Vol.2019-SLP-128 No.8
2019/7/19

国際会議ICASSP2019報告

秋田 祐哉[1]　大町 基[2]　岡本 拓磨[3]　落合 翼[4]　小川 厚徳[4]　神田 直之[5]　郡山 知樹[6]　鈴木 雅之[7]
太刀岡 勇気[8]　俵 直弘[4]　増村 亮[4]　渡部 晋治[9]

概要：2019 年 5 月 12 日から 17 日にかけて，英国ブライトンにて IEEE 主催の国際会議 ICASSP2019 (The 44th IEEE International Conference on Acoustics, Speech and Signal Processing) が開催された．ICASSP は Interspeech と並んで音声言語処理分野におけるトップカンファレンスであるが，ICASSP はよりスコープが広く信号処理分野全体を包含する懐の深さが特色である．本稿では音声言語処理に関する 5 分野に注目し，ICASSP2019 の採択論文を中心に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．

キーワード：ICASSP, 技術動向レビュー

1. はじめに

2019 年 5 月 12 日から 17 日にかけて，英国ブライトンにて IEEE 主催の国際会議 ICASSP2019 (The 44th IEEE International Conference on Acoustics, Speech and Signal Processing) が開催された．ICASSP は Interspeech と並んで音声言語処理分野におけるトップカンファレンスであるが，ICASSP はよりスコープが広く信号処理分野全体を包含する懐の深さ・投稿件数の多さが特色である．本年の論文の投稿数は 3,150 件あり，うち 1,725 件が採択された (採択率 49.1%)．本稿では音声言語処理に関する 5 分野に注目し，本 ICASSP の採択論文を中心に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]．（太刀岡）

2. 音源分離・音声認識

音源分離および音声認識の関連技術に関する発表を紹介する．2.1 では音源分離の研究を概観し，2.2 で，音声認識のフロントエンドに関する取り組みをいくつか紹介する．2.3 では，音声認識の関連技術として，キーワード・スポッティング (Keyword spotting; KWS) の研究を紹介する．

2.1 音源分離

セッション名に separation を含む音源分離に関するセッションは合計で 9 セッションあり，盛んに研究が行われていた．研究動向としては，従来からのマルチチャネル (MC) 信号に基づく研究 (e.g., Independent Component Analysis) に加えて，モノラル信号を対象とした深層学習ベースの手法 (e.g., Deep Clustering (DC), Permutation Invariant Training) に関する研究が多く見られた．ここでは，上記深層学習ベースの手法を，以下の 4 つの観点から，紹介する．

1) 位相推定　近年の深層学習技術の活用によって，音源分離における振幅成分の推定精度が大きく向上し，モノラル音源分離においても高い分離精度 (e.g., Signal-to-Distortion Ratio) が達成可能となった．そういった状況の中で，更なる分離精度の向上を目指し，位相成分の改善を目指す研究に関心が集まっていた．例えば文献 [37] では，混合信号の各振幅成分が既知と仮定すると，対応する位相成分を陽に解くことができ，結果として 2 つの位相候補が得られる点に着目し，位相推定における回帰問題を 2 つの位相候補に対する識別問題へと変換する手法を提案した．振幅の推定精度が十分に高いという前提条件はあるものの，現時点で state-of-the-art の時間領域で波形推定をする手法 (e.g., TasNet [23]) と同程度の分離精度を，時間周波数領域の枠組みで達成することに成功している．

2) オンライン (低遅延) 化　従来，分離性能の向上のため，深層学習モデルとして未来の情報も予測に利用する non-causal なモデル構造 (e.g., BLSTM) が活用されていた．そういった状況の中で，実時間性が重要となる応用 (e.g., リアルタイム音声認識) への拡張を目指し，分離手法のオンライン (低遅延) 化を目指す研究が複数発表されていた．例えば文献 [7] では，Deep Attractor Network の枠組みで，各信号源に対応する attractor をフレーム毎の移動平均に基づき逐次更新する手法を提案している．実験により提案法は，フレーム毎に処理する causal なモデル構造ながら，発話単位で処理を行う従来の non-causal なモデル

[1] 京都大学
[2] ヤフー株式会社
[3] 情報通信研究機構
[4] 日本電信電話株式会社
[5] 日立製作所
[6] 東京大学
[7] IBM
[8] デンソーアイティーラボラトリ
[9] ジョンズホプキンス大学
[*1] 著者は 50 音順である．

