Vol.2020-SLP-131 No.6
2020/2/13

国際会議Interspeech2019報告

秋田 祐哉[1]　岡本 拓磨[2]　塩田 さやか[3]　俵 直弘[4]　角尾 衣未留[5]　増村 亮[4]

概要：2019年9月15日から19日にかけて，オーストリア・グラーツにてISCA主催の国際会議Interspeech 2019が開催された．Interspeechは ICASSP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．本稿では音声言語処理に関する分野に注目して，Interspeech2019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. はじめに

2019年9月15日から19日にかけて，オーストリア・グラーツにてISCA主催の国際会議Interspeech 2019 (The 20th Annual Conference of the International Speech Communication Association) が開催された．Interspeechは ICASSP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，Interspeechは音声言語処理をより幅広く扱う特色がある．2019年の論文の投稿数は2,180件あり，1,855件が査読の対象となった．これらは過去のInterspeechにおける最多件数からそれぞれ20%の増加とのことである．このうち採択されたのは914件で，採択率は49.3%であった[*1]．本稿では音声言語処理に関する分野に注目して，Interspeech2019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[*2]．（秋田）

2. 音声認識

近年，入力特徴量から直接文字列を出力するEnd-to-end (E2E) の手法が活発に議論されているが，その性能は従来のハイブリッド音声認識に匹敵する程に向上し，今年のInterspeechではE2E音声認識とハイブリッド音声認識の直接的な性能比較が注目を集めた．また，トレンドとして学習データを合成したり加工処理を行ったりするオーグメンテーションに関する発表が散見された．さらには，E2Eが実用レベルの認識性能に達してきていることもあり，オンライン処理化などの実用化に向けた研究も多く見られた．

2.1 End-to-end音声認識とハイブリッド音声認識

文献 [1] ではRWTHのチームによるLibrispeechデーターセットでの両手法の比較実験に関して報告されている．ハイブリッド手法の様々なセットアップのうちでも，GMM/HMMとDNN/HMMや識別学習の有無，HMM状態の音素ツリーによるクラスタリング粒度などについても慎重に比較したうえで最高性能のものを選び出している．その際の音響モデルは双方向1000ユニットずつの6層のbidirectional LSTMを音素ツリー (CART) ラベルで学習し，声道長正規化 (VTLN) と話者適応を適用し，識別学習を行っている．これを単語単位の4-gram言語モデルと2層4096ユニットのLSTM言語モデルを組み合わせてデコードを行い，さらに96層のTransformer言語モデルでリスコアリングを行っている．この結果，test cleanでWER 2.3%，test otherでWER 5.0%と，当時のstate-of-the-art (SOTA) の性能を実現した．

一方，E2E音声認識はencoder decoder attentionモデルを利用し，それぞれ1024ユニットの6層LSTMを利用している．これらはバイト対符号化 (BPE) されたユニットのラベルで学習されている．こちらに対しても24層のTransformer言語モデルが適用され，test clean 2.8%，test other 9.3%とE2Eでの当時のSOTA性能の2.5%，5.8%(後述する文献 [2]) には至らなかったが，オーグメンテーション無しE2Eの条件ではSOTAであった．まだハイブリッド音声認識のほうが認識性能は高いことを示した．

この研究は引き続き，Transformer E2Eモデルも含めてさらなる評価が行われ [3]，文献 [4] とともに複数タスクセットでの比較実験がASRU2019で報告されている．

2.2 SpecAugment

文献 [2] はarXivに掲載された時点でE2Eが当時のハイブリッド音声認識の性能を超えてSOTAを達成したことで

[1] 京都大学
[2] 情報通信研究機構
[3] 首都大学東京
[4] 日本電信電話株式会社
[5] ソニー株式会社

[*1] これらの数値は会議の概要集および開会式での説明による．
[*2] 著者は50音順である．以降，担当した範囲ごとに末尾にその担当者を示す．

話題になっていたが，“How you train is as important as what you train.” から始まるプレゼンテーションも人が部屋から溢れんばかりの盛況であった．

