Vol.2015-SLP-105 No.7
2015/2/27

国際会議INTERSPEECH2014, SLT2014参加報告

浅見 太一[1]　岩野 公司[2]　小川 哲司[3]　駒谷 和範[4]　齋藤 大輔[5]　篠田 浩一[6]　太刀岡 勇気[7]
東中 竜一郎[1]　福田 隆[8]　増村 亮[1]　渡部 晋治[9]

概要：2014 年 9 月 14 日から 18 日にかけシンガポールで開催された ISCA 主催の INTERSPEECH2014, 及び，同年 12 月 14 日から 18 日にかけて米国レイク・タホで開催された IEEE 主催の SLT2014 に参加した．ともに，音声言語処理分野で一流の国際会議である．ここでは，海外からの発表を中心に，これらの会議における最新の技術動向，注目すべき発表について報告する．

1. はじめに

2014 年 9 月 14 日から 18 日にかけシンガポールで開催された ISCA 主催の INTERSPEECH2014，及び，同年 12 月 14 日から 18 日にかけて米国レイク・タホで開催された IEEE 主催の SLT2014 に参加した．ともに，音声言語処理分野で一流の国際会議である．INTERSPEECH2014 への投稿数は 1173 で採択数は 614，受理率は 52%であった．SLT2014 への投稿数は 216 で採択数は 104，受理率は 48%であった．以下，海外からの発表を中心に，INTERSPEECH2014 について分野ごとに最新の技術動向および注目すべき発表について紹介した上で，最後に SLT2014 について報告する．

2. 音響モデル

音響モデルの研究動向としては，引き続き Deep neural network(DNN) の研究発表が多く見られた．その中で注目されるのは，リカレントニューラルネットワーク (RNN) 及びその発展系である Long Short-Term Memory network(LSTM) である [1], [2]．[1] では，ネットワークの時間方向への展開・固定フレーム数分の履歴の切捨てにより，RNN を時間遅れの入力を考慮した DNN と同様のトポロジーで表現できるため，DNN と類似の実装で RNN を実現できている．LSTM は RNN よりも時間依存性を柔軟に表現できるため，音声や言語のモデル化に適している．[2] では、LSTM の分散型学習を提案しており，1900 時間に及ぶ大量の音声データを用いた学習を実現している．また特筆すべきはその総パラメータ数であり，LSTM の柔軟な時間依存性により、DNN よりも少ないパラメータで同程度の性能を達成しており、結果として学習時間の削減にも寄与している．Google は LSTM を音声認識を含む多くのアプリケーション (音声区間検出，機械翻訳) に利用しており，今後 DNN が LSTM に置き換わる可能性を秘めている．上記の RNN や LSTM を含む，Deep learning の様々なアイデアを実装するためには, GPU プログラミングを含めた多くの労力を必要とする．Microsoft Research は, 多様な Deep network をシンプルなスクリプトの記述により実現する Computational Network ToolKit (CNTK)[3][*1]を公開した．CNTK を使うことにより，例えば新規に考案した複雑なネットワークトポロジーを容易に実装できるため, 自前の大規模なソフトウェアインフラストラクチャーを持たない大学や小規模な企業研究所でも，最先端の Deep network を実現できるため, この分野での研究がさらに加速すると期待される．原稿執筆現在[*2], CNTK は Windows による開発環境のみを提供しているが，近日中に Linux への porting も実現する予定である．

最後に紹介する文献 [4] は，DNN による特徴表現の学習能力が，従来の発見的な特徴量である MFCC や PLP にどこまで迫っているかを考察した文献である．興味深い結果として、特に大量データで学習した場合は，時間領域の音声信号を入力に用いた場合でも，MFCC との性能差は大きくは変わらず (MFCC:21.1%, time signal: 23.5%)，また最も入力に近い層の重み係数は、バンドパスフィルター