構造と同程度の分離精度を達成し得ると報告されている．

3) MC 学習　従来，音源分離モデルの学習に際し，単一話者の音声信号を混合した模擬データの利用が標準的であった．そういった状況の中で本 ICASSP では，実観測信号を学習に利用することを目指した研究が，複数の研究機関から報告されていた．上記研究に共通するアイデアは，モノラル音源分離モデルの学習のために，MC 信号を活用することである．例えば，文献 [31] では，DC の枠組みにおいて，学習の際に必要となる時間周波数ビン毎の各信号源の割り当て (e.g., バイナリ時間周波数マスク) を，MC 信号から得られる時間周波数ビン毎の位相差特徴量を k-means クラスタリングした結果から作成する方法を提案している．実験によって提案法は，模擬データを利用した従来法と同程度の分離精度を，MC の実観測信号のみを用いた学習によって達成し得ることを報告している．

4) 目的話者抽出　音源分離の派生研究である「目的話者抽出」に関する研究も散見された．「目的話者抽出」は，事前情報なしに混合信号を各信号源へと分離する「ブラインド音源分離」とは異なり，事前情報として目的話者の情報 (e.g., 事前録音された発話) を活用することで，混合信号から目的話者の音声のみを抽出する枠組みである．例えば文献 [26] では，目的話者の情報として，音声情報ではなく動画情報を活用する枠組みに関する研究に取り組んでいる．入力顔画像からの特徴抽出手続きにおいて，画像分野で古くから研究されているランドマーク検出器を活用することで，GRID コーパスのような小規模なデータセットにおいても，動画情報に基づいた目的話者抽出が実現可能であることが実験によって確認されている．(落合)

2.2　フロントエンドと音響モデルの一体型学習

フロントエンドと音響モデルを同時に学習することにより，認識率が向上することが知られており，本 ICASSP でもいくつかの発展的な取り組みが提案されていた．

文献 [24] では，1ch の雑音下音声認識をタスクとし，ニューラルネットワーク (NN) による音声/雑音マスクの推定と Parametric Wiener Filter (PWF) を組み合わせる手法が提案された．実験では，特に PWF の雑音推定に，NN による雑音マスクを用いることが効果的であることを示した．一方で NN による音声マスクを単純に音声にかけるだけでは，たとえ音響モデルとの一体型の学習を行っても精度劣化が生じることも確認された．比較的単純な手法ながら 10%以上の誤り削減が得られている点は興味深い．

文献 [9] は，残響除去として広く音声認識への有効性が知られている Weighted Prediction Error (WPE) を音響モデルと同時最適化する試みが提案されていた．WPE では雑音抑圧後の Power Spectral Density (PSD) の推定値が重要な役割を持つため，PSD を NN によって逐次推定する方法が過去に提案されていた [8], [14]．提案法はこの従来法を，音響モデルと同時に学習するように拡張したものに相当する．実験では，逐次推定で残響抑圧を行う場合に特に有効であること，また提案法の枠組みでは PSD 推定部はランダムな初期値からでも学習可能であり，このことから学習に実データを用いることができる点を示した．

文献 [15] はマイクロホンアレイでの音声認識のための MC ビームフォーマと音響モデルの一体型学習が提案されていた．ここでのポイントは，ビームフォーマに相当するパラメータの初期値を特定のアレイ配置，特定の方向にビームを形成するように設定することである．様々なアレイ配置，様々な方向にビームを向けた形で初期化された複数のネットワークブロックを配置してから一体型学習をすることで，そのモデルを利用する際にも複数のアレイ配置に対応でき，さらにほとんどの場合で単一のアレイ配置だけで学習した場合よりも精度が向上することを示した．

最後に筆者らの研究 [12] についても簡単に紹介する．ここではマイクロホンアレイでの音声認識を指向し，MC 音声を受け付ける入力部と 1ch 入力を受け付ける入力部の 2 つの入力部を持つ音響モデルを提案した．MC 入力部が音響モデルと同時学習されるフロントエンドの役割を担うのに対し，1ch 入力部は別個のフロントエンドから得られる音声を受け付けることができる．2 種類のフロントエンドが相補的に働き，高い認識精度が得られることを示した．当該モデルは CHiME-5 と AMI それぞれのデータセットにおいて最高精度を達成している [11], [12].（神田）

2.3　キーワード・スポッティング (KWS)