アイディアは非常にシンプルなもので，入力のスペクトルを直接画像処理的に加工することによって入力特徴量のバリエーションを増やし，頑健性を向上させるというものであった．オーグメンテーションは 3 つの処理を組み合わせて行われる．一つは時間軸の中心が左，もしくは右に移動するようにスペクトル伸縮させるタイムワーピングである．この画像処理によって疑似的に話速の変化を与えている．二つめは周波数マスキングである．ランダムに周波数を選び，0 平均の正規分布に従う幅のマスクをかける．この処理はスペクトル値の平均が 0 になるように正規化されたのちに 0 を埋めることによってマスキングされるので，スペクトルの平均値で埋めていることと等価の処理になっている．三つめは，時間マスキングで，0 から上限値までの一様分布で幅が決定され，マスキングが行われる．

これらを組み合わせたオーグメンテーションをイテレーション毎にランダムに変化させることによって学習を行い，上述の Librispeech タスクで当時の SOTA の test clean 2.5%，test other 5.8%を達成した．このためには大きなモデルサイズのセットアップで 3 週間以上の学習をおこなっている．

このオーグメンテーション手法は文献 [4] など，多くの研究で効果が見られ，汎用的に利用可能であることが示されている．

2.3 オンライン End-to-end 音声認識

End-to-end 音声認識として提案されているものの多くは bidirectional LSTM やアテンション機構を利用しており，発話全体を必要としていたが，音声認識をストリームで処理したい場合，これらの手法では実用的ではなかった．この問題に対し，数々のオンライン化のアプローチが提案された．

Bidirectional LSTM を単方向に置き換えるアプローチとして，文献 [5] ではハイブリッド音声認識で一定の性能を示している TDNN-LSTM モデルを検討し，さらに Linear 層と LSTM 層を入れ替えた time-delay LSTM (TDLSTM) モデルと，複数の独立した LSTM を並列して利用する parallel time-delay LSTM (PTDLSTM) を提案し，PTDLSTM が最も効果があったと報告している．Librispeech タスクにおいて test clean 5.7%，test other 16.9%の WER であった．

アテンションの逐次化も多くみられ，文献 [6] では固定窓長で単方向にアテンションを逐次的に形成する monotonic chunk-wise attention (MoChA) に対して，適応的に窓長を変化させる Adaptive MoChA (AMoChA) を提案している．また，MoChA に対しても演算式に変更を加え，学習が安定するような工夫をした stable MoChA (sMoChA) が提案された [7]．ここでは，オンライン処理可能な bidirectional LSTM としての latency controlled biLSTM が導入されている．一方で，シンプルなアテンションのメディアンや最大値から単方向に窓をシフトさせるアプローチも，文献 [8] で改めて比較されている．

文献 [9] では，逐次処理可能な手法としての RNN-T と，バッチ処理である LAS との性能のギャップを埋めるため，2 パスの演算とし，最初のパスで RNN-T を用いて認識結果を得た後に，次のパスで全体の情報を利用して LAS によってそれを補正するアプローチが提案された．

2.4 ドロップアウトを利用した半教師学習

近年は半教師学習もまた大きな関心を集めているが，ここでは特にハイブリッド音声認識の学習におけるユニークな手法を一つ取り上げておく．半教師学習において，finite-state transducer (FST) の教師ラティスを用いて lattice-free MMI 基準で学習する方法があるが，この場合最尤推定された尤度に偏った教師ラティスが生成されてしまう．この問題に対し，文献 [10] は異なるドロップアウトを適用して複数回試行することでベイズ推定を行い，偏りのない分布の教師ラティスを生成することで，半教師学習での性能を向上させた．このアプローチはニューラルネットワークベースの言語モデルにも適用でき，これを組み合わせることによって更に性能向上が得られている．また，End-to-end 音声認識にも汎用的に利用できる可能性が言及されている．（角尾）

