Vol.2020-MUS-127 No.7
Vol.2020-SLP-132 No.7
2020/6/7

会話音声を対象とした実環境向けニューラル音声認識技術

福田 隆[1,a)]

概要：音声認識はディープラーニングと時系列信号のモデリング技術の進展によって，飛躍的な性能改善を成し遂げた．音声認識研究には，計算機の能力を最大限に活かしたオフライン処理向けの技術と，リアルタイム性を考慮した比較的軽量なオンライン向けの技術がある．実用の観点からはどちらのアプローチも重要であるが，本稿では特に，オンライン向けのアプローチについて音響処理技術を中心に，知識蒸留法などのニューラルネットワーク技術や学習アルゴリズム，アーキテクチャー，データ拡張処理など実環境を考慮した音声認識研究を紹介する．

Neural Speech Recognition Techniques Targeting Conversational Speech for Real World Environments

1. はじめに

近年の音声認識技術は，ディープラーニングと時系列信号のモデリング技術の進展によって飛躍的な性能改善を成し遂げ，コールセンターにおけるエージェントアシストなどの応用で，音声認識が真に役立つ技術として世の中に浸透し始めた [1]．音声認識における音響処理研究の観点からは，HMM と GMM の組み合わせを基本とした統計モデルの学習アルゴリズムの検討が長らく続いた．最近では，それが HMM とニューラルネットワークの組み合わせによる音響モデリング技術の研究にシフトし，さらには音響モデル・発音辞書・言語モデルといった個々の要素技術を一括的に扱う End-to-End 音声認識に数多くの研究が見られるようになった．本稿では，近年のディープラーニング時代の音声認識技術について実用サービスの観点から興味深い研究を，主に音響的な処理を対象として俯瞰的に紹介する．

2. 音声認識フレームワーク

2.1 音声認識とディープラーニング

音声認識技術は，音声信号分析・音響モデリング・言語モデリング・発音辞書・デコーダに大別され，それぞれの要素技術について個別に研究が進められてきた．従来の音響モデリングの研究においては，音声データに含まれる時間的な変化情報を追随する HMM と，どのような時系列データが生成されやすいかを表現する確率モデル（GMM）を組み合わせたハイブリッドアプローチが典型的であった．ハイブリッドアプローチにおいて，GMM の代わりにニューラルネットワークを用いるという研究は古くから存在するが，ディープラーニングの登場によって，HMM とニューラルネットワークの組み合わせの研究が再加速した．そして近年では，LSTM を代表とするリカレント構造を持ったニューラルネットワークの利用により，音声データに含まれる時間的な特性とその音声らしさの確率を同時にモデリングするような End-to-End アプローチの研究が急激に増え，本稿執筆段階（2020 年 4 月）では，End-to-End アプローチの研究がハイブリッドアプローチと比べて相対的に多くなった．End-to-End アプローチの代表例として，Connectionist Temporal Classfication (CTC) [2] や Encoder-decoder モデルなどの Sequence to sequence 関連の基礎研究が積極的に行われている．End-to-End 音声認識の研究には，音響分析・音響モデル・言語モデル・発音辞書全てを単一のネットワークで表現するフレームワークも含まれるが，本稿で紹介する End-to-End アプローチは主に音響モデリングに関するものに限定する．

2.2 ハイブリッド vs. End-to-End アプローチ

ハイブリッドアプローチは，音声信号分析・音響モデ

[1] IBM 東京基礎研究所 / IBM Research AI
19–21, Nihonbashi Hakozaki-cho, Chuo-ku, Tokyo 103–8510, Japan
[a)] fukuda1@jp.ibm.com

表 1　Hybrid vs. End-to-End アプローチ（文献 [3] からの抜粋）

paper	model	label unit		LM	WER [%]			
		AM	LM		dev		test	
					clean	other	clean	other
Han et al. [3]	hybrid, seq. disc., single	CDp	word	RNN	3.0	8.8	3.6	8.7
	hybrid, seq. disc., ensemble				2.6	7.6	3.2	7.6
Zeghidour et al. [8]	end-to-end GCNN	chars	words	GCNN	3.2	10.1	3.4	11.2
Irie et al. [9]	end-to-end attention	Word Piece Model		LSTM	3.3	10.3	3.6	10.3
Zeyer et al. [5]		BPE			3.5	11.5	3.8	12.8
this work				None	4.3	12.9	4.4	13.5
				LSTM	2.9	8.9	3.2	9.9
				Transformer	**2.6**	**8.4**	**2.8**	**9.3**
	hybrid	CDp	word	4-gr	4.0	9.6	4.4	10.0
	hybrid, seq. disc.				3.4	8.3	3.8	8.8
				+ LSTM	**2.2**	**5.1**	**2.6**	**5.5**
				Transformer resc.	**1.9**	**4.5**	**2.3**	**5.0**
Park et. al. [10]	end-to-end attention/SpecAugment	Word Piece Model		LSTM	-	-	2.5	5.8