本 ICASSP では，スマートスピーカーなどの音声 UI における音声認識のトリガーとして広く用いられている KWS についてもいくつか発表があった．KWS は計算資源が限られたデバイスで利用されることが多い．近年では低演算量で高精度に KW を検出することを目指した研究が盛んにおこなわれており，本 ICASSP でも関連した取り組みが報告されていた．例えば文献 [1] では，音響特徴量から特徴表現を抽出する encoder と，特徴表現を KW または非 KW に分類する decoder の 2 つの DNN で構成される KWS システムにおいて，全結合層を低ランクの行列で近似する構造に置き換えることで演算量を削減する方法を提案していた．さらに，encoder と decoder を 1 つの DNN とみなして同時に学習することで，従来のモデルの 20%程度の演算量で誤棄却を 60%程度削減できると報告している．

学習済みのモデルの精度を改善するための取り組みもいくつか報告されていた．文献 [35] は，既存のシステムが誤検知，または，誤棄却しやすい *adversarial examples* を生成し，モデルを再学習する方法を提案しており，誤棄却を 45%削減できることを示している．文献 [17] は，モデルの学習環境と利用環境のミスマッチの影響を解消するために，クライアント上に置かれたユーザデータを用いてモデルを更新する federated learning を，KWS の枠組みで動かすためのアルゴリズムを提案している．文献 [17] の著者ら

は，クラウドソーシングで収集したデータセット[*2]を公開しており，学術・研究目的であれば利用できる．(大町)

3. End-to-End 音声認識

本 ICASSP では，End-to-End 音声認識に関してオーラル 2 セッション、ポスター 3 セッションが開催され，本研究テーマの注目度の高さが伺えた．

3.1 Transformer

本 ICASSP では，ニューラル機械翻訳の分野で 2017 年初出の Transformer を応用した End-to-End 音声認識の検討がいくつか見られた．Transformer は条件付きの自己回帰生成モデルとして表される Encoder-Decoder モデルの 1 種であり，特に Encoder 部分に強みを持つモデル化である．具体的には，Self-Attention と呼ばれる機構を用いることで，入力系列の長距離の関係性を精緻に捉えたベクトル埋め込みを行うことができ，RNN に基づく Encoder-Decoder よりも高い性能を実現できることが報告されてきている．

文献 [45] では，Transformer に基づく End-to-End 音声認識の改良が検討されている．Encoder における音響特徴量系列のサブサンプリングの工夫，Exposure Bias 問題を緩和するための Scheduled Sampling の改良，そしてトークン頻度の不均衡問題を緩和するための Focal Loss の適用を検証することで，相対誤り改善率 10-20%を達成できることを報告している．

Transformer は通常条件付きの自己回帰生成モデルを意味するが，その他の End-to-End 音声認識のモデル化にも知見を適用する検討がなされている．文献 [6] では，Transformer の Encoder ブロックと Connectionist Temporal Classification (CTC) を組み合わせが検討されており，Encoder に RNN を用いた CTC と比較して，同等または上回る性能が報告されている．また文献 [29] では，Recurrent Neural Aligner の Encoder に Transformer の Encoder ブロックを用いた Self-Attention Aligner を提案しており，Recurrent Neural Aligner と比較して，大幅な性能改善が報告されている．

3.2 新しい系列変換モデル

また，Transformer 以外では，文献 [27] において，CTC や Attention-based Encoder-Decoder とは異なる新しい系列変換モデルが提案されている．系列変換モデルでは入力 (t)・出力 (n) で異なる時刻をどのようにモデル化するかが問題となる．提案手法は二次元 LSTM を用いて一時刻前のラベル w_{n-1} と encoder の出力 $\mathbf{h}_t$ から，状態ベクトルの漸化式 $\mathbf{s}_{t,n}$ を次のように求める．

$$\mathbf{s}_{t,n} = \mathrm{2DLSTM}([\mathbf{h}_t, w_{n-1}], \mathbf{s}_{t,n-1}, \mathbf{s}_{t-1,n}) \qquad (1)$$

認識時には全ての入力時刻における状態ベクトル $\mathbf{s}_{1:T,n-1}$ の Max Pooling により，次時刻の出力単語列の事後確率 $p(w_n|w_{n-1}, \mathbf{h}_{1:T})$ を求める．本手法は計算量に問題を抱えるものの，Switchboard タスクにおいて，Attention-based Encoder-Decoder と同等の性能を達成している．