3. 音声の自己教師あり表現学習

ラベル付きデータを用いることなく，ラベルなしデータのみから音声処理に有用な表現を獲得する自己教師あり表現学習の検討が増えてきている．文献 [11] では，ログメルフィルタバンク係数に対して LSTM をエンコーダに用いて，未来の情報を自己回帰に基づき予測するタスクを自己教師あり学習のタスクとする方法を提案している．評価実験では，LibriSpeech から表現学習を行い，音素認識と話者認証のタスクに適用する評価を行っており，代表的な従来手法である Contrastive Predictive Coding よりも良い表現学習を実現できることを報告している．

時間領域の波形情報に対する表現学習も検討されている．文献 [12] では，時間領域の波形情報に対して SincNet をエンコーダに用いて，対数パワースペクトルや MFCC，F0 や零交差率を予測するタスクを自己教師あり表現学習のタスクとする方法を提案している．評価実験では，LibriSpeech から表現学習を行い，話者認識，感情分類，音声認識のタスクに適用する評価を行っており，MFCC やログメルフィルタバンク係数を用いる場合と比較して高い性能を実現で

きることを報告している．文献 [13] では，先程と同様に時間領域の波形情報に対して SincNet をエンコーダに用いて，同一発話内の音声に対してエンコーダの出力が類似するように，そして異なる発話から取り出された音声に対してエンコーダの出力が類似しないように学習する方法を提案している．VoxCeleb1 を用いた評価実験において，提案手法に基づく表現学習を行ってから d-vector ベースの話者認証を学習することにより，高い性能を達成できることを報告している．

その他にも，文献 [14], [15], [16] 等において自己教師あり表現学習に注目した研究が検討されており，今後も注目すべき研究分野と言える．（増村）

4. 音声合成・声質変換

音声合成セッション (声質変換，歌声合成を含む) は，5 つのオーラルセッションと 4 つのポスターセッションがあり，大変盛況であった．研究動向としてはやはり，WaveNet から始まったニューラルボコーダによる音声波形生成ネットワークおよび Tacotron 2 の成功による End-to-end 音響モデルが主流であった．声質変換では，ノンパラレルデータを用いた研究が多かった印象がある．また今回は，3 年に 1 度の音声合成サテライトワークショップ SSW10 もウィーンで開催され，300 人以上のレジストレーションがあり，音声合成業界の活気がうかがえた．以下では，話者適応型リアルタイムニューラルテキスト音声合成 [17]，ユニバーサルニューラルボコーダ [18]，End-to-end テキスト音声合成の音響モデル [19], [20]，およびマルチリンガル・クロスリンガル音声合成 [21] についてを紹介する．

話者適応型リアルタイムニューラルテキスト音声合成 [17] では，SLT 2018 で発表された従来法である話者適応型テキスト音声合成は WORLD ボコーダのため音質の頭打ちがある問題に対して，ICASSP 2019 で発表されたリアルタイムニューラルボコーダ LPCNet を導入することにより，単一 CPU でもリアルタイムかつ高品質な音声合成を実現している．また，話者適応では，10 時間以上の単一話者で学習されたモデルに 20 分程度の新しい話者のデータを適用することにより，高品質な合成を実現できることも示されおり，デモ音声も公開されている[*3]．

ユニバーサルニューラルボコーダ [18] では，複数の話者のみならず複数の言語の音声データを用いて WaveRNN ボコーダを学習することにより，学習データには含まれていない話者はもちろんのこと，学習データに含まれていない言語の音声についても合成できるかについての検討が行われた．通常のニューラルボコーダは音声特徴量を直接アップサンプリング層に入力するのに対して，このモデルではアップサンプリング層の前段に双方向の GRU 層を追加することにより，時間変化を捉えた特徴量を扱っている (=時間方向に圧縮してより重要な情報にエンコードしている) 所がポイントである．約 15 万発話 (17 言語・74 話者) を用いることにより，学習データに含まれない言語の音声も原音と同等の品質で合成可能である．

End-to-end テキスト音声合成の音響モデルでは，学習データに含まれないドメインのテキストや長い発話に対する頑健性が重要であり，以下の 2 件はその頑健性を向上させる試みである．