リング・言語モデリング・発音辞書構築を個別に行うため，それぞれの段階での作り込み作業が比較的容易であり，ターゲットの音声認識環境が事前にわかっている場合には，個々のモデルをある程度その環境に合わせ込むことができる．ただし，それぞれの設計について高い専門性が必要で，技術の横展開は必ずしも容易ではない．一方，End-to-End は学習用の入力音声データと，そのデータが何と発話しているかの書き起こしテキストの組み合わせを与えることによってネットワークの学習が進められるので，ハイブリッドアプローチを比べると「技術の手離れ」は良いと言える．しかし，認識対象の音響環境で十分な性能を引き出すためには，事実上多くのノウハウや工夫，ハイパーパラメータチューニングの努力が必要であり，End-to-End アプローチの元々の思想を鑑みると，基礎研究はまだまだ必要のように思われる．

さて，上述のハイブリッドアプローチと End-to-End アプローチはどちらが最高性能を引き出すことができるであろうか？これは多くの研究者にとって興味深い点であると思われる．ここでハイブリッドアプローチと End-to-End アプローチの比較を包括的に行った一例を紹介したい [3]．Lüscher らは，シーケンス学習までを取り入れた DNN/HMM ハイブリッドシステムと，注意機構を持つ Encoder-decoder フレームワークを End-to-End システムとして，LibriSpeech タスクにおいて様々な条件で両者を比較した．また同時に，ハイブリッドシステムにおいて GMM と DNN の比較と，言語モデルについても Transformer [4] を利用するなど様々なケースで比較を行っている．表 1 に文献 [3] から実験結果を一つ抜粋する．実験条件や表記の詳細は文献 [3] を参照されたい．ハイブリッドと End-to-End アプローチは基本となるフレームワークが異なるのでフェアな比較は難しいところであるが，この論文では DNN/HMM ハイブリッドシステムが注意機構付きの End-to-End システムと比較して，LibriSpeech の Clean セットで 15%，Other セットで 40%の相対的改善が得られることを示した．ハイブリッドシステムと End-to-End システムはそれぞれ一長一短があり，一概にどちらが良いとは言い難いが，多くの研究者にとって興味深い一つの比較結果であるように思われる．

3. 実環境向けニューラル音声認識

本節ではリアルタイム性を考慮したオンライン音声認識について，ハイブリッドと End-to-End アプローチの双方からいくつか研究例を紹介したい．オンライン向け音声認識の研究はハイブリッドアプローチが主流であったが，近年では End-to-End の研究においてもリアルタイム性を意識した研究例が増えている．ここでは，画像処理の分野で提案された知識蒸留法を用いた方法を中心に，筆者らの研究機関で取り組んでいる研究も紹介しつつ，ハイブリッドと End-to-End アプローチ，さらには会話音声を対象とした音声認識の問題点と研究の方向性，その他周辺技術の研究について簡単にまとめる．

3.1 ハイブリッド音声認識と知識蒸留法

知識蒸留法（knowledge distillation）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的としている [5]．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る．これを音声認識の枠組みに当てはめると以下の損失関数によって学習を進めることになる．

$$\mathcal{L}(\theta) = -\sum_i q_i \log p_i, \tag{1}$$

ここで q_i は教師ネットワークから生成されるソフトラベルであり，生徒ネットワークのためのターゲット，すなわち擬似ラベルとして役割を果たす．p_i は生徒ネットワークの

出力確率であり，i は音素コンテキストクラスを表すインデックスである．知識蒸留法では one-hot ベクトルをハードターゲットとして用いる代わりに，各学習サンプルに対して対立候補の音素クラスに非ゼロの確率を持つようなソフトラベルを利用する [6–11]．一般に，教師側には認識精度を重視した複雑かつ大規模なネットワークを用い，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．この枠組みでは，速度重視のコンパクトなネットワークに教師ネットワークのより強力な識別能力を転移できるため，リアルタイム性を重視したネットワークの構築において大きな役割を果たす．知識蒸留法は必ずしも同種のネットワーク間でのみ学習が実現されるわけではなく，例えば LSTM から畳み込み処理主体のネットワークへの知識蒸留など，基本構成の異なるネットワーク間での性能改善も報告されている [7,8,12]．

