必要であり，単語出現数の偏りや未知語の問題がある．文献 [5] は，Connectionist Temporal Classification (CTC)[6] ベースの A2W モデル [7] に対して様々な手法を用いることで，2,000 時間程度の学習データにおいて，従来のハイブリッドモデルやサブワード単位 E2E モデルと同等の性能を実現しており，今後の E2E モデル構築時のレシピとして，またベンチマークとして大いに参考となる文献である．本文献では，モデルの初期値としてサブワード単位 CTC を用いること，学習データを短い発話順に並べること，Dropout による正則化が重要としており，また単語数=ノード数の巨大な出力層の直前に，ノード数の少ない層を挟むことで，収束速度と性能を向上させている．さらに未知語の問題に対して，単語と文字列 (スペル) を併記した文章を出力ラベルとすることで，単語と文字列両方を出力するモデルを学習させる手法を提案しており，従来の A2W と同等の認識性能を保ったまま，未知語を文字単位で認識可能にしている．文献 [8] や [9] でもサブワード単位 E2E モデルと併用することで同様の問題を解決する手法を提案しており，A2W モデルの重要テーマの一つになっていると言える．

サブワード単位 E2E モデルでは上記に挙げた A2W モデルの問題に対しては比較的頑健であるが，辞書や言語モデル，複雑なデコーダを用いない限り，スペルミスにより WER が悪化するという問題がある．文献 [10] では，サブワード単位 CTC に対して，フレーム間，ラベル間の依存関係のモデリングを強化することで，上記の問題を抑制し性能を改善する手法を提案している．基本的なアイディアは，現在フレームの前後数フレームの隠れ層出力を畳み込んだ上で識別層に渡すことであるが，単なる時間方向畳み込みではフレーム毎の重みが時不変なのに対し，提案法では Attention 機構 [11] を用いることで，事変の重みを得ている．Attention 機構の改良として，識別層前後の値を入力とした LSTM を導入し，その出力を Attention 機構の入力とすることで，擬似的な言語モデルの働きを持たせる手法や，Attention 機構の出力をベクトルへ拡張することでノード毎に異なる重みを持たせる手法を提案しており，従来の CTC に対して相対的に 20%の性能向上を示している．本文献はサブワード単位 E2E モデル，特に CTC において，改良の方向性を示す文献として参考になる．（高島）

文献 [12] は様々なテクニックを包括的に検討し，Listen Attend and Spell (LAS) [13] 型の手法で従来型 (識別学習された DNN-HMM 型音声認識) を上回る精度を達成したことで注目を集めた．本文献では構造上の工夫として Word Piece Model に基づくサブワード単位とマルチヘッド Attention 機構の導入で 11%の誤り削減を行った．さらに学習面での工夫として単語誤り最小化学習，正解ラベルの代わりに推定値を学習時に利用する Scheduled Sampling 法，正解ラベルのスムージング，及び同期型の並列学習の導入によって 27.5%の誤り削減を行った．

なお，文献 [12] で導入された重要な技術である単語誤り最小化学習については，本 ICASSP で（少なくとも本章著者が把握している限り）同時に 4 研究機関から類似の提案がなされていた [14], [15], [16], [17]．特に文献 [15], [16], [17] は強化学習の枠組みで単語誤り最小化学習を捉え，様々な評価関数を比較している点で興味深い取組みである．またユーザの教示を報酬と見立て，やはり強化学習によって音声認識率を改善する試み [18] も提案されていた．強化学習と音声認識の組み合わせは今後の研究動向として注目される．

E2E 型音声認識のもうひとつの重要な研究の方向性は，ひとつのモデルで複数のタスクを同時に行おうというものである．文献 [19] では複数の言語の音声をひとつのネットワークに学習させることで発話途中で言語が切り替わる，いわゆる Code-Switching を扱うことのできる音声認識が提案されていた．学習時に多言語文字セットに加え言語ラベルも出力させるようにすることで実行時には言語識別と音声認識が同時に行われるようになる [20]．さらに，先に Code-Switching のないデータでの学習を行い，その後 Code-Switching のあるデータで学習を継続することで高精度に Code-Switching を扱う音声認識が可能であることを示した．Code-Switching 問題に対する一貫した定式化を与えており，非常に価値の高い論文と言える．（神田）