文献 [19] では，Tacotron 型モデルはデコーダは過去の情報しか用いないのに対して，未来の情報も考慮した学習法を提案している．過去からデコードした系列と未来から逆方向にデコードした系列は理論的には一致することに着目し，双方向のデコーダの出力およびその隠れ状態を一致するように学習させることにより，未来の情報も考慮した学習法となるため，従来の Tacotron よりも高品質な合成を実現しており，デモ音声も公開されている[*4]．

文献 [20] では，単調型注意機構を発展させた段階的単調型注意機構を提案している．通常の Tacotron 等では滑らかな注意重みを採用しているため，注意機構予測が失敗すると発話が崩壊してしまう問題があり，単調型注意機構によりそれを防ぐことができる．しかし，音素をスキップしてしまう問題はこれでは解決できないため，1 段階しか遷移しない制約を加えた段階的単調型注意機構により，高品質かつ発話崩壊や音素の繰り返し，スキップの少ない頑健なモデルを実現しており，デモ音声も公開されている[*5]．

マルチリンガル・クロスリンガル音声合成 [21] では，複数話者の英語，スペイン語，中国語を 1 つのモデルで合成可能なマルチリンガル音声合成が提案されており，さらに，英語話者の音声でスペイン語や中国語も合成可能なクロスリンガル音声合成を実現しており，デモ音声も公開されている[*6]．合成時は言語コードと話者コードを指定することにより言語と話者を制御できるが，ポイントは言語成分と話者成分を如何に分離するかである．例えばある言語が 1 話者分のデータしかない場合，そのまま Tacotron を学習してしまうと言語と話者の分離は難しい．そのために，敵対的損失を加えており，エンコードされたテキスト入力から話者情報を差し引くことを行っている．また，韻律や録音環境の違いによるノイズ等を吸収するために，差分エンコーダも加えている．（岡本）

5. 話者認識

話者認識関連では，3 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，4 つのポスターセッションがあり計 105 件の発表があった．

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

[*4] https://vancycici.github.io/fbdecode/
[*5] https://dy-octa.github.io/interspeech2019/index.html
[*6] http://google.github.io/tacotron/publications/multilingual

5.1 話者ダイアライゼーション

スペシャルセッションの 1 つは話者ダイアライゼーションのコンペティションである DIHARD II に関するものであった．本セッションで報告されたシステムは，いずれも事前に切り出した固定長セグメントから得られた x-vector 等の話者表現に対し，クラスタリングを適用することで，同一話者の発話区間を同定する手法を導入したものであった．このアプローチにより得られる結果は，最初に切り出された発話単位に依存するため，特に話者交代点付近の性能が低下してしまう．そのため，文献 [22] では話者を状態とした HMM を推定し，フレーム単位でアラインメントを取ることで，結果を修正する手法がよく用いられるが，これを拡張して x-vector の空間でも HMM を作成し，この結果に基づきクラスタリングを行う手法が提案されている．論文中では DIHARD I のみの評価だったが，同機関による DIHARD II のトップシステムでも用いられたことが後の報告で示されている [23].

また，このようなクラスタリングに依らない新たな手法として，文献 [24] では音源分離に用いられてきた permutation invariant training (PIT) を話者ダイアライゼーションタスクに適用することで end-to-end でダイアライゼーションを行う手法が提案されている．2 話者の音声を重畳し作成した擬似的な対話データに適用した結果，従来の x-vector とクラスタリングに基づく手法よりも，高い性能を達成することを示した．投稿時点では BLSTM を使用していたため，実環境データに対しては従来手法よりも低い性能であったが，BLSTM の代わりに self-attention 機構を導入することで，従来手法よりも高い性能を達成できることが，後に同著者により報告されている [25].