知識蒸留法に関しては，筆者らもいくつか関連の研究発表を行っている [12,13]．文献 [12] では学習時にのみ利用可能な情報を特権情報（privileged information [14]）として考え，狭帯域音声認識（サンプリング周波数 8kHz）の性能改善に，より豊富な情報を持つ広帯域ネットワーク（サンプリング周波数 16kHz）から生成されるソフトラベルを利用した一般化知識蒸留法を提案した．知識蒸留法の本来の考え方に基づくと，通常は生徒と教師ネットワークで同じサンプリング周波数の音声信号を利用したモデル化を行うところ，広帯域ネットワークでしか表現できない補足的な情報を，知識蒸留の枠組みで狭帯域モデルに組み込んだ部分に特徴がある．また，この考えを拡張して文献 [13] では，広帯域・狭帯域の両方をカバーする単一のニューラル音響モデルを構築する方法を提案した．通常，広帯域と狭帯域の音声信号を混ぜ合わせてモデルを学習すると，広帯域か狭帯域のどちらかの認識性能劣化が免れないが，知識蒸留法と特権情報の考え方をうまく利用して，広帯域と狭帯域のどちらの信号に対しても高精度に動作する音響モデリングを実現した．

上記の他にもいくつか知識蒸留法についての先行研究を紹介したい．Li らは知識蒸留法の枠組みを利用した音響ドメイン適応法を提案している [15]．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ノイズ混合音声の音響モデルへの適用および，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で実証した．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．同著者らはこの枠組みを End-to-End アプローチに発展させ，教師なしドメイン適応の改善を試みた [16]．

他方，従来の知識蒸留法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知識蒸留法の能力や，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知識蒸留法を提案した [17]．生徒側に存在しない音素状態クラスについては，教師の音素状態クラスから事後確率を推定している，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．

3.2 End-to-End 音声認識

End-to-End の研究においても，最近はオンライン音声認識での稼働を意識した研究例が増えている．例えば，End-to-End システムでよく用いられる双方向 LSTM（BLSTM: bidirectional LSTM）は，様々な実験条件で高い認識性能をもたらす可能性が示唆され，新しい学習アルゴリズムやトポロジーの提案など多くの研究が行われている．BLSTM は時間的に順方向と逆方向の入力を必要とするため，一般にはユーザの発声の終了（もしくは，短文などの区切りの良い単位）まで待ってからの処理開始となる．ユーザの発話終了を待ってからの認識開始となると処理に大きな遅延が生じるため，リアルタイム性を重視するオンライン音声認識には不向きである．解決策の一つとして，オンライン向けには逐次処理が可能な単方向 LSTM（ULSTM: unidirectional LSTM）の利用が考えられる．文献 [18,19] では，知識蒸留法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 BLSTM ネットワークから ULSTM に知識を転移する学習方法を提案している．これらの方法では，教師ネットワークである BLSTM と生徒ネットワーク側である ULSTM 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，End-to-End 音声認識学習の利点を損なわずに ULSTM の性能を改善することができる．このうち，文献 [18] 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．文献 [19] では，CTC が出力する事後確率のスパイクのタイミングを調整することによって，BLSTM から ULSTM への知識蒸留をより効果的に行う方法を提案している．そして文献 [20,21] では，CTC のフレームワークにおいて N-best 仮説を利用したシーケンスレベルの知識蒸留法ついて述べている．これらの知識蒸

留法の他，文献 [22] では，ULSTM と TDNN の枠組みを考慮した新しいネットワーク構成を提案し，遅延を 200ms 以下に抑えつつ認識性能も高い水準に維持できる可能性を示した．また，逐次処理の可能な ULSTM の利用ではなく，レイテンシーを小さく抑える工夫を組み込んだ BLSTM もいくつか提案されている [23].