3.2 教師なし学習

教師なし学習の枠組みとしては，生成モデルの隠れ変数を利用する方法の他，オートエンコーダ (AE) を用いる方法や特定の構成のニューラルネットを対象に triplet loss 学習や敵対的学習を行う方法がある．生成モデルとしては一般に，GMM や HMM，AE のコードに事前分布を導入する変分オートエンコーダ (VAE) などがある．文献 [21] では，ベイジアン HMM およびベイジアン HMM を変分オートエンコーダ (VAE) と組み合わせたモデルを音素の教師なし学習に応用している．F 値と正規化相互情報量 (NMI) により評価を行い，一定程度の学習が可能であることを示している．VAE と HMM の組み合わせは文献 [22] と類似しているが，両者を交互にではなく同時に学習している点に違いがある．文献 [23] では，階層化した VAE である Factorized Hierarchical Variational Autoencoder (FHVAE) をドメイン非依存の特徴量の教師なし学習に応用している．文献 [24] では triplet loss 学習を音響イベント分類のための特徴量学習に用いている．triplet loss 学習は特徴量抽出の教師なし学習という点でデノイジング AE と類

似するが，decoder ネットワークを用いない点に特徴がある．文献 [25] では認識条件の識別器のロスと敵対させる形で Teacher-Student 学習を行う方法を提案している．この他，Generative Adversarial Network(GAN) についてのチュートリアル講演 [*2]において，サイクル GAN をアライメントの無い音声とテキストからの認識器学習等に用いるアイデアが紹介された．（篠崎）

3.3 耐雑音

耐雑音音声認識に関しては，スマートスピーカを中心とする音声ホームデバイスの普及に伴い，「遠隔発声」,「複数マイク」,「複数話者」を主なキーワードとして議論が進められていた．この分野の研究においては従来，MVDR-BF 等の物理モデルに基づく歪み無し音源分離をフロントエンド処理として用いることが主流であったが，現在では少数派となり Neural-mask beamformer 等に取って代わられたという印象が強い．特に音声強調 NN（音源分離）と音声認識 NN（音響モデル）を統合的に学習，最適化する Joint training に関する発表が盛んであり，例えば文献 [26] では，音源分離，音響モデルそれぞれに複数の最適化基準を導入し，最適な組み合わせを見出すことにより性能改善を得た．文献 [27] では，話者情報を積極的に用いて性能改善を得ている．また単なる遠隔音声認識の性能改善に関する発表だけでなく，デバイスを起動するためのアクティベーションコマンドの識別，認識性能改善，それに伴う音声区間検出（音響イベント検出を含む）等についても数多くの発表があった．アクティベーションコマンドの認識は基本的にオフライン（デバイス本体）で行われる処理であるため，処理量の軽減が極めて重要であり，Teacher-student training 等を用いて，いかにコンパクトかつ高精度な NN を構築するかということが盛んに議論されていた [28]．

（藤本，太刀岡）

4. 話者認識

話者認識については, オーラル 1 セッション (SP-L5: Neural Methods in Speaker Recognition and Verification)，ポスター 3 セッション (SP-P4: Speaker, Dialect, and Language ID and Multilinguality, SP-P5: Speaker Diarization & Identification, SP-P7: Deep Learning for Speaker Recognition & Verification) において，計 27 件の発表が行われた．

話者認識では，発話から話者情報を表現するベクトル (話者表現) を抽出したのち，2 つの話者表現の同一性を評価することで話者の照合や識別を行う枠組みが一般的である．話者表現には i-vector を，同一性評価には Probabilistic linear discriminative training (PLDA) を用いる手法がこれまでの主流であったが，今回の ICASSP では i-vector に代わり，DNN に基づく話者表現が発表の過半数を占めた．