1 NTT
2 東京都市大学
3 早稲田大学
4 大阪大学
5 東京大学
6 東京工業大学
7 三菱電機株式会社
8 日本アイ・ビー・エム株式会社
9 Mitsubishi Electric Research Laboratories

*1 https://cntk.codeplex.com/
*2 2015 年 1 月 28 日

の役割を担っていることがわかった．これらのアプローチは，音声認識における信号処理を用いた特徴抽出部を，Data-driven な手法に置き換える可能性を示したものであると言える．(渡部)

3. 耐雑音音声認識

音声認識の耐雑音性に関する研究は古くから盛んに行われており，信号処理・特徴抽出，音響モデル，適応処理など様々な要素技術に対して幅広い検討がなされてきた．また，近年のディープラーニング分野の目覚しい進展は耐雑音性の研究にも波及し始め，DNN における耐雑音性の効果が確認されると共に，耐雑音性に焦点を当てた DNN の関連研究も見受けられるようになってきた．一般に，DNN に代表される学習ベースの方法では，学習と評価の不適合に起因する性能低下が問題になっている．本章では，これを改善する研究を 2 つ紹介した後，耐雑音性検証実験を中心とした文献を 2 つ取り上げる．

近年，画像認識発祥の畳み込みネットワーク (CNN) を，音声認識に適用した研究が多くなっている．CNN は画像のずれに対応するため，局所的な特徴量にフィルタを畳み込むことで性能を向上させている．フィルタは学習データのみから何らの事前知識なく学習されるので，学習と評価で不適合がある場合には，学習がうまくいかない可能性がある．[5] は，聴覚モデルに基づく 2D-Gabor フィルタをフィルターの初期値 (事前知識) として使うことで，フィルタ推定の頑健性を向上させている．2 つの音声認識タスクで手法の有効性を示している．

[6] では DNN の入力に，観測特徴量に加えて，いくつかの騒音推定の手法で予測した騒音のみの特徴量を合わせたものを使い，騒音の推定と DNN による音声強調処理を分けることで，頑健性を向上させている．騒音推定にも DNN を使い，バイナリマスクを推定する方法が，最良であったことを，PESQ の評価で示している．

一方，文献 [7] では雑音下およびチャネルミスマッチな環境における CNN と DNN の挙動について，これまでに提案されてきた特徴量と数種類のネットワーク構造の組み合わせを用いて検証実験を行っている．Aurora4 タスクにより CNN と DNN の性能を比較した場合，CNN には DNN よりも一貫した頑健性が確認されたことを述べている．他方 [8] では，クリッピング，低ビットレート信号圧縮，および過渡な雑音除去に伴う非線形歪みに対するハイブリッド DNN/HMM の頑健性を調査し，チャネル補正を施した従来の GMM/HMM システムとの比較を行っている．ハイブリッド DNN/HMM は特にスペクトルをマスクしてしまうような非線形歪みに対して大きな改善があり，HEQ などのスペクトル補正手法の併用は効果が限定的であったことを述べている．

以上で取り上げたように，DNN の耐雑音性に関する研究は，現在ではネットワークの構造に深く入り込んだ提案は少なく，DNN で用いられる特徴量についての検討がほとんどである．(福田・太刀岡)

4. 言語モデル・発音辞書

言語モデルの全体的な傾向としては，音響モデルと同様に Neural Network に基づく手法の報告が多く，特に RNN 言語モデルに関しては，その拡張や実装に関わる報告が 6 件あった．また発音辞書についての報告も多く，音声認識における確率モデルの 1 つとしての再考が伺えた．本節では，特に興味深い文献をピックアップして概説する．

まず言語モデルを発展させる話題として，2 つの文献を紹介する．[9] では，LSTM に基づく言語モデルをサポートするツールキット `rwthlm` の紹介とその評価を報告している．Treebank コーパスを用いたパープレキシティによる評価では，N-gram 言語モデルが 141，RNN 言語モデルが 122 である一方で，LSTM 言語モデルが 108 という高い性能を示しており，長距離文脈の考慮が言語モデルにおいて重要であることを再確認できる報告と言える．一方 [10] では，Sum-Product Network(SPN) に基づく言語モデルを提案している．SPN は加算ノードと乗算ノードを持つ DNN の 1 種であり，[9] と同一の実験条件でパープレキティによる評価を実施したところ，SPN 言語モデルは 100 という高い性能 (N-gram 言語モデルと組み合わせると 81) を実現している．SPN 言語モデルは，N-gram 言語モデルと同様に直前の数単語のみを文脈として考慮する言語モデルである．長距離文脈を考慮することなく高い性能を実現できている点は，乗算ノードを持つモデル構造に起因したものと考えられ，注目に値する．