多くのダイアライゼーション手法では音響情報のみが用いられるが，文献 [26] では言語情報も統合することで，コンテキストを考慮したダイアライゼーションシステムを提案している．本手法では音声認識結果から話者交代確率を推定し，これをセグメントをクラスタリングする際の話者類似度に反映させることで，語彙情報を考慮した話者ダイアライゼーションを実現する．話者数が未知の場合と既知の場合の両方において，提案手法は音響情報のみ用いた場合よりも高い性能を達成することが示されている．

5.2 話者認識

話者認識に関するコンペティションである NIST-SRE18 に関する報告が 3 件あった．SRE18 の主な特色は，学習時と評価時における言語，収録環境等のミスマッチであったため，これらドメインの違いの解決を試みるアプローチが多かった．SRE18 で最も高い性能を達成したシステムの詳細は文献 [27] で報告されている．本報告によると，現在広く使われる x-vector の抽出に用いられている time dilated neural network (TDNN) について，より広い受容野を持つように拡張した Extended TDNN と multi-head attention を組み合わせたモデルが，単体で最も良い性能を達成した．本システムを Resnet や factorized TDNN などを用いた複数のシステムと組み合わせ，さらに，開発セットに対し話者クラスタリングを適用し，得られた疑似話者ラベルでバックエンドの PLDA を再学習することで，SOTA の性能が得られたことが示されている．

また，近年では見られなかった傾向として，音源分離に用いられてきた手法を導入することで，混合音声を対象とした話者照合を行う手法が多く見られた．例えば，文献 [28] では話者分離手法である SpeakerBeam により分離した音声に対し i-vector を抽出することで，WSJ から作成した 2 話者混合音声に対する話者照合実験において，混合音声をそのまま用いた場合よりも高い性能が得られることが示されている．また，文献 [29] では，雑音環境下音声の話者照合法として，Ratio mask に基づく音声強調 DNN と，強調音声からの話者特徴抽出 DNN を連結し，話者認識ロスを用いて全体を最適化することで，話者照合に特化した強調音声と，その話者表現が得られることが示されている．バブルノイズや音楽を付加した雑音環境下音声に対する話者照合実験により，音声強調用の Denoising autoencoder と話者特徴抽出 DNN を別々に学習する従来の手法よりも，高い性能が得られることが示されている．（俵）

5.3 なりすまし検出

スペシャルセッションの 1 つは，話者照合への攻撃として問題視されているなりすまし攻撃の検出に関するコンペティション ASVSpoof2019 へ投稿されたシステムに関するポスターセッションと統括および議論のためのオーラルセッションで構成されていた．ASVSpoof は 2015 年から隔年で開催されており，2015 年は論理攻撃（Logistic attack; LA）と呼ばれる攻撃方法に，2017 年は物理攻撃（Physical attack;PA）と呼ばれる攻撃方法に着目して開催された．2019 年は LA および PA 両方の攻撃それぞれを想定したシナリオのデータベースが公開された．コンペティションの参加者は攻撃シナリオを選択し各自のシステムを構築することになっていた．全体的な分析や精度に関しては [30] にまとまっているが，一部抜粋すると，LA 攻撃を想定したシナリオに 48 チーム，PA 攻撃を想定したシナリオには 50 チームがシステムを投稿している．これまでのコンペティションでは識別器を工夫することよりも様々な特徴量を併用することで性能向上を目指すシステムがほとんどであったが，2017 年の結果において最も性能が良かったシステムが CNN を使った手法であったことから，2019 年の投稿システムの多くがニューラルネットワークをフロントエンドもしくはバックエンドに用い

る手法で投稿され，高い性能が得られたことが報告されていた．2017 年の最高性能を得たシステムは Light CNN を識別器に用いており，同じチームが 2019 年にもシステムを投稿し LA および PA どちらのシナリオについても 2 位を得ている [31]．2017 年からの主な変更点は Angular margin based softmax loss という損失関数を Light CNN に導入したことである．これは顔認識の分野で提案されたものである．また，ネットワーク構造も大規模化し，パラメータ数は前回の約 30 倍となっている．さらにテストセットのような unseen な環境に対する頑健性を上げるために Cepstral mean and variance normalization （CMVN）が有効であることが報告されている．LA シナリオの 1 位になったシステムに関しては論文が投稿されていなかったがプレスリリースが公開されていた [32]．PA シナリオの 1 位になったシステムに関しては APSIPA2019 で報告がされていた [33]．こちらに関しては，CQT に基づく群遅延を用いたこと及び ResNet を改良した ResNeWt を用いたことが特徴となっている．Interspeech のスペシャルセッションだけでなく ASRU2019 のスペシャルセッションにおいても ASVSpoof2019 の投稿システムに関する論文が投稿されている． （塩田）

参考文献

[1] Lüscher, C., Beck, E., Irie, K., Kitza, M., Michel, W., Zeyer, A., Schlüter, R. and Ney, H.: RWTH ASR Systems for LibriSpeech: Hybrid vs Attention, *Proc. Interspeech*, pp. 231–235 (2019).