DNN に基づく話者表現の代表例として，d-vector と呼ばれる，多数の話者を含む学習データから話者分類を行うネットワークを学習し，そのボトルネック特徴を話者表現とする手法が挙げられる．Eric らは，LSTM による d-vector 抽出モデルの学習にカリキュラム学習を導入する手法を提案した [29]．具体的には，テキスト一致発話，テキスト部分一致発話，テキスト不一致発話，全発話の順に d-vector 抽出モデルを学習することで，一度に全データを学習する場合に比べて話者照合誤りが 30%以上低減することが示されている．

また DNN に基づく話者表現では，同じ話者の発話 2 つと別話者の発話 1 つの組を与え，同じ話者の話者表現に対しては距離が小さく，別話者の話者表現に対しては距離が大きくなるようネットワークを学習する triplet loss に基づく手法も存在する．Li らは，この triplet loss に基づく手法をミニバッチ学習可能な Generalized End-to-End モデルに拡張した [30]．テキスト依存/非依存話者照合の両タスクにおいて，提案モデルは従来の triplet loss に基づく手法に比べて学習時間を 60%削減しつつ話者照合誤りを 10%以上削減している．また論文中では，対象ドメインに類似したドメインの大量データを補助的に利用して話者表現抽出の学習を行う手法も提案され，話者照合誤りを 30%以上削減することが報告されている．

さらに，新たな試みとして，発話長が長くなるほど話者表現が正確になるように話者表現抽出を行う Collective Network (CLNet) が提案された [31]．CLNet は RNN による d-vector 抽出モデルと似た構造を取るが，話者表現を逐次更新する，すなわち直前の話者表現と現時刻の特徴量から求めた差分に基づいて現在の話者表現を決定する点が異なる．CLNet は従来の CNN/RNN による d-vector や i-vector に比べて話者照合精度が高く，さらに発話長の増加に伴う精度低下が発生しないことが実験的に示されている．（安藤）

5. 感情認識

感情認識については，オーラル 2 セッション (SP-L7: Emotion recognition and Biometrics, SS-L8: Deep Learning for Computational Paralinguistics)，ポスター 2 セッション (SP-P1: Emotion, Sentiment and Speech Analysis, SP-P2: Prosody and Emotion) を中心に計 31 件の発表があった．

感情認識における重要課題の一つとして，ラベルあり学習データがきわめて少量である点が挙げられる．今回の ICASSP では，この課題を解決するための試みが数多く発表された．

[*2] http://sigport.org/2863

少量のラベルあり学習データからの学習では正則化がよく用いられることから，感情認識においても正則化に着目した手法がいくつか提案された．例えばSaurabhらは，入力特徴量に微小雑音を加えた場合でも正解ラベルや識別結果が変わらないように学習を行うAdversarial TrainingやVirtual Adversarial Trainingを感情認識モデル学習に適用し，同一コーパス評価およびクロスコーパス評価の両方で感情分類精度が改善することを示した[32]．Cheらは，畳み込み層が分岐するResidual Networkを用いて，各分岐をランダムな重み付け和で組み合わせたものを学習するShake-Shake regularizationを感情認識に応用した[33]．さらに，感情情報は周波数帯域ごとに偏りをもって表れるという仮説に基づき，低域/高域ごとにShake-Shake regularizationを適用することで感情分類精度が向上することを示した．これらの正則化に基づく手法は一定の効果が得られたが，誤り削減率は10%以下と限定的であった．

別のアプローチとして，表現学習の枠組みを応用し，感情認識に有効な特徴量を教師なしで獲得する枠組みも提案されている．Sefikらは，入力発話を低次元の潜在変数空間に射影するエンコーダを教師なしで学習させ，その後エンコーダ出力から感情カテゴリを推定するモデルを少量の教師ありラベルで学習する手法を提案した[34]．潜在変数空間への射影モデルにはVariational AutoencoderやAdversarial Variational Bayesなどの幾つかの手法が試され，いずれも一般的なCNNに基づく手法に比べて高い感情分類精度を示したことが報告されている．Lixingらは，Autoencoderに基づく教師なし特徴量獲得において，発話から話者性と感情という直行する二つの潜在変数を同時に求めることを試みるOrthogonal Autoencoderを提案した[35]．提案手法は幾つかのAutoencoderによる教師なし特徴量獲得に比べて高い感情分類精度を示し，また話者性 (男性/女性) と感情 (緊張/平常/リラックス) が潜在変数空間において分離されたことが確認されている．