3.3 学習方法の工夫

さらに，End-to-End 音声認識の学習方法に工夫を入れ込む検討も増えてきている．文献 [20] では，敵対的生成ネットワーク (GAN) を用いて，End-to-End 音声認識の出力ラベル系列がより自然に生成されるよう補正する手法が提案されている．より具体的には，Criticizing Language Model (CLM) と呼ばれる音声認識結果と実際のテキストデータの識別器と，End-to-End 音声認識（生成器と見なされる）を交互に学習することにより上記の補正を実現する．通常の GAN と同様，本手法は音声・書き起こしのペアデータを必要とせずに音声認識性能を向上することができる（もちろん End-to-End 音声認識自体の学習には十分なペアデータが必要である）．実験には Libirspeech の 100 時間ペアデータサブセットと 360/860 時間相当の書き起こしを用いた半教師あり学習タスクが用いられた．100 時間ペアデータのみを用いたモデル及び，360/860 時間相当の書き起こしを用いて学習された言語モデルを併用する従来手法から，相対的に 10 %以上の誤り削減を達成している．（増村，渡部）

4. 音声合成

リアルタイム波形生成モデルと End-to-End 音声合成に向けた取り組みについて紹介する．

4.1 リアルタイム波形生成モデル

2016 年の WaveNet の登場以来，波形直接生成モデルの研究が盛んに行われている．本 ICASSP でも，音声合成はもちろん，AASP のセッションにおいては，符号化や帯域拡張，また，音声認識での音声データ拡張にも WaveNet や FFTNet を用いた発表があった．本項では，従来の自己回帰型 WaveNet での生成時間問題を解決した，リアルタイム波形生成モデル 3 件を報告する．

1) 自己回帰モデル　1 件目は，LPCNet [32] と呼ばれる方式で，WaveRNN の発展系モデルである．WaveRNN は過去の波形サンプルから次の波形サンプルをリアルタイムに推定する自己回帰モデルであるが，LPCNet では過去の波形サンプルとその線形予測係数から次のサンプルの「予測残差」を推定するモデルである．予測残差を用いることにより，μ-law 量子化誤差に頑健になり，モデルサイズが小さい場合でも WaveRNN(μ-law 量子化版) よりも高品質な生成をリアルタイムで実現できる．WaveRNN と同様，スパース化することにより，モバイル端末でもリアルタイムな生成が可能となる．学習 (TensorFlow) および生成 (C 言語) のソースコードが公開されていることも重要である[*3]．

[*2] http://research.snips.ai/datasets/keyword-spotting

[*3] https://github.com/mozilla/LPCNet/

2) パラレル生成型モデル　後の 2 件は，LPCNet のような自己回帰モデルではなく，Parallel WaveNet のように全ての波形を同時に生成するパラレル生成型モデルである．Neural source-filter (NSF) [36] は，従来のソースフィルタボコーダで用いられる基本周波数 f_o とメルケプストラムを入力とし，f_o に対応する高調波を含んだ正弦波とホワイトノイズから多段の変数変換によって音声波形を出力するモデルである．また，WaveGlow [28] は，Flow 型画像生成モデル Glow のアフィン変換部分に WaveNet を導入し，メルスペクトログラムとホワイトノイズを入力とし，可逆演算可能な多段の変数変換によって音声波形を生成可能なモデルである．NFS では，周波数領域のパワーロスのみで学習が可能であり，WaveGlow では，学習時は音声波形を入力としてホワイトノイズを出力するモデルを学習し，生成時はその逆変換によりホワイトノイズから音声を出力する革新的モデルである．Parallel WaveNet は自己回帰型の教師モデルを学習する必要があったが，これらのモデルはパラレル生成モデルを直接学習できることが最大のポイントである．両モデル共に，自己回帰型 WaveNet と同等の高品質な音声波形を GPU 演算によりリアルタイムで生成可能である．それぞれソースコードが公開されている[*4] [*5]．