本節の最後に，その他オンライン End-to-End 音声認識技術についていくつか紹介する．発音辞書や（理想的には）言語モデルを必要としない End-to-End 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 [24] 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 [25] では End-to-End 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 [25] 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 End-to-End 音声認識学習にも効果が期待できる方法として紹介した．

3.3 会話音声認識

本節では主にオフライン処理のケースを想定して，先行研究についてまとめる．複数人の会話音声を認識対象とする場合，認識性能や学習データの収集コストなどの課題もさることながら，会話音声特有の事象にも対応しなければならない．例えば電話音声の場合でも 2 名間の通話をモノラルで録音するケースがあり，単一話者の発声と比べて認識の難易度が格段に増す．（ただし，複数人会話の音声認識であっても，各話者ごとに音声を録音できるような状況であれば，音響処理の観点からは数多くの最高性能を成し遂げた先行研究が転用できることが多い [26,27].）

複数人会話でよく起こる音声認識上の問題は，まず発話の衝突があげられる．計算コストが問題にならないとするなら，マイクロホンアレイを用いた話者分離によって音声を個々の話者に切り分け，それぞれの話者についてさらに音声強調をするなどのアプローチが考えられる．例えば，End-to-End 音声認識においては，音声分離・強調・認識処理を全て一括に End-to-End で行う方法が提案された [28]．一方，電話音声のモノラル録音など複数マイクを扱えない状況や，マイクが複数利用可能であっても音声の回り込み現象によって複数話者の声が重なってしまうような場合は，ターゲット話者の音声に焦点を当てるバイナリマスクなどの音声強調手法の研究 [29,30] と共に，発話衝突（混合音声）そのものに頑健な音響モデリング手法の検討も望まれる [31]．また，複数人会話を単一マイクで集音する場合，話者の頻繁な切り替わりにより，CMS などの話者特性正規化処理において性能劣化が起こりやすい問題がある．これについては，混合音声の正規化手法の検討と共に，各話者の発話区間を同定する speaker diarization の研究も重要になってくる．近年ではニューラルネットワークに基づく diarization の研究例が増えてきている [32,33]．ここまでに紹介してきた音声認識技術の良し悪しは，一般にテストデータに対する単語誤り率（もしくは文字誤り率）で評価されることが多いが，会話音声のような自然発話の認識タスクでは単語誤り率の改善が必ずしも人間の感覚に一致しないことがある．文献 [34] では，人間の知覚を考慮した音声認識の評価方法を提案している．

さて，会話音声を対象とした音声認識システムでは，認識対象が数分から数十分の比較的長い音声に至ることが珍しくなく，このようなケースについては発話のチャンキングの研究が重要であると考えられ，発話区間検出 [35] や発話を言語的に意味のある単位で区切る手法などとのリンクが望まれる．例えば文献 [36] では，話者の息継ぎ位置を利用して発話を分割することの効果を音声認識のタスクで検証している．これまでの End-to-End 音声認識は 1 発話を数十秒程度のものと見なした研究が大部分で，コールセンター会話などの長時間録音音声をそのまま認識対象とするようなケースは少なかったが，そのような現実によく起こりうる長い録音データに対する検討も始まりつつある [37,38].

3.4 データ拡張

データ拡張（Data augmentation）の研究は単なる性能の底上げのみならず，音声コーパスの作成コストの観点から非常に重要な研究課題の一つと位置付けられ，様々な場面でその他技術と組み合わせて用いられている．音声認識におけるデータ拡張処理は，典型的にはクリーン音声に対する雑音付加やチャネル特性の重畳であるが，近年では話速変換もよく用いられるようになった [39]．またごく最近では，音声スペクトルにブロックタイプの人工マスクをかけるデータ拡張法も提案され，注意機構付き Encoder-decoder モデルの大幅な性能向上に役立つとして大きな注目を浴びている [40]．その他，2 つのデータサンプルに対して，データとそのラベルの双方を線形補間して

新しいデータとラベルのペアを生成する mixup 法なども注目され，音声認識での効果が検証されている [41, 42].