半教師あり学習によりラベル無しデータを学習に利用する手法も提案された．Rahulらは，文書からの感情分類において，近傍に存在するデータは同じラベルが表れるようにラベル付きデータとラベル無しデータを組み合わせて学習するmanifold regularizationを導入し，ラベル付き学習データが1/1000の場合でも感情分類精度が66%から62%までしか低下しないことを示した[36]． （安藤）

6. 音声合成・音響信号処理・声質変換

音声合成・声質変換を中心としたセッションはオーラルセッションが1つ (SP-L2: Neural Network based Speech Synthesis)，ポスターセッションが2つ (SP-P6: Voice Transformation, SP-P14: Speech Synthesis, Generation and Coding) で26件の講演があった．特にWaveNet関連の発表は全体の約1/3を占め，さらに音声符号化や音声強調の他分野にも応用されていた．日本からは，SP-L2がNIIの2件とパクーシャテクノロジー1件の計3件，SP-P6が東大，NII，名工大，神戸大の計4件，SP-P14がNICTの2件と，日本勢は約1/3と大貢献したと言える．また，音声認識も含めた今回のトレンドはやはりend-to-end(E2E) とGANと言えるが，テキスト音声合成ではTacotron 2(含WaveNet) のE2Eがあり，声質変換では昨年度のInterspeechでの発表からのGANの発表が多数見受けられ，音声認識と同様の傾向が見受けられた．以下では，主に海外からの発表について報告する．

6.1 sequence-to-sequence(seq2seq) 音声合成

伝統的な統計的パラメトリック音声合成は，発音やアクセントなどのコンテキストから音声パラメータを予測する枠組みであったが，近年はテキストの文字列から音声波形を直接合成するend-to-end音声合成の枠組みが注目を集めている．

文献[37]では，attentionに基づくseq2seq音声合成において後述するWaveNetをボコーダとして用いるTacotron 2を提案し，自然音声とほぼ同等の自然性を持つ音声の合成を実現している．また，Tacotronとの差分としてネットワーク構造の変更やattentionのアラインメント精度向上のためのlocation-sensitive attentionの導入を行っている．

seq2seq音声合成において音響特徴量列と音素列は基本的に同じ順番で並んでいることから，attentionのアラインメントは時間軸に対し単調に変化することが望ましい．文献[38]ではseq2seq音声合成におけるアラインメント精度向上に向けて，前の時刻のattentionと類似したattentionになるように強い制約を持たせたforward-attentionが提案された．提案手法によって，合成音声における音素の欠落や繰り返しが減少することが示された．

文献[39]では，encoderおよびdecoderのRNNをCNNに置き換えることにより，高速な学習が可能なattention-seq2seq音声合成が提案された．提案法では15時間程度の学習で，約12日学習したTacotronより自然性の高い音声が合成可能であることを示している．また，この手法では単調変化するattentionの実現を目的として，attentionが対角線上から離れるとコストが大きくなるようなロスを導入している． （郡山）

6.2 WaveNetの性能分析

文献[40]では，WaveNetを始めとして近年発表された様々な波形生成・音響モデルの比較評価を行った．実験結果から，波形生成モデルにはWaveNetボコーダが有効であり，サンプリング周波数16kHzの波形を生成するWaveNetが48kHzの他手法より高性能であることが示された．ま

た，音響モデルの比較実験では自己回帰構造を持つモデルの有効性が示された．

文献 [41] では，コンテキストから波形を直接生成する WaveNet TTS の枠組みにおいて学習データ量や学習データに含まれるエラーの影響を調査した．実験の結果，学習データ量を 10000 文 (14 時間) から 2000 文 (3 時間) 程度まで減らしても合成音声の品質が大きく劣化しないことを示した．（郡山）