4.2 sequence-to-sequence 音声合成モデル

音声合成モデルとしては，End-to-End 音声合成に向けたモデルとして提案された，attention(注意機構) を用いた sequence-to-sequence モデルの応用が主なトピックであった．特に Tacotron および Tacotron2 のモデル構造を用いた研究が多く見られた．Yasuda らの研究では Tacotron の日本語音声合成への有効性が検討された [40]．英語の End-to-End 音声合成では，テキストの文字列から直接単語の発音やストレスを推定できることが報告されているが，日本語の場合，漢字など文字の多様性の問題やアクセントがテキスト中に明記されないなどの問題があり，End-to-End モデルを学習することは困難である．この研究では音素とアクセント型の系列を入力として学習するネットワークを提案している．実験ではアクセント型を入力に加えることで MOS の値が上昇することが示されたが，一方で従来法の LSTM-RNN に基づく音声合成の性能には至っていないことが報告された．またこの研究では，並列計算が可能であり長時間の情報を考慮した変換の可能な self-attention の利用が検討され MOS 値が向上するという結果を得た．

機械翻訳で提案された self-attention の音声合成への適用は文献 [22], [39] でも検討された．文献 [22] では，中国語のテキストが与えられたときの韻律語や単語の境界の推定において self-attention が有効であることを示している．self-attention は長期間の情報を考慮できるが，位置情報が構造に含まれないため，隣接する音素やフレームの情報を重要視できないという問題点がある．そこで文献 [39] では，self-attention に明示的に相対位置情報を加えることで，入力に絶対位置情報を与える場合に比べ，合成音声の品質が向上することを示している．

また，End-to-End 音声合成の学習には，クリーンな音声とその書き起こしが大量に必要になるという問題点がある．文献 [3] では，Tacotron のエンコーダとデコーダを，テキストと音声のペアを用いずに，個別に事前学習する手法が提案されている．Tacotron を一から学習した場合と比べ，データが数十分程度であれば品質が大きく向上することを示している．(郡山、岡本)

5. Human Language Technology (HLT)

音声翻訳に用いられる手法と音声認識における言語処理に関して述べる．

5.1 音声翻訳

音声翻訳に関しては，音声認識と機械翻訳を接続したカスケード型の手法ではなく，翻訳元言語の音声と翻訳先言語のテキストのみによる，教師なしの手法が提案されている [4]．本手法では，翻訳元音声の speech embedding と翻訳先テキストの word embedding をそれぞれ独立に学習したのち，これらの embedding space 間の対応関係を構成する．入力音声に対しては，セグメンテーション後の各音声区間について speech embedding への変換および word embedding への写像を行い，この空間上で写像先と近接する単語を翻訳候補として得る．最近傍の単語が正しいとは限らないので，候補の単語に言語モデルを適用してスコアを計算することで翻訳結果を決定する．また，あらかじめノイズを加えたデータにより作成した denoising autoencoder を用いて，翻訳結果の訂正を行う．LibriSpeech の英仏翻訳における実験では，教師あり学習によるカスケード型のシステムとおおむね同等の BLEU 値を得ている．

[10] では End-to-End 音声翻訳モデル学習のために以下の二つの半教師あり data augmentation 手法を提案している．(1) 事前学習された機械翻訳モデルを用いて翻訳元言語テキストを翻訳先言語テキストに変換する．(2) 事前学習された音声合成モデルを用いて翻訳元言語モデルテキストを翻訳元言語音声データに変換する．これら二つの手法で増強された翻訳元音声データと翻訳先言語テキストのペアデータを用いて End-to-End 音声翻訳モデルの学習を行う．英語音声からスペイン語テキストへの自然発話音声翻訳タスクにおいて，in-domain データに対する評価で 60 弱，out-of-domain データに対する評価で 26 強と，非常に高い BLEU 値を達成している．

5.2 音声認識における言語処理

近年では音声認識結果単語の信頼度推定も NN を用いて行われるようになってきている．[18] では，従来，1-best 音声

[*4] https://github.com/nii-yamagishilab/project-CURRENNT-scripts
[*5] https://github.com/NVIDIA/waveglow

認識結果に対して行っていた bidirectional RNN (BiRNN) を用いた信頼度推定を，confusion network (CN) 及び lattice 形式の音声認識結果に対しても行えるようにしている．その際の主な課題は，音声認識仮説の分岐点における隠れ状態ベクトルの伝搬をどのように行うかである．本研究ではこの課題を標準的な forward-backward algorithm と同様の方法で効率的に解き，実験によりその有効性を示している．1-best 音声認識結果は誤認識単語を多く含む可能性があり，N-best は音声認識結果の表現形式としては冗長である．このため音声認識結果の効率的な表現形式である CN 又は lattice 上の単語の信頼度を高精度に推定できれば，後段の音声認識アプリケーションの品質を向上させられる．