[2] Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D. and Le, Q. V.: SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition, *Proc. Interspeech*, pp. 2613–2617 (2019).

[3] Zeyer, A., Bahar, P., Irie, K., Schlüter, R. and Ney, H.: Comparison of Transformer and LSTM Encoder Decoder Models for ASR, *Proc. ASRU*, pp. 8–15 (2019).

[4] Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., Someki, M., Soplin, N. E. Y., Yamamoto, R., Wang, X., Watanabe, S., Yoshimura, T. and Zhang, W.: A Comparative Study on Transformer vs RNN in Speech Applications, *Proc. ASRU*, pp. 449–456 (2019).

[5] Moritz, N., Hori, T. and Roux, J. L.: Unidirectional Neural Network Architectures for End-to-End Automatic Speech Recognition, *Proc. Interspeech*, pp. 76–80 (2019).

[6] Fan, R., Zhou, P., Chen, W., Jia, J. and Liu, G.: An Online Attention-Based Model for Speech Recognition, *Proc. Interspeech*, pp. 4390–4394 (2019).

[7] Miao, H., Cheng, G., Zhang, P., Li, T. and Yan, Y.: Online Hybrid CTC/Attention Architecture for End-to-end Speech Recognition, *Proc. Interspeech*, pp. 2623–2627 (2019).

[8] Merboldt, A., Zeyer, A., Schlüter, R. and Ney, H.: An Analysis of Local Monotonic Attention Variants, *Proc. Interspeech*, pp. 1398–1402 (2019).

[9] Sainath, T. N., Pang, R., Rybach, D., He, Y., Prabhavalkar, R., Li, W., Visontai, M., Liang, Q., Strohman, T., Wu, Y., McGraw, I., and Chiu, C.-C.: Two-Pass End-to-End Speech Recognition, *Proc. Interspeech*, pp. 2773–2777 (2019).

[10] Tong, S., Vyas, A., Garner, P. N. and Bourlard, H.: Unbiased Semi-Supervised LF-MMI Training using Dropout, *Proc. Interspeech*, pp. 1576–1580 (2019).

[11] Chung, Y.-A., Hsu, W.-N., Tang, H. and Glass, J.: An Unsupervised Autoregressive Model for Speech Representation Learning, *Proc. Interspeech*, pp. 146–150 (2019).

[12] Pascual, S., Ravanelli, M., Serra, J., Bonafonte, A. and Bengio, Y.: Learning Problem-agnostic Speech Representations from Multiple Self-supervised Tasks, *Proc. Interspeech*, pp. 161–165 (2019).

[13] Ravanelli, M. and Bengio, Y.: Learning Speaker Representations with Mutual Information, *Proc. Interspeech*, pp. 1153–1157 (2019).

[14] Schneider, S., Baevski, A., Collobert, R. and Auli, M.: wav2vec: Unsupervised Pre-training for Speech Recognition, *Proc. Interspeech*, pp. 3465–3469 (2019).

[15] Lian, Z., Tao, J., Liu, B. and Huang, J.: Unsupervised Representation Learning with Future Observation Prediction for Speech Emotion Recognition, *Proc. Interspeech*, pp. 3840–3844 (2019).

[16] Stafylakis, T., Rohdin, J., Plchot, O., Mizera, P. and Burget, L.: Self-Supervised Speaker Embeddings, *Proc. Interspeech*, pp. 2863–2867 (2019).

[17] Kons, Z., Shechtman, S., Sorin, A., Rabinovitz, C. and Hoory, R.: High Quality, Lightweight and Adaptable TTS using LPCNet, *Proc. Interspeech*, pp. 176–180 (2019).