6.3 リアルタイム波形直接生成型ボコーダ

WaveNet は上記にある通り，英語音声合成において自然音声と同等の品質を実現できるが，1 サンプルずつを巨大なネットワークで逐次計算するために，生成に時間を要する問題がある．これに対して，WaveNet の dileted causal convolution によるダウンサンプリングが Wavelet 変換に相当することに着目し，高速フーリエ変換 (FFT) に相当するダウンサンプリングを導入した FFTNet が提案された．WaveNet に比べればモデルサイズは約 1/20 であり，リアルタイム生成が可能である．しかし，音質的にはまだ課題が残る [42]．（岡本）

6.4 WaveNet の音声符号化や音声強調への応用

従来の低ビットレート (2.4 kbps) の符号化情報 (サンプリング周波数 8 kHz) を条件としてサンプリング周波数 16 kHz の音声を WaveNet 学習させることにより，WaveNet を帯域拡張型音声復号器とする発表があり，非常に注目を集めていた [43]．従来の 2.4 kbps の方式と比べて，格段に品質の向上があった．また，音声強調への応用については 2 節で紹介された [3] があった．（岡本）

7. Human Language Technology

7.1 音声言語理解

音声言語理解の技術トレンドは，例年と変わらずニューラルネットワークを用いた技術が支配的であったが，音声認識誤りを考慮した手法が数多く見られた．音声言語理解の技術検討は，音声認識誤りを含まない正解の書き起こしを用いることが一般的であるが，音声認識誤りを含むテキストに対しては大きく性能劣化してしまうことが知られている．以下では，音声認識誤りへの対処に着目した 4 つの文献を紹介する．

文献 [44] では，スロットフィリングや発話意図推定のためのニューラルネットワークに，音声認識誤りを訂正するためのニューラルネットワークを連結し，一体で学習するアプローチを提案している．音声言語理解に対する改善効果に加え，4 ポイント程度の単語誤り率の改善も報告しており，音声認識の観点においても注目されるアプローチである．文献 [45] では，音声認識結果のコンフュージョンネットワークから発話意図推定を直接行うためのニューラルネットワークを提案しており，n-best リスト等を使う場合と比較して高い性能が得られることを報告している．コンフュージョンネットワークを扱うニューラルネットワークは，発話意図推定以外の様々な音声言語理解に適用可能であり，興味深いアプローチと言える．文献 [46] では音声認識誤りに頑健なスロットフィリングのために，スロットフィリング用のニューラルネットワークと共に，音声認識結果自体を再構成するニューラルネットワークを一体でモデル化するアプローチを提案しており，これにより，音声認識結果が入力の場合でも頑健に動作することを報告している．このアプローチの利点は，ラベル付きの音声認識結果を準備する必要がない点であり，単語単位のラベリングが必要なスロットフィリングにとっては，特に有望な方法といえる．文献 [47] では，音声認識結果に左右されない処理を行うために，音響特徴量系列から直接意図ラベルを推定するモデル化を提案している．音響特徴量系列から単語系列を推定するニューラルネットワークと，単語系列から意図ラベルを推定するニューラルネットワークをそれぞれプリトレーニングしておくことで効率的なモデル化が可能であり，音声認識用の学習データも間接的に利用できる点は非常に実用的である．

これらの技術動向から，今後は音声翻訳や音声要約などの他の音声言語処理技術についても，音声認識誤りを考慮したニューラルネットワークベースのアプローチの検討が進むと考えられる．

7.2 言語モデル

言語モデルにおいても，引き続きニューラルネットワークを用いた技術が中心であった．

文献 [48] では，分野ごとの “Expert” モデルと，それらに対する補間重みを推定する “Mixer” モデルの組み合わせを提案している．Experts と Mixer は両方ともに LSTM を利用してモデル化をされ，予測する単語ごとに補間重みが動的に変更されることが特徴であり，YouTube ビデオの書き起こしタスクにおいて，認識率の改善が報告されている．文献 [49] では，音響モデルでよく用いられている Teacher-student modeling を，言語モデルに適用している．LSTM に比べて Feed-forward 型ニューラルネットワークは音声認識時の利用が高速化しやすいことを考慮し，LSTM から Feed-forward 型ニューラルネットワークへの Teacher-student modeling が提案されている．電話会話書き起こしタスクにおいて，LSTM と同等の性能を得るためには，10-gram までを利用した Feed-forward 型ニューラルネットワークが必要であることが示されている．