従来の N-gram 言語モデルの範囲内においても，[11] が興味深い結果を示している．コンテキストに合わせて単語やフレーズをトークン化して扱ってモデリングすることで，通常の単語ベースの N-gram モデルと比較して単語誤り率で 2 ポイント程度の改善効果を得ている．モデル化方法以外で高い改善効果が得られる点は，重要な知見である．

また言語モデルにおいては，技術とともに大量の学習データを扱うことも重要である．その点において，[12] では 10 億単語の学習データを用い，代表的な技術についてベンチマークの結果を報告している．このデータはパブリックに公開されており，技術比較を実施する際の評価対象として，今後は有用な存在となるであろう．

発音辞書については，書記素音素変換や複数の音素表記付与の検討が近年の傾向である．その中で [13] では，音素の概念を用いずに，発音辞書を構築する方法を提案している．具体的には，書記素による発音辞書を初期モデルとし，各単語の発音を最適化していくことで，専門家による音素ベースの発音辞書に近い性能を実現している．日本語等の書記素数の多い言語への適用は難しいが，少資源言語の音

声認識において有用な報告と言える．(増村)

5. 応用システム

近年の音声処理技術の進歩により様々なアプリケーションが現実的となり，INTERSPEECH でも音声アプリケーションを構築するための数々の技術が発表された．

放送コンテンツや講義などを音声認識でテキスト化し，話題分類やインデキシングにより内容把握を容易にする音声ドキュメント処理は有用な音声認識アプリケーションの一つであり，多くの報告が行われた．[14] はコールセンタの電話応対の話題分類に取り組んでいる．音声認識誤りによるテキストの変動に対して頑健に話題を捉えるために，まず Latent Dirichlet Allocation (LDA) のハイパーパラメタを変えて観点の異なる複数のトピック空間でドキュメントを表現する．それらを連結したスーパーベクトルを因子分析で次元削減することにより主要な変動要因である話題の違いを取り出す手法を提案し，単一トピック空間でドキュメントを表現するよりも高い話題分類精度を達成している．

[15] は多人数会話の話題境界検出に取り組んでいる．使われている単語の分布の変化点を話題境界とする考え方が一般的だが，著者らはそれに加えて，会話に参加している（発言している）話者の分布の変化点も話題境界として尤もらしいという考えを導入する．この考えに基づく，単語分布と話者分布をともに考慮したスコアにより，単語のみを考慮した場合よりも高い話題境界検出精度が得られている．

[16] は，チェコ／チェコスロバキアの 90 年分のラジオ放送を検索可能なアーカイブとして公開したプロジェクトの成果報告である．チェコ語とスロバキア語が混在し，BGM 等が重畳し，狭帯域と広帯域が混在する 10 万時間超のラジオ放送に，全文検索インデックスのほか言語 ID，話者 ID，ジングル等の音響イベントラベルを付与するタスクに取り組んでいる．システムは音声認識に加えて音楽検出や話者ダイアライゼーション，帯域識別，言語識別など多数のモジュールから為る．各モジュールは基本的な技術で構成されているが，実際に音声認識システムを構築する際に直面する課題と対処法を広く含む内容となっている．

スマートフォンやウェアラブル端末を対象にしたユニークなアプリケーションも見られた．[17] は，個人のスマートフォンに撮り貯められた大量の写真に対して，「ビルおじさんの家」のような極めて個人的なクエリで検索を行う実現難度の高いアプリケーションに挑戦している．撮影場所や撮影対象が何であるかを発声した音声が一部の写真に付随しているという前提のもと，画像特徴量と音声認識結果をトピック空間で対応付けることで，音声が付随していない写真の検索も可能にする手法が検討されている．十分な検索精度は得られていないもののアプリケーションは魅力的であり，採用されているアプローチも興味深い．