これらのデータ拡張法に対し，我々は子供音声と大人音声の音響的特性の違いに着目し，母音区間に焦点を当てたデータ拡張法を提案した [43]．子供音声はプライバシーや録音環境のコントロールの面で一般に収集が難しく，また喋り方も様々で低学年と高学年では大きく異なるため，大人音声に匹敵する認識性能を実現することが非常に難しい．提案法は子供音声を対象とした自由発話音声認識タスクにおいて，単純な話速変換法と比較して 2%の性能改善があることを確認した．また興味深いことに，成人女性の音声に提案法の変形を加えた音声が，子供音声認識のための学習データとして有効であることも見出した．

ところで，データ収集の困難さの観点から別の音声認識タスクを見ると，外国語なまりのアクセント（非ネイティブ，L2 話者）を対象とした音声認識も未だチャレンジングな認識対象と言える．これは発声に関してネイティブ話者との音響的差分が大きく，また言語の習熟度も話者や時期によって様々であるため，たとえ同一話者内であっても発音の揺らぎが非常に大きくなってしまうことが主たる要因である．この課題に対する最も効果的なアプローチは，可能な限り大量の外国語なまり音声を収集し，発話内容の書き起こしと共に学習データに含めることに他ならないが，前述のとおり非ネイティブ話者の音声は習熟度の問題もあり，ネイティブ話者（L1）に比べて，必ずしも話者のレベルに応じた発話を幅広く豊富に収集できる訳ではない．たとえ一般的な音声コーパスと同等サイズの非ネイティブ話者のデータが収集されたとしても，非ネイティブ話者の発声の音響的な広がりはネイティブ話者の音響空間と比較して極めて大きいため，非ネイティブ話者の学習データがもたらす性能改善効果には限りがあった．そこで筆者らは，非ネイティブ話者の音声に対する人工的データ拡張処理が，音声認識にどのような効果をもたらすかを実験的に検証した [44]．実験では，ラテンアメリカ英語とアジア英語の 2 種類を対象に，声質変換（声帯振動と声道特性の変換），話速変形，雑音付与によって複数のコピーを生成し，教師あり学習と教師なし学習の両方のシナリオで，人工的に生成されたデータの効果を確認した．特に，非ネイティブ話者専用の音響モデルをスクラッチから構築する場合にデータ拡張の効果が大きく，話速変換を用いたデータの生成によって，30%以上の相対的誤り削減が得られるケースもあることを報告した．

4. おわりに

本稿では，近年の実用サービスの観点から，ディープラーニングに基づく音声認識技術について簡単に紹介した．車載機器やコールセンターなど，利用環境をある程度コントロールできるような状況においては，人間の作業を効率よくアシストできるような性能にまで至ってきている．しかし，実用アプリケーションに対して十分な性能を実現するためにはモデルの構築に専門家の手が必要であり，現状では技術の手離れは決して良くない．実サービスを意識した技術は様々な視点からさらなる研究の進展が望まれる一方で，音声認識に関する高度な専門知識がなくても，データさえあれば実稼動可能な音声認識が実現できるようなモデリング手法の確立も重要な研究課題である．筆者らの研究機関では，本稿で紹介したもの以外にも多くの基礎研究 [45–48] を行っているが，紙面の都合上，文献の出典も含めて紹介は別の機会に譲りたい．

参考文献

[1] 長野徹, 吉田一星, 壁谷佳典, 岡原勇朗, 倉田岳人, 立花隆輝，“音声認識を用いたリアルタイムコンタクトセンターオペレータ支援”，電子情報通信学会論文誌，J102-D(9)，pp. 597–608，2019.

[2] K. Audhkhasi, G. Saon, Z. Tüske, B. Kingsbury, M. Picheny, *“Forget a Bit to Learn Better: Soft Forgetting for CTC-Based Automatic Speech Recognition”*, Proc. Interspeech, 2019.

[3] C. Lüscher, E. Beck, K. Irie, M. Kitza, W. Michel, A. Zeyer, R. Schüter, and H. Ney, *“RWTH ASR Systems for LibriSpeech: Hybrid vs Attention”*, Proc. Interspeech, 2019.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N.Gomez, Ł. Kaiser, and I. Polosukhin, *“Attention is all you need”*, Proc. NeurIPS, Long Beach, CA, USA, Dec. 2017.

[5] G. Hinton, O. Vinyals and J. Dean, *“Distilling the Knowledge in a Neural Network”*, arXiv:1503.02531v1, 2015.

[6] J. Li, R. Zhao, J. Huang and Y. Gong, *“Learning Small-Size DNN with Output-Distribution-Based Criteria”*, Proc. Interspeech, pp. 1910–1914, 2014.

[7] W. Chan, N. R. Ke and I. Lane, *“Transferring Knowledge from a RNN to a DNN”* Proc. Interspeech, 2015.