通常の言語モデリングでは，一文の生起確率 $P(\mathbf{W})$ を直接求めることは困難であるため，これを単語履歴で条件付

けられた一単語ごとの単語の生起確率の積として求めていた．これに対して，文を単位とする方法も提案されている．文献 [50] では一文の構造を的確に捉えるために，$P(\mathbf{W})$ を直接求める LSTM ベースの一文 (whole sentence) 言語モデルを提案している．一文言語モデルは，正解文と，従来の n-gram 言語モデルを用いて正解文から生成された誤りを含む文を用いて，Noise Contrastive Estimation (NCE) の枠組みで学習される．Switchboard 音声認識タスクにおける n(=100)-best リスコアリングで，着実な単語誤り率削減を実現している．文献 [51] では，n-best 仮説対に対して，文全体の情報から，どちらの仮説が良いかを判断するモデルを提案している．従来研究されてきた識別的リランキングと同様に，言語的情報以外に音響的情報なども柔軟に組み込めることが特徴であり，日本語話し言葉コーパスを対象とした実験で，大きな改善が報告されている．

（倉田，増村，小川）

参考文献

[1] Lin, S.: Jointly tracking and separating speech sources using multiple features and the generalized labeled multi-Bernoulli framework, *Proc. ICASSP*, pp. 3211–3215 (2018).
[2] Chen, Z., Yoshioka, T., Xiao, X., Li, J., Seltzer, M. L. and Gong, Y.: Efficient integration of fixed beamformers and speech separation networks for multi-channel far-field speech separation, *Proc. ICASSP*, pp. 5384–5388 (2018).
[3] Rethage, D., Pons, J. and Serra, X.: A Wavenet for speech denoising, *Proc. ICASSP*, pp. 5069–5073 (2018).
[4] Gao, T., Du, J., Dai, L. R. and Lee, C. H.: Densely connected progressive learning for LSTM-based speech enhancement, *Proc. ICASSP*, pp. 5054–5058 (2018).
[5] Audhkhasi, K., Kingsbury, B., Ramabhadran, B., Saon, G. and Picheny, M.: Building competitive direct acoustics-to-word models for English conversational speech recognition, *Proc. ICASSP*, pp. 4759–4763 (2018).
[6] Graves, A., Fernández, S., Gomez, F. and Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, *Proc. ICML*, ACM, pp. 369–376 (2006).
[7] 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).
[8] Li, J., Ye, G., Das, A., Zhao, R. and Gong, Y.: Advancing acoustic-to-word CTC model, *Proc. ICASSP*, pp. 5794–5798 (2018).
[9] Ueno, S., Inaguma, H., Mimura, M. and Kawahara, T.: Acoustic-to-word attention-based model complemented with character-level CTC-based model, *Proc. ICASSP*, pp. 5804–5808 (2018).
[10] Das, A., Li, J., Zhao, R. and Gong, Y.: Advancing connectionist temporal classification with attention modeling, *Proc. ICASSP*, pp. 4769–4773 (2018).
[11] Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K. and Bengio, Y.: Attention-based models for speech recognition, *Proc. NIPS* (2015).
[12] Chiu, C.-C., Sainath, T. N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R. J., Rao, K., Gonina, E., Jaitly, N., Li, B., Chorowski, J. and Bacchiani, M.: State-of-the-art speech recognition with sequence-to-sequence models, *Proc. ICASSP*, pp. 4774–4778 (2018).
[13] Chan, W., Jaitly, N., Le, Q. and Vinyals, O.: Listen, attend and spell: a neural network for large vocabulary conversational speech recognition, *Proc. ICASSP*, pp. 4960–4964 (2016).