本 ICASSP では句読点挿入 (punctuation) が複数報告されていた．音声認識結果に対して句読点を自動挿入する手法は以前から検討があるが，今回は attention 機構を用いた枠組みが 2 件報告されていた [13], [41]. [41] では multi-head self-attention による層を重ねて，[13] では BiRNN の各層に multi-head attention をそれぞれ加えて，NN を構成している．いずれも IWSLT の英語 TED 講演データセットを使用して評価しており，これまでの RNN による手法に対して性能の改善を（特に [41] が）実現している．

言語モデルに関しては，今回 HLT としての報告は多くなく，また特定の傾向もみられなかったが，ニューラル言語モデルについて知識蒸留 [30]，汎化性能のための活性化関数の拡張 [16]，RNN モデルにおける学習アルゴリズム (SGD) の拡張 [42] などが報告されていた．(秋田、小川)

6. 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 3 セッション，ポスター 2 セッションで合計 53 件の発表が行われた．

6.1 話者エンベディング

X-vector をはじめとする NN を用いた話者エンベディングに関しては，学習時の工夫による精度の改善が、同時多発的に報告された [2], [19], [38], [43]．これらは，N 名の話者を識別する NN を学習し，その softmax 層に入力されるベクトルを話者エンベディングとして用いる．提案手法では，学習時にマージンを設定することで，同じ話者が空間のより狭い部分に集中するよう学習している．この手法はもともと顔画像認識の分野で提案され高い効果が知られており [5], [34] ，本 ICASSP において、話者認識にも有効であることが明らかになった．また [44] では，話者エンベディング法について，アクティベーション関数や構造の違いによる性能の違い，そして性能改善に不可欠なデータ拡張の方法など様々なトリックが紹介されている．

6.2 話者認識のアプリケーション

例年あまり見られないタイプの研究として，話者認識のアプリケーションに関する発表が注目されていた [33]．本研究では，ユーザがウェブインタフェースを通じて話者認識システムを簡単に体験できるプラットフォームが提案されている．これはユーザが自分の音声をアップロードすると，ユーザに似た話者を YouTube から検索し，類似度が最も高い順に 5 名表示するというシステムである．バックエンドは旧来の i-vector と PLDA に基づくオーソドックスな手法だが，テスト運用の結果では精度，速度，使用感全てにおいて使用者から高評価を得ており，話者認識手法の可能性を感じさせる研究であった．

6.3 Speaker spoofing

Speaker spoofing に関する研究も例年通り多くの発表が見られた．本分野特有の特徴量として，話者認識ではあまり用いられない位相情報が replay attack の検出に有効であることを示す発表が複数あった [21], [25]．[21] ではラッピングの問題を相対位相を用いることで解消し, さらに適応的な重み付けを行うことで，従来のパワーに基づく特徴量と相補的な性能が得られることを示した．一方，[25] では，瞬時周波数を離散コサイン変換することでラッピングの問題を解消している．さらに，生音声の位相特徴量を事前にディクショナリとして学習しておき，評価音声から生音声の成分を取り除くことで，再生音声特有の歪成分を強調できることを示した．(俵，鈴木)

参考文献

[1] Alvarez, R. and Park, H.: End-to-end Streaming Keyword Spotting, *ICASSP*, pp. 6336–6340 (2019).
[2] Bhattacharya, G., Alam, J. and Kenny, P.: Adapting end-to-end neural speaker verification to new languages and recording conditions with adversarial training, *ICASSP*, pp. 6041–6045 (2019).
[3] Chung, Y., Wang, Y., Hsu, W., Zhang, Y. and Skerry-Ryan, R.: Semi-supervised Training for Improving Data Efficiency in End-to-end Speech Synthesis, *ICASSP*, pp. 6940–6944 (2019).
[4] Chung, Y.-A., Weng, W.-H., Tong, S. and Glass, J.: Towards Unsupervised Speech-to-text Translation, *ICASSP*, pp. 7170–7174 (2019).
[5] Deng, J., Guo, J., Xue, N. and Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition, *CVPR*, pp. 4690–4699 (2019).
[6] Dong, L., Wang, F. and Xu, B.: Self-Attention ALigner: A latency-Control End-to-End Model For ASR using Self-Attention Network and Chunk-Hopping, *ICASSP*, pp. 5656–5660 (2019).
[7] Han, C., Luo, Y. and Mesgarani, N.: Online Deep Attractor Network for Real-time Single-channel Speech Separation, *ICASSP*, pp. 361–365 (2019).
[8] Heymann, J., Drude, L., Haeb-Umbach, R., Kinoshita, K. and Nakatani, T.: Frame-online DNN-WPE dereverberation, *IWAENC*, pp. 466–470 (2018).
[9] Heymann, J., Drude, L., Haeb-Umbach, R., Kinoshita, K. and Nakatani, T.: Joint Optimization of Neural Network-based WPE Dereverberation and Acoustic Model for Robust Online ASR, *ICASSP*, pp. 6655–6659 (2019).
[10] Jia, Y., Johnson, M., Macherey, W., Weiss, R. J., Cao,