[18] Lorenzo-Trueba, J., Drugman, T., Latorre, J., Merritt, T., Putrycz, B., Barra-Chicote, R., Moinet, A. and Aggarwal, V.: Towards Achieving Robust Universal Neural Vocoding, *Proc. Interspeech*, pp. 181–185 (2019).

[19] Zheng, Y., Wang, X., He, L., Pan, S., Soong, F. K., Wen, Z. and Tao, J.: Forward-backward Decoding for Regularizing End-to-end TTS, *Proc. Interspeech*, pp. 1283–1287 (2019).

[20] He, M., Deng, Y. and He, L.: Robust Sequence-to-sequence Acoustic Modeling with Stepwise Monotonic Attention for Neural TTS, *Proc. Interspeech*, pp. 1293–1297 (2019).

[21] Zhang, Y., Weiss, R. J., Zen, H., Wu, Y., Chen, Z., Skerry-Ryan, R., Jia, Y., Rosenberg, A. and Ramabhadran, B.: Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-language Voice Cloning, *Proc. Interspeech*, pp. 2080–2084 (2019).

[22] Diez, M., Burget, L., Wang, S., Rohdin, J. and Černockỳ, J.: Bayesian HMM Based x-Vector Clustering for Speaker Diarization, *Proc. Interspeech*, pp. 346–350 (2019).

[23] Landini, F., Wang, S., Diez, M., Burget, L., Matějka, P., Žmolíková, K., Mošner, L., Plchot, O., Novotnỳ, O., Zeinali, H. et al.: BUT System Description for DIHARD Speech Diarization Challenge 2019, *arXiv preprint arXiv:1910.08847* (2019).

[24] Fujita, Y., Kanda, N., Horiguchi, S., Nagamatsu, K. and Watanabe, S.: End-to-End Neural Speaker Diarization with Permutation-Free Objectives, *Proc. Interspeech*, pp. 4300–4304 (2019).

[25] Fujita, Y., Kanda, N., Horiguchi, S., Xue, Y., Nagamatsu, K. and Watanabe, S.: End-to-End Neural Speaker Diarization with Self-Attention, *Proc. ASRU*, pp. 296–303 (2019).

[26] Park, T., Han, K., Huang, J., He, X., Zhou, B., Geor-

giou, P. and Narayanan, S.: Speaker Diarization with Lexical Information, *Proc. Interspeech*, pp. 391–395 (2019).

[27] Villalba, J., Chen, N., Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Borgstrom, J., Richardson, F., Shon, S., Grondin, F. et al.: State-of-the-art Speaker Recognition for Telephone and Video Speech: the JHU-MIT Submission for NIST SRE18, *Proc. Interspeech*, pp. 1488–1492 (2019).

[28] Rao, W., Xu, C., Chng, E. S. and Li, H.: Target Speaker Extraction for Overlapped Multi-talker Speaker Verification, *Proc. Interspeech*, pp. 1273–1277 (2019).

[29] Shon, S., Tang, H. and Glass, J.: VoiceID Loss: Speech Enhancement for Speaker Verification, *Proc. Interspeech*, pp. 2888–2892 (2019).

[30] Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., Yamagishi, J., Evans, N., Kinnunen, T. and Lee, K. A.: ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection, *Proc. Interspeech*, pp. 1008–1012 (2019).

[31] Lavrentyeva, G., Novoselov, S., Tseren, A., Volkova, M., Gorlanov, A. and Kozlov, A.: STC Antispoofing Systems for the ASVspoof2019 Challenge, *Proc. Interspeech*, pp. 1033–1037 (2019).

[32] ID R&D: ID R&D Ranks First in Detecting Synthetic Speech in Global ASVspoof Challenge, https://www.idrnd.ai/id-rd-ranks-first-in-detecting-synthetic-speech-in-global-asvspoof-challenge/.

[33] Cheng, X., Xu, M. and Zheng, T. F.: Replay Detection using CQT-based Modified Group Delay Feature and ResNeWt Network in ASVspoof 2019, *Proc. APSIPA ASC*, pp. 540–545 (2019).