[14] Prabhavalkar, R., Sainath, T., Wu, Y., Nguyen, P., Chen, Z., Chiu, C.-C. and Kannan, A.: Minimum word error rate training for attention-based sequence-to-sequence models, *Proc. ICASSP*, pp. 4839–4843 (2018).
[15] Karita, S., Ogawa, A., Delcroix, M. and Nakatani, T.: Sequence training of encoder-decoder model using policy gradient for end-to-end speech recognition, *Proc. ICASSP*, pp. 5839–5843 (2018).
[16] Tjandra, A., Sakti, S. and Nakamura, S.: Sequence-to-sequence ASR optimization via reinforcement learning, *Proc. ICASSP*, pp. 5829–5833 (2018).
[17] Zhou, Y., Xiong, C. and Socher, R.: Improving end-to-end speech recognition with policy learning, *Proc. ICASSP*, pp. 5819–5823 (2018).
[18] Kato, T. and Shinozaki, T.: Reinforcement learning of speech recognition system based on policy gradient and hypothesis selection, *Proc. ICASSP*, pp. 5759–5763 (2018).
[19] Seki, H., Watanabe, S., Hori, T., Le Roux, J. and Hershey, J. R.: An end-to-end language-tracking speech recognizer for mixed-language speech, *Proc. ICASSP*, pp. 4919–4923 (2018).
[20] Watanabe, S., Hori, T. and Hershey, J. R.: Language independent end-to-end architecture for joint language identification and speech recognition, *Proc. ASRU*, pp. 265–271 (2017).
[21] Ondel, L., Godard, P., Besacier, L., Larsen, E., Hasegawa-Johnson, M., Scharenborg, O., Dupoux, E., Burget, L., Yvon, F. and Khudanpur, S.: Bayesian models for unit discovery on a very low resource language, *Proc. ICASSP*, pp. 5939–5943 (2018).
[22] Ebbers, J., Heymann, J., Drude, L., Glarner, T., Haeb-Umbach, R. and Raj, B.: Hidden markov model variational autoencoder for acoustic unit discovery, *Proc. Interspeech*, pp. 488–492 (2017).
[23] Hsu, W.-N. and Glass, J.: Extracting domain invariant features by unsupervised learning for robust automatic speech recognition, *Proc. ICASSP*, pp. 5614–5618 (2018).
[24] Jansen, A., Plakal, M., Pandya, R., Ellis, D., Hershey, S., Liu, J., Moore, R. C. and Saurous, R. A.: Unsupervised learning of semantic audio representations, *Proc. ICASSP*, pp. 126–130 (2018).
[25] Meng, Z., Li, J., Gong, Y. and Juang, B.-H.: Adversarial teacher-student learning for unsupervised domain adaptation, *Proc. ICASSP*, pp. 5949–5953 (2018).
[26] Settle, S., Le Roux, J., Hori, T., Watanabe, S. and Hershey, J. R.: End-to-end multi-speaker speech recognition, *Proc. ICASSP*, pp. 4819–4823 (2018).
[27] Delcroix, M., Zmolikova, K., Kinoshita, K., Ogawa, A. and Nakatani, T.: Single channel target speaker extraction and recognition with speaker beam, *Proc. ICASSP*, pp. 5554–5558 (2018).
[28] Li, J., Zhao, R., Chen, Z., Liu, C., Xiao, X., Ye, G. and Gong, Y.: Developing far-field speaker system via