Y., Chiu, C.-C., Ari, N., Laurenzo, S. and Wu, Y.: Leveraging Weakly Supervised Data to Improve End-to-end Speech-to-text Tranlsation, *ICASSP*, pp. 7180–7184 (2019).

[11] Kanda, N., Boeddeker, C., Heitkaemper, J., Fujita, Y., Horiguchi, S., Nagamatsu, K. and Haeb-Umbach, R.: Guided Source Separation Meets a Strong ASR Backend: Hitachi/Paderborn University Joint Investigation for Dinner Party ASR, *INTERSPEECH* (2019 予定).

[12] Kanda, N., Fujita, Y., Horiguchi, S., Ikeshita, R., Nagamatsu, K. and Watanabe, S.: Acoustic Modeling for Distant Multi-talker Speech Recognition with Single-and Multi-channel Branches, *ICASSP*, pp. 6630–6634 (2019).

[13] Kim, S.: Deep Recurrent Neural Networks with Layer-wise Multi-head Attentions for Punctuation Restoration, *ICASSP*, pp. 7280–7284 (2019).

[14] Kinoshita, K., Delcroix, M., Kwon, H., Mori, T. and Nakatani, T.: Neural Network-Based Spectrum Estimation for Online WPE Dereverberation., *INTERSPEECH*, pp. 384–388 (2017).

[15] Kumatani, K., Minhua, W., Sundaram, S., Strom, N. and Hoffmeister, B.: Multi-Geometry Spatial Acoustic Modeling for Distant Speech Recognition, *ICASSP*, pp. 6635–6639 (2019).

[16] Lam, M. W. Y., Chen, X., Hu, S., Yu, J., Liu, X. and Meng, H.: Gaussian Process LSTM Recurrent Neural Network Language Models for Speech Recognition, *ICASSP*, pp. 7235–7239 (2019).

[17] Leroy, D., Coucke, A., Lavril, T., Gisselbrecht, T. and Dureau, J.: Federated Learning for Keyword Spotting, *ICASSP*, pp. 6341–6345 (2019).

[18] Li, Q., Ness, P., Ragni, A. and Gales, M.: Bi-directional Lattice Recurrent Neural Networks for Confidence Estimation, *ICASSP*, pp. 6755–6759 (2019).

[19] Li, R., Li, N., Tuo, D., Yu, M., Su, D. and Yu, D.: Boundary Discriminative Large Margin Cosine Loss for Text-independent Speaker Verification, *ICASSP*, pp. 6321–6325 (2019).

[20] Liu, A. H., Lee, H. and Lee, L.: Adversarial Training of End-to-end Speech Recognition Using a Criticizing Language Model, *ICASSP*, pp. 6176–6180 (2019).

[21] Liu, M., Wang, L., Dang, J., Nakagawa, S., Guan, H. and Li, X.: Replay Attack Detection Using Magnitude and Phase Information with Attention-based Adaptive Filters, *ICASSP*, pp. 6201–6205 (2019).

[22] Lu, C., Zhang, P. and Yan, Y.: Self-attention Based Prosodic Boundary Prediction for Chinese Speech Synthesis, *ICASSP*, pp. 7035–7039 (2019).

[23] Luo, Y. and Mesgarani, N.: Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation, *IEEE/ACM Trans. on ASLP* (2019).

[24] Menne, T., Schlüter, R. and Ney, H.: Investigation into Joint Optimization of Single Channel Speech Enhancement and Acoustic Modeling for Robust ASR, *ICASSP*, pp. 6660–6664 (2019).

[25] Mohammad Rafi B, S. and Sri Rama Murty, K.: Importance of Analytic Phase of the Speech Signal for Detecting Replay Attacks in Automatic Speaker Verification Systems, *ICASSP*, pp. 6306–6310 (2019).