[18] は，ウェアラブル端末で収録した個人の一日分の音声ライフログから，その人が一日に発声した単語の個数をカウントするシステムを提案している．人混みや自動車内，オフィスなど様々な雑音環境で収録される音声ライフログからの単語カウントは容易ではなく，一般的な音声認識のようなモデルベースの手法は学習データと収録音声のミスマッチが避け難い．著者らは学習が不要な信号処理ベースの音節検出を使い，得られた音節カウントから線形回帰で単語カウントを推定する手法を提案し，実際のライフログデータから誤差 20%程度で単語数を推定するシステムを実現している．(浅見)

6. 話者認識

話者認識の分野でもニューラルネットの利用の試みが進み，INTERSPEECH2014 においても関連する報告が目立つようになった．特に，DNN を話者照合のスコアリングに用いるのではなく，i-vector の抽出過程における統計量の計算に用いるアプローチは，UBM/i-vector に代わる ASR/i-vector アプローチと呼ばれ成功を収めており，複数件の報告があった．従来の UBM/i-vector アプローチでは，UBM により入力空間を分割し，混合要素の事後確率を計算する．一方，ASR/i-vector アプローチでは，HMM の状態によって入力空間を分割し，ASR の結果得られる状態の事後確率を DNN によって計算している [19]．しかし，DNN は高精度であるがゆえに，雑音等音環境のミスマッチに頑健ではないため，DNN の代わりに CNN を適用することで，雑音に対する頑健性を向上させる試みがなされている [20]．同様に，DBN の隠れ層の要素に対する事後確率を計算し，i-vector や Gaussian supervector を抽出する試みもなされ，従来の枠組みと同等の性能を得ている [21]．また，現在の話者認識フロントエンドの state-of-the-art である i-vector に基づく話者照合の性能向上と，機械学習研究者を話者認識研究に呼び込むことを目的に昨年開催された “NIST Speaker Recognition i-vector Machine Learning Challenge” の結果が報告されたが [22]，ここでも，DBN に基づくフロントエンドを利用したシステムが性能改善に寄与している．

前年の INTERSPEECH2013 でスペシャルセッションが組まれた，「各種の攻撃に対する話者照合の頑健性の向上に関する研究」については，本年は 1 件の研究報告に留まった [23]．この報告では GMM を利用した声質変換攻撃を想定した対策手法を提案しており，「登録話者を目標とした声質変換音声を作成し，それらを学習データに含めた PLDA の枠組みで（変換を行っていない）本人の声かどうかの判定を行う」方式の効果が高く，成りすまし音声に対する FAR が 5〜6 割ほど削減されることを示している．

また，音響的な話者認識とは異なるが，多人数会話の

音声認識結果を利用した「話者の役割認識（speaker role recognition）」についての報告もあった [24]．AMI コーパスの会議音声を対象に，各話者が「プロジェクトマネージャー」「インターフェースデザイナー」といった役割のどれであるかを推定するタスクで，特徴量には各話者の発言の構成単語から得られる LDA に基づくトピック混合比を，識別器には SVM を利用している．正解精度は約 7 割となり，話者決定後のメタ情報付与の一手法として期待される．(岩野・小川)

7. 対話システム

今回の INTERSPEECH では, 主に Spoken Dialogue System と 特別セッション Open Domain Situated Conversational Interaction の 2 つで，対話システム関連の研究が発表された．本章では，応用と基礎技術の両方の視点から，これらの発表の一部を紹介する．

特定のタスクを遂行するタスク指向型音声対話システムの研究は，Apple の Siri や NTT ドコモのしゃべってコンシェルの商用化が行われるなど一定の成果を上げたと言える．フォームを埋めるような単純なタスクは Web ベースのアプリケーションで実現でき，音声対話を行う必要性は薄いため，状況依存対話とオープンドメイン対話が現在注目されている．状況依存対話とは，自動車の運転中やロボットとの対話中といった，特定の物理的状況における対話を指す．このような対話はハンズフリー・アイズフリーが望まれ，Web ベースのアプリケーションとは異なる．また特定の状況を仮定できることから，発話理解なども行いやすい．オープンドメイン対話とは，制約を設けずに対話を行う技術の総称である．オープンドメイン対話には，タスク指向型対話において自由なユーザ発話を許容するオープンドメインの発話理解と，非タスク指向型対話（いわゆる雑談対話）においてオープンな話題を扱うものに大別され，どちらの研究も盛んになりつつある．