[8] K. J. Geras, A. Mohamed, R. Caruana, G. Urban, S. Wang, O. Aslan, M. Philipose, M. Richardson, and C. Sutton, *“Blending LSTMs into CNNs”*, ICLR Workshop, 2016.

[9] Z. Tang, D. Wang, and Z. Zhang, *“Recurrent neural network training with dark knowledge transfer”*, Proc. IEEE ICASSP, pp. 5900–5904, 2016.

[10] Y. Chebotar and A. Waters, *“Distilling knowledge from ensembles of neural networks for speech recognition”*, Proc. Interspeech, pp. 3439–3443, 2016.

[11] K. Markov and T. Matsui, *“Robust speech recognition using generalized distillation framework”*, Proc. interspeech, 2016.

[12] T. Fukuda, M. Suzuki, G. Kurata, S. Thomas, and B. Ramabhadran, *“Efficient Knowledge Distillation from an Ensemble of Teachers”*, Proc. Interspeech, pp. 3697–3701, 2017.

[13] T. Fukuda and S. Thomas, *“Mixed Bandwidth Acoustic Modeling Leveraging Knowledge Distillation”*, Proc. IEEE ASRU, 2019.

[14] V. Vapnik and R. Izmailov, *“Learning using privileged information: Similarity control and knowledge transfer”*, Machine Learning Research, Vol. 16, pp. 2023–