teacher-student learning, *Proc. ICASSP*, pp. 5699–5703 (2018).
[29] Marchi, E., Shum, S., Hwang, K., Kajarekar, S., Sigtia, S., Richards, H., Haynes, R., Kim, Y. and Bridle, J.: Generalised discriminative transform via curriculum learning for speaker recognition, *Proc. ICASSP*, pp. 5324–5328 (2018).
[30] Wan, L., Wang, Q., Papir, A. and Moreno, I. L.: Generalized end-to-end loss for speaker verification, *Proc. ICASSP*, pp. 4879–4883 (2018).
[31] Wen, Y., Zhou, T., Singh, R. and Raj, B.: A corrective learning approach for text-independent speaker verification, *Proc. ICASSP*, pp. 4894–4898 (2018).
[32] Sahu, S., Gupta, R., Sivaraman, G. and Espy-Wilson, C.: Smoothing model predictions using adversarial training procedures for speech based emotion recognition, *Proc. ICASSP*, pp. 4934–4938 (2018).
[33] Huang, C.-W. and Narayanan, S.: Shaking acoustic spectral sub-bands can better regularize learning in affective computing, *Proc. ICASSP*, pp. 6827–6831 (2018).
[34] Eskimez, S. E., Duan, Z. and Heinzelman, W.: Unsupervised learning approach to feature analysis for automatic speech emotion recognition, *Proc. ICASSP*, pp. 5099–5103 (2018).
[35] Liu, L., Ghosh, S. and Scherer, S.: Towards learning nuisance-free representations of speech, *Proc. ICASSP*, pp. 6817–6821 (2018).
[36] Gupta, R., Sahu, S., Espy-Wilson, C. and Narayanan, S.: Semi-supervised and transfer learning approaches for low resource sentiment classification, *Proc. ICASSP*, pp. 5109–5113 (2018).
[37] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R. J., Saurous, R. A., Agiomyrgiannakis, Y. and Wu, Y.: Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions, *Proc. ICASSP*, pp. 4779–4783 (2018).
[38] Zhang, J.-X., Ling, Z.-H. and Dai, L.-R.: Forward attention in sequence-to-sequence acoustic modeling for speech synthesis, *Proc. ICASSP*, pp. 4789–4793 (2018).
[39] Tachibana, H., Uenoyama, K. and Aihara, S.: Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention, *Proc. ICASSP*, pp. 4784–4788 (2018).
[40] Wang, X., Lorenzo-Trueba, J., Takaki, S., Juvela, L. and Yamagishi, J.: A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis, *Proc. ICASSP*, pp. 4804–4807 (2018).
[41] Vit, J., Hanzlicek, Z. and Matousek, J.: On the analysis of training data for WaveNet-based speech synthesis, *Proc. ICASSP*, pp. 5684–5688 (2018).
[42] Jin, Z., Finkelstein, A., Mysore, G. J. and Lu, J.: FFTNet: a real-time speaker-dependent neural vocoder, *Proc. ICASSP*, pp. 2251–2255 (2018).
[43] Kleijn, W. B., Lim, F. S. C., Luebs, A., Skoglund, J., Stimberg, F., Wang, Q. and Walters, T. C.: WaveNet based low rate speech coding, *Proc. ICASSP*, pp. 676–680 (2018).
[44] Schumann, R. and Angkititrakul, P.: Incorporating ASR errors with attention-based jointly trained RNN for intent detection and slot filling, *Proc. ICASSP*, pp. 6059–6063 (2018).
[45] Masumura, R., Ijima, Y., Asami, T., Masataki, H. and Higashinaka, R.: Neural ConfNet classification: fully neural network based spoken utterance classification using word confusion networks, *Proc. ICASSP*, pp. 6039–6043 (2018).
[46] Zhu, S., Lan, O. and Yu, K.: Robust spoken language understanding with unsupervised ASR-error adaptation, *Proc. ICASSP*, pp. 6179–6183 (2018).
[47] Chen, Y.-P., Price, R. and Bangalore, S.: Spoken language understanding without speech recognition, *Proc. ICASSP*, pp. 6189–6193 (2018).
[48] Irie, K., Kumar, S., Nirschl, M. and Liao, H.: RADMM: recurrent adaptive mixture model with applications to domain robust language modeling, *Proc. ICASSP*, pp. 6079–6083 (2018).
[49] Irie, K., Lei, Z., Schlüter, R. and Ney, H.: Prediction of LSTM-RNN full context states as a subtask for n-gram feedforward language models, *Proc. ICASSP*, pp. 6104–6108 (2018).
[50] Huang, Y., Sethy, A., Audhkhasi, K. and Ramabhadran, B.: Whole sentence neural language models, *Proc. ICASSP*, pp. 6089–6093 (2018).
[51] Ogawa, A., Delcroix, M., Karita, S. and Nakatani, T.: Rescoring N-best speech recognition list based on one-on-one hypothesis comparison using encoder-classifier model, *Proc. ICASSP*, pp. 6099–6103 (2018).