[25] は物理的状況における音声認識や発話理解の改善に関するものである．具体的には，Townsurfer と呼ばれる車載対話システム（周りの建物などについてドライバーの質問に答えることができる）において，言語モデル・発話理解モデルの改善のためのデータをクラウドソーシングを用いて収集する手法を提案している．物理的状況を伴う状況依存対話ではデータ収集が機材の関係もあり難しい．そこで，本研究では，写真と動画をワーカーに提示して発話データを収集し，その有効性を検証している．結果として，写真を用いた場合の方が，動画よりも発話が集めやすく，予想される改善も大きかった．ただし，動画でないと取得できないタイプの発話もいくつか見られた．

[26] はコールセンターにおける発話の理解に関するものである．コールセンターには一般に様々な問い合わせがあり，オープンドメインの発話理解系が望まれる．そこで，著者らはフレーム意味論におけるフレーム（出来事を表す汎用的な意味表現）をユーザ発話から抽出することに取り組んでいる．フレームの抽出には発話内の単語の関係性を把握することが重要であるため，本研究では，話し言葉用の依存構造解析器を構築し，コールセンタの実データの書き起こしについて，一定の精度でフレームが抽出できることを報告している．

一方, 音声対話システムの基礎技術としては，その知識を扱う研究と，ターンテイキングを扱う研究が見られた．これまで対話システムの知識は，人手でドメインごとに定義したオントロジ（スロット構造）を用いる「狭く深い」場合と，構造を仮定せず大量のテキストを保持しそれに対する検索を行う「広く浅い」場合のいずれかが多い．前者は，近年の POMDP に基づく統計的対話管理を含む多くのタスク指向型対話システムの知識構造であり，後者は call routing から始まる，情報検索に基づく応答選択手法における知識構造である．これらに対して「広く深い」知識を扱う取り組みとして，知識グラフ (Knowledge Graph) を対話システムの知識として用いる試みが行われており，チュートリアル（T8）でも取り上げられた．これは，Freebase など，知識を大量に記述し共有する LOD (Linked Open Data) の流れとも呼応している．

このような流れの中で，言語理解における曖昧性の解消を目的として，知識グラフ上のエンティティに対して重みを与える研究が行われた [27]．ここでは，公開されている知識グラフや Wikipedia 内のテキストを用いた重みづけにより，入力発話のタイプ推定の性能向上を示している．さらに，このように共有される知識は一般的なものであるため，ドメイン固有の知識は，対話を通じて獲得できることが望ましい．[28] では，大学における講演の検索をタスクとした対話システムにおいて，被験者の発話から知識グラフを獲得する試みについて報告している．ここでは，発話中に現れる研究トピックと研究者名の共起から知識グラフを作成し，得た知識グラフの分析や，それを用いた質問拡張 (query expansion) による検索性能の向上を示している．

音声対話システム (Spoken Dialogue System) セッションでは，応答内容を扱う研究というより，ターンテイキングや韻律などパラ言語情報を扱う発表が大半を占めた．そのひとつとして [29] では，Let's Go!バスシステムにおいて，ユーザの長い発話はシステムに理解されにくいことから，長い発話の後半はタスク成功に貢献しない傾向をまず示した．この傾向をもとに，漸次的対話処理 (incremental dialogue processing) の枠組みに基づき，ユーザの発話中の 175 ミリ秒ごとに，長い発話に対してシステム側から割り込み（バージイン）を行うかどうかを決定する手法を提案した．このシステムを実ユーザに対して公開し，ユーザ発話が短くなる傾向や，タスク成功率などのシステム性能の向上を示した．

(東中・駒谷)