2049, 2015.
[15] J. Li, M. Seltzer, X. Wang, R. Zhao, and Y. Gong, *"Large-Scale Domain Adaptation via Teacher-Student Learning"* Proc. Interspeech, pp. 2386–2390, 2017.
[16] Z. Meng, J. Li, Y. Gaur, and Y. Gong, *"Domain Adaptation via Teacher-Student Learning for End-to-End Speech Recognition"*, Proc. IEEE ASRU, 2019.
[17] J. Wong and M. Gales, *"Sudent-teacher training with diverse decision tree ensembles"*, Proc. Interspeech, pp.117–121, 2017.
[18] S. Kim, M. Seltzer, J. Li, and R. Zhao, *"Improved Training for Online End-to-end Speech Recognition Systems"*, Proc. Interspeech, pp. 2913–2917, 2018.
[19] G. Kurata and K. Audhkhasi, *"Guiding CTC Posterior Spike Timings for Improved Posterior Fusion and Knowledge Distillation"*, Proc. Interspeech, 2019.
[20] R. Takashima, S. Li and H. Kawai, *"An Investigation of a Knowledge Distillation Method for CTC Acoustic Models"*, Proc. IEEE ICASSP, pp. 5809–5813, 2018.
[21] R. Takashima, S. Li and H. Kawai, *"Investigation of Sequence-level Knowledge Distillation Methods for CTC Acoustic Models"*, Proc. IEEE ICASSP, pp. 6156–6160, 2019.
[22] V. Peddinti, Y. Wang, D. Povey, and S. Khudanpur, *"Low Latency Acoustic Modeling Using Temporal Convolution and LSTMs"*, IEEE Signal Processing Letters, Vol. 25, Issue 3 , March 2018.
[23] S. Xue and Z. Yan, *"Improving latency-controlled BLSTM acoustic models for online speech recognition"*, Proc. IEEE ICASSP, 2017.
[24] R. Pang, T. Sainath, R. Prabhavalkar, S. Gupta, Y. Wu, S. Zhang and C. Chiu, *"Compression of End-to-End Models"* Proc. Interspeech, pp. 27–31, 2018.
[25] A. Zeyer, K, Irie, R. Schluter and H. Ney, *"Improved Training of End-to-end Attention Models for Speech Recognition"*, Proc. Interspeech, pp. 7–11, 2018.
[26] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L. Lim, B. Roomi, and P. Hall, *"English Conversational Telephone Speech Recognition by Humans and Machines"*, Proc. Interspeech, 2017.
[27] S. Thomas, M. Suzuki, Y. Huang, G. Kurata, Z. Tuske, G. Saon, B. Kingsbury, M. Picheny, T. Dibert, A. Kaiser-Schatzlein, and B. Samko, *"English Broadcast News Speech Recognition by Humans and Machines"*, Proc. IEEE ICASSP, 2019.
[28] X. Chang, W. Zhang, Y. Qian, J. Le Roux and S. Watanabe, *"MIMO-SPEECH: End-to-end Multi-channel Multi-speaker Speech Recognition"*, Proc. IEEE ASRU, 2019.
[29] D. Wang, *"On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis"*, Speech Separation by Humans and Machines, pp. 181–197, 2005.
[30] D. Wang and J. Chen, *Supervised Speech Separation Based on Deep Learning: An Overview"*, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, pp. 1702–1726, 2018.
[31] O. Ichikawa, T. Fukuda, G. Kurata, and S. J. Rennie, *"Factorial Modeling for Effective Suppression of Directional Noise"*, Proc. Interspeech, pp. 389–393, 2017.
[32] Y. Higuchi, M. Suzuki, and G. Kurata, *"Speaker Embeddings Incorporating Acoustic Conditions for Diarization"*, Proc. IEEE ICASSP, 2020.
[33] N. Kanda, S. Horiguchi, Y, Fujita, Y. Xue, K. Nagamatsu, and S. Watanabe, *"Simultaneous Speech Recognition and Speaker Diarization for Monaural Dialogue Recordings with Target-speaker Acoustic Models"*, Proc. IEEE ASRU, 2019.
[34] N. Itoh, G. Kurata, R. Tachibana, and M. Nishimura, *"A Metric for Evaluating Speech Recognizer Output Based on Human-perception Model"*, Proc. Interspeech, 2015.
[35] T. Fukuda, O. Ichikawa, and M. Nishimura, *Long-term Spectro-temporal and Static Harmonic Features for Voice Activity Detection"*, IEEE Journal of Selected Topics in Signal Processing, Vol.4, No.5, pp.834–844, 2010.
[36] T. Fukuda, O. Ichikawa, and M. Nishimura, *"Detecting Breathing Sounds in Realistic Japanese Telephone Conversations and Its Application to Automatic Speech Recognition"*, Speech Communication, Vol. 98, pp.95-103, April 2018.
[37] A. Narayanan, R. Prabhavalkar, C. Chiu, D. Rybach, T. Sainath, and T. Strohman, *"Recognizing Long-form Speech Using Streaming End-to-end Models"*, Proc. IEEE ASRU, 2019
[38] C. Chiu, W. Han, Y. Zhang, R. Pang, S. Kishchenko, P. Nguyen, H. Soltau, A. Narayanan, H. Liao, S. Zhang, A. Kannan, R. Prabhavalkar, Z. Chen, T. Sainath, and Y. Wu, *"A comparison of end-to-end models for long-form speech recognition"*, Proc. IEEE ASRU, 2019
[39] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, *"Audio augmentation for speech recognition"*, Proc. Interspeech pp.3586–3589, 2015.
[40] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, *"SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition"*, Proc. Interspeech, 2019.
[41] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, *"mixup: Beyond Empirical Risk Minimization"*, Proc. ICLR, 2018.
[42] I. Medennikov, Y. Khokhlov, A. Romanenko, D. Popov, N. Tomashenko, I. Sorokin, and A. Zatvornitskiy, *"An Investigation of Mixup Training Strategies for Acoustic Models in ASR"*, Proc. Interspeech, 2018.
[43] T. Nagano, T. Fukuda, M. Suzuki, and G. Kurata, *"Data Augmentation Based on Vowel Stretch for Improving Children's Speech Recognition"*, Proc. IEEE ASRU, 2019.
[44] T. Fukuda, R. Fernandez, A. Rosenberg, S. Thomas, B. Ramabhadran, A. Sorin, and G. Kurata, *"Data Augmentation Improves Recognition of Foreign Accented Speech"*, Proc. Interspeech, pp. 2409–2413, 2018.
[45] T. Fukuda, M. Suzuki, and G. Kurata, *"Direct Neuron-wise Fusion of Cognate Neural Networks"*, Proc. Interspeech, 2019.
[46] F. Iwama and T. Fukuda, *"Automated Testing of Basic Recognition Capability for Speech Recognition Systems"*, Proc. IEEE International Conference on Software Testing, Verification and Validatation (ICST), 2019.
[47] M. Heck, M. Suzuki, T. Fukuda, G. Kurata, and S. Nakamura, *"Ensembles of Multi-scale VGG Acoustic Models"*, Proc. Interspeech 2017, pp. 1616–1620, 2017.
[48] T. Fukuda, O. Ichikawa, G. Kurata, R. Tachibana, S. Thomas, and B. Ramabhadran, *"Effective Joint Training of Denoising Feature Space Transforms and Neural Network Based Acoustic Models"*, Proc. IEEE ICASSP, pp. 5190–5194, 2017.