[26] Morrone, G., Bergamaschi, S., Pasa, L., Fadiga, L., Tikhanoff, V. and Badino, L.: Face Landmark-based Speaker-Independent Audio-Visual Speech Enhancement in Multi-Talker Environments, *ICASSP*, pp. 6900–6904 (2019).

[27] Parnia, B., Zeyer, A., Schlüter, R. and Ney, H.: On Using 2D Sequence-to-sequence Models for Speech Recognition, *ICASSP*, pp. 5671–5675 (2019).

[28] Prenger, R., Valle, R. and Catanzaro, B.: WaveGlow: A Flow-based Generative Network for Speech Synthesis, *ICASSP*, pp. 3617–3621 (2019).

[29] Salazar, J., Kirchhoff, K. and Huang, Z.: Self-Attention Networks for Connectionist Temporal Classification in Speech Recognition, *ICASSP*, pp. 7115–7119 (2019).

[30] Shi, Y., Hwang, M.-Y., Lei, X. and Sheng, H.: Knowledge Distillation for Recurrent Neural Network Language Modeling with Trust Regularization, *ICASSP*, pp. 7230–7234 (2019).

[31] Tzinis, E., Venkataramani, S. and Smaragdis, P.: Unsupervised deep clustering for source separation: Direct learning from mixtures using spatial information, *ICASSP*, pp. 81–85 (2019).

[32] Valin, J.-M. and Skoglund, J.: LPCNet: Improving Neural Speech Synthesis Through Linear Prediction, *ICASSP*, pp. 5826–7830 (2019).

[33] Vestman, V., Soomro, B., Kanervisto, A., Hautamäki, V. and Kinnunen, T.: Who do I sound like? showcasing speaker recognition technology by YouTube voice search, *ICASSP*, pp. 5781–5785 (2019).

[34] Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z. and Liu, W.: Cosface: Large margin cosine loss for deep face recognition, *CVPR*, pp. 5265–5274 (2018).

[35] Wang, X., Sun, S., Shan, C., Hou, J., Xie, L., Li, S. and Lei, X.: Adversarial Examples for Improving End-to-end Attention-based Small-footprint Keyword Spotting, *ICASSP*, pp. 6366–6370 (2019).

[36] Wang, X., Takaki, S. and Yamagishi, J.: Neural source-filter-based waveform model for statistical parametric speech synthesis, *ICASSP*, pp. 5916–5920 (2019).

[37] Wang, Z.-Q., Tan, K. and Wang, D.: Deep learning based phase reconstruction for speaker separation: A trigonometric perspective, *ICASSP*, pp. 71–75 (2019).

[38] Xie, W., Nagrani, A., Chung, J. S. and Zisserman, A.: Utterance-level Aggregation For Speaker Recognition In The Wild, *ICASSP*, pp. 5791–5795 (2019).

[39] Yang, S., Lu, H., Kang, S., Xie, L. and Yu, D.: Enhancing Hybrid Self-attention Structure with Relative-position-aware Bias for Speech Synthesis, *ICASSP*, pp. 6910–6914 (2019).

[40] Yasuda, Y., Wang, X., Takaki, S. and Yamagishi, J.: Investigation of Enhanced Tacotron Text-to-speech Synthesis Systems with Self-attention for Pitch Accent Language, *ICASSP*, pp. 6905–6909 (2019).

[41] Yi, J. and Tao, J.: Self-attention Based Model for Punctuation Prediction Using Word and Speech Embeddings, *ICASSP*, pp. 7270–7274 (2019).

[42] Yu, J., Lam, M. W. Y., Chen, X., Hu, S., Liu, S., Wu, X., Liu, X. and Meng, H.: Recurrent Neural Network Language Model Training Using Natural Gradient, *ICASSP*, pp. 7260–7264 (2019).

[43] Yu, Y.-Q., Fan, L. and Li, W.-J.: Ensemble Additive Margin Softmax for Speaker Verification, *ICASSP*, pp. 6046–6050 (2019).

[44] Zeinali, H., Burget, L., Rohdin, J., Stafylakis, T. and Cernocky, J. H.: How to improve your speaker embeddings extractor in generic toolkits, *ICASSP*, pp. 6141–6145 (2019).

[45] Zhao, Y., Li, J., Wang, X. and Li, Y.: The Speech-Transformer for Large-Scale Mandarin Chinese Speech Recognition, *ICASSP*, pp. 7095–7099 (2019).