8. 音声合成・声質変換

本年の INTERSPEECH では，音声合成および声質変換に関連するセッションとして 7 つのセッションがあった．特筆すべき内容として，音声合成および生成における DNN に関するスペシャルセッションが組まれた事が挙げられる．音声合成においては最終的に音声を生成する観点から，ディープラーニング関連技術の使用に関して，音響モデリング以外にも様々な段階での応用が考えられ，今回のスペシャルセッションにおいても，多岐に渡る利用が紹介されていた．一方，音声合成分野における個々の研究ターゲットとしては，Grapheme-to-Phoneme (G2P) や焦点推定などのテキスト処理，クロスリンガル，アクセント（訛り）や発音変動，Expressive な音声合成等，従来からの課題となっているトピックがまんべんなく取り組まれている印象があった．声質変換およびその関連タスクにおいては，DNN の他，GMM に基づく回帰・変換も引き続き多く見られた．以下本章では DNN 関連のものを中心にいくつか興味深かった文献について個別にとりあげる．

[30] では，近年 ASR においてデファクトスタンダードになりつつある Kaldi[*3] をベースとしてパラメトリック音声合成システム全体を構築しようというプロジェクトである Idlak について紹介され，そのうちのテキスト処理に関するフロントエンドについて発表されていた．Kaldi のようなパイプライン式の処理を採用し，テキスト処理に関しては XML フォーマットを介してやり取りする．フルコンテキストラベルについても，HTK/HTS が採用しているようなモデル名をベースにしたものではなく，XML による構造化が図られており，可読性や柔軟性の向上が見られる．Idlak は Kaldi のブランチとして公開されている[*4]．

[31] では，HMM 音声合成における過剰平滑化の問題に対する，DNN を用いたアプローチを紹介している．HMM 音声合成では，統計的アプローチを用いたモデリングの結果としてスペクトルが過剰に平滑化する問題がある．これまで，系列内変動（Global Variance; GV）やパラメータの変調スペクトル（Modulation Spectrum; MS）を補償・考慮するアプローチが試みられてきたが，この文献では合成音声と自然音声のパラレルデータから，この過剰平滑化を補償するフィルタを直接学習しようというアプローチになっている．実験の結果，従来法よりも自然性の改善が見られたと報告されている．

[32] では，LSTM を Bidirectional RNN に用いた BLSTM-RNN を，DNN-based の音声合成のモデルとして使うことを検討している．DNN-based 音声合成ではコンテキストラベルから音響パラメータへのマッピングを直接 DNN で学習するフレームワークによって，従来の HMM 音声合成のガウス分布+決定木のフレームワークを置き換えている．しかしパラメータの時間依存性は未だにデルタパラメータを考慮したトラジェクトリ生成が中心であった．この文献ではパラメータ系列の時間依存性を学習できる BLSTM-RNN を，従来の DNN-based の音声合成のうち，出力層部分に用いる事で，デルタパラメータおよびラベルのコンテキストを明示的に用いる事なく，滑らかなパラメータ生成を実現している．

[33] では，RBM および，離散変数間の Bernoulli Bidirectional associative memory (BBAM) によって，生成モデル的に学習された DNN を用いた声質変換について述べている．この文献の主張として，声質変換のようなタスクにおいては，スペクトルのモデリングエラーを最小化する識別的な学習では，必ずしも聴感上の音質の向上が見られないとしている．本文献では，デルタ特徴量の代わりにマルチフレームを連結した特徴量を提案モデルによって直接変換することを提案し，最終的にファインチューニングをする場合よりも主観評価がよい傾向が見られた．(齋藤)

9. SLT2014

SLT(Spoken Language Technology)2014 は隔年に開催される，IEEE 主催の音声言語処理分野に重点を置いた国際ワークショップで，同じく IEEE 主催の音声認識・理解の国際ワークショップ IEEE ASRU(Automatic Speech Recognition and Understanding) と交互に開催されている．今年は，参加者が 240 名ほどで例年に比べ多く，また，2/3 が SLT に初参加，1/4 以上が学生であった．全体的に若く活気がある．旧知の研究者が少なく，代替わりをした印象をもった．また，他の国際会議に比べ，米国在住の研究者の占める割合が多く，APSIPA2015 と時期的に重なったこともあり，アジア，特に，日本からの参加者は少なかった．

発表は，従来の音声言語理解とその応用に関する発表が半分程度を占めたが，それ以外に，他の近年の国際会議の例に漏れず，音響モデルの Deep Learning に関する発表が目立った．基調講演，招待講演，Tutorial では，活躍している若手研究者がビギナー向けに各分野の動向について丁寧に説明していた．また，Big Data と研究者キャリアに関するパネルディスカッションでも活発に議論がされていた．これらの企画は若手のために若手自身で立てられており，好感が持てた．日本の国内会議でも見習う点が多いように思う．企画の中では，Tara Sainath の Deep Learning に関する基調講演が印象的であった．同氏は IBM から Google に移籍したばかりである．CNN, RNN を音声認識に用いた結果を報告している．DNN からのエラー削減率は各々は 4-5%程度ではあったが組み合わせにより一割程度している．また，並列計算などを用いた高速化についても，これ

[*3] http://kaldi.sf.net/
[*4] https://svn.code.sf.net/p/kaldi/code/sandbox/idlak

らの研究機関のリッチな計算資源を活用して，例えばIBMのBlue Gene/Qを使うと，Hessian Free学習でも，GPUによるSGDよりも10倍早く計算できる，などの結果を見せてくれた．曰く「Googleは毎日10年分の音声を処理している」とのことである．

一般発表の中では特にBest Paperに選ばれたエジンバラ大からの話者適応の論文 [34] を紹介する．この手法では，DNNの各ノードの出力を乗算するパラメータ (スカラ) を新たに加え，話者適応ではその乗算する値を学習する．ノード数を N としたとき，通常のDNNの学習では，N^2 のオーダーの重み係数を推定する必要があるが，この手法では N 個の乗算係数を推定するのみであり，少量のデータでも頑健な学習ができる．従来のfMLLRを用いた手法と同程度の10%程度の誤り削減率を達成し，また，それと組み合わせることにより更に5%誤りを削減している．(篠田)

参考文献

[1] H. Sak *et al.*, "Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling," *Proc. Interspeech*, pp. 338-342, 2014.
[2] G. Saon *et al.*, "Unfolded Recurrent Neural Networks for Speech Recognition," *Proc. Interspeech*, pp. 343-347, 2014.
[3] D. Yu *et al.*, "An Introduction to Computational Networks and the Computational Network Toolkit," *Proc. Interspeech*, pp. 895-899, 2014.
[4] Z. Tüske *et al.*, "Acoustic Modeling with Deep Neural Networks Using Raw Time Signal for LVCSR," *Proc. Interspeech*, pp. 890-894, 2014.
[5] S. Chang *et al.*, "Robust CNN-based Speech Recognition with Gabor Filter Kernels," *Proc. Interspeech*, pp. 905-909, 2014.
[6] Y. Xu *et al.*, "Dynamic Noise Aware Training for Speech Enhancement Based on Deep Neural Networks," *Proc. Interspeech*, pp. 2670-2674, 2014.
[7] V. Mitra *et al.*, "Evaluating robust features on Deep Neural Networks for speech recognition in noisy and channel mismatched conditions," *Proc. Interspeech*, pp. 895-899, 2014.
[8] L. Seps *et al.*, "Investigation of Deep Neural Networks for Robust Recognition of Nonlinearly Distorted Speech," *Proc. Interspeech*, pp. 363-367, 2014.
[9] M. Sundermeyer *et al.*, "rwthlm–The RWTH Aachen University Neural Network Language Modeling Toolkit," *Proc. Interspeech*, pp. 2093-2097, 2014.
[10] W. Cheng *et al.*, "Language Modeling with Sum-Product Networks," Interspeech, *Proc. Interspeech*, pp. 2098-2102, 2014.
[11] M. Levit *et al.*, "Word-Phrase-Entity Language Models: Getting More Mileage out of N-grams," *Proc. Interspeech*, pp. 666-670, 2014.
[12] C. Chelba *et al.*, "One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling," *Proc. Interspeech*, pp. 2635-2639, 2014.
[13] D. Harwath *et al.*, "Speech Recognition without a Lexicon - Bridging the Gap between Graphemic and Phonetic Systems," *Proc. Interspeech*, pp. 2655-2659, 2014.
[14] M. Morchid *et al.*, "I-vector based representation of highly imperfect automatic transcriptions," *Proc. Interspeech*, pp. 1870-1874, 2014.
[15] A. Bouchekif *et al.*, "Speech cohesion for topic segmentation of spoken contents," *Proc. Interspeech*, pp. 1890-1894, 2014.
[16] J. Nouza *et al.*, Speech-to-text technology to transcribe and disclose 100,000+ hours of bilingual documents from historical Czech and Czechoslovak radio archive," *Proc. Interspeech*, pp. 964-968, 2014.
[17] Y. Liou *et al.*, "Semantic retrieval of personal photos using matrix factorization and two-layer random walk fusing sparse speech annotations with visual features," *Proc. Interspeech*, pp. 1762-1766, 2014.
[18] A. Ziaei *et al*, "A speech system for estimating daily word counts," *Proc. Interspeech*, pp. 880-884, 2014.
[19] Y. Lei *et al.*, "A deep neural network speaker verification system targeting microphone speech," *Proc. Interspeech*, pp. 681-685, 2014.
[20] M. McLaren *et al.*, "Application of convolutional neural networks to speaker recognition in noisy conditions," *Proc. Interspeech*, pp. 686-690, 2014.
[21] W. M. Campbell, "Using deep belief networks for vector-based speaker recognition," *Proc. Interspeech*, pp. 676-680, 2014.
[22] D. Bansé *et al.*, "Summary and initial results of the 2013-2014 speaker recognition i-vector machine learning challenge," *Proc. Interspeech*, pp. 368-372, 2014.
[23] E. Khoury *et al.*, "Introducing i-vector for joint anti-spoofing and speaker verification," *Proc. Interspeech*, pp. 61-65, 2014.
[24] A. Sapru and H. Bourlard, "Detecting speaker roles and topic changes in multiparty conversations using latent topic models," *Proc. Interspeech*, pp. 2882-2886, 2014.
[25] T. Misu, "Crowdsourcing for Situated Dialog Systems in a Moving Car," *Proc. Interspeech*, pp. 125-129, 2014.
[26] F. Bechet *et al.*, "Adapting Dependency Parsing to Spontaneous Speech for Open Domain Spoken Language Understanding," *Proc. Interspeech*, pp. 135-139, 2014.
[27] D. Hakkani-Tür *et al.*, "Probabilistic Enrichment of Knowledge Graph Entities for Relation Detection in Conversational Understanding," *Proc. Interspeech*, pp. 2113-2117, 2014.
[28] A. Pappu and A. I. Rudnicky, "Learning Situated Knowledge Bases through Dialog," *Proc. Interspeech*, pp. 120-124, 2014.
[29] F. Ghigi *et al.*, "Incremental Dialog Processing in a Task-Oriented Dialog," *Proc. Interspeech*, pp. 308-312, 2014.
[30] M. P. Aylett *et al.*, "A Flexible Front-End for HTS," *Proc. Interspeech*, pp. 1283-1287, 2014.
[31] L. Chen *et al.*, "DNN-Based Stochastic Postfilter for HMM-Based Speech Synthesis," *Proc. Interspeech*, pp. 1954-1958, 2014.
[32] Y. Fan *et al.*, "TTS Synthesis with Bidirectional LSTM Based Recurrent Neural Networks," *Proc. Interspeech*, pp. 1964-1968, 2014.
[33] L. Chen *et al.*, "Voice Conversion Using Generative Trained Deep Neural Networks with Multiple Frame Spectral Envelopes," *Proc. Interspeech*, pp. 2313-2317, 2014.
[34] P. Swietojanski *et al.*, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," *Proc. IEEE SLT*, pp. 171-176, 2014.