Vol.2014-SLP-100 No.2
2014/1/31

「音声認識」は今後こうなる！

篠田 浩一[1,a)]　堀 貴明[2,b)]　堀 智織[3,c)]　篠崎 隆宏[1,d)]

概要：情報処理学会音声言語情報処理 (SLP) 研究会が 100 回を迎えた．音声認識・理解はこの 20 余年の間に当初は予想もできないほど飛躍的な進歩を遂げた．本研究会は日本における音声認識・理解研究の議論・発表の場としてその進歩に大きく貢献してきた．本稿では，この記念すべき 100 回目の研究会における一連の企画の 1 つとして，この 100 回の歩みを踏まえた上で，今後音声認識・理解研究が進むべき方向性について，4 人の研究者が提言を行う．

キーワード：音声認識，音声理解

Where is Speech Recognition Going?

Koichi Shinoda[1,a)]　Takaaki Hori[2,b)]　Chiori Hori[3,c)]　Takahiro Shinozaki[1,d)]

Abstract: We celebrate the 100th meeting of IPSJ Spoken Language Processing (SLP) SIG. The technology of speech recognition and spoken language understanding has advanced beyond our expectation in these 20 years. SLP SIG has greatly contributed to this advancement as one of the communities to discuss research topics in this field. In this article, four researchers discuss the directions to which speech recognition and understanding researches should go in the future.

Keywords: speech recognition, spoken language understanding

1. 20 年前と 20 年後 (篠田浩一)

1.1 今後の方向性

過去の 20 年余りの音声認識の進歩を今振り返ってみると「予定調和」，つまり，進むべき方向に進んだだけのように思える．当初から統計モデルとその学習アルゴリズムは提案されており，計算機技術の進歩にしたがって，より自由で強力な実装が許されるようになり，より汎用な（したがって分かりやすく教科書的な）方法がスタンダードとなった．ここから学べることはパターン認識技術の発展はハードウェアの制約を強く受けている，ということである．つまり今後の方向性についても，我々の選択の余地は実はあまりなく，むしろ将来のハードウェアの進歩を予想し，それに沿った方向を目指すことが大事である．端的な例が携帯電話である．私が研究を始めたころは，現在の携帯電話のあり方を誰も予想していなかった．予想もしない進歩が予想もしなかったサービスを産み，新しい需要が生まれ，新しい技術が開発される．

また，我々はこの 20 年で，「量は質を凌駕する」ことをまざまざと見せつけられてきた．精緻なモデリングよりもデータをたくさん集めることのほうが往々にして認識性能の向上に寄与した．我々はデータ量に対してスケールする方法を目指してきたが，それでも「桁が違う」話になると方法論自体が変わってくる．今後も爆発的な勢いでデータが増加する．新たな方法論が求められている．

[1] 東京工業大学
Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo
[2] 日本電信電話 (株)
NTT, 2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto
[3] (独) 情報通信研究機構
NICT, 3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto
[a)] shinoda@cs.titech.ac.jp
[b)] hori.t@lab.ntt.co.jp
[c)] chiori.hori@nict.go.jp
[d)] http://www.ts.ip.titech.ac.jp

1.2 個別の予想

まず，センサ技術は今後も進歩していくであろう．小型化が進み，無線通信，自動給電の機能を備えた，現在の RFID のサイズ程度のセンサが世の中にばらまかれる．我々も様々なセンサを身に着ける．それらの信号を蓄積するクラウドの技術も進歩する．プライバシの問題が大きな障害になるが，我々はプライバシについて 20 年前には想像できないほど鈍感になった．今後も，プライバシと利便性をはかりにかけ，益々鈍感になっていくと予想する．もちろんセキュリティ技術の進歩が伴うことは言うまでもない．音声について言えば，人間の数をはるかに凌駕するマイクロホンから常時受音する．マイクロホンの品質は一般的に低い．またその位置も厳密には不明で，さらに時々移動する．反面，極めて長時間の録音が入手できる．音源位置推定，音検出，話者認識，音声認識，などの要素技術に新しい課題を提供する [1].

また，Deep Learning 技術にも期待がかかる．計算量が大きいが, ハードウェアの進歩により問題ではなくなっていく．データをどのように手に入れるか，ラベルをどのように付けるかが重要な課題となる．また，以前から Mixture of Experts (MoE) よりも Product of Experts (PoE) のほうが効率が良いことは知られていた．Deep Learning は PoE の研究の端緒として位置付けられる．今後は，多数の Expert からなる PoE の最適な構造をどのように求めるか，について研究が進むと予想する．

さらに，音声意味理解の分野は応用は進んだものの，方法論ではここ 20 年で目立ったブレークスルーはなかったように思える．一重にデータ量が少なすぎたことが問題である．今後，データが爆発的に増えるに従い，確率・統計的なアプローチが遅ればせながら威力を発揮してくると予想している．

これら，センサ，PoE, 意味理解は，特にメディアを限らない．すなわち，今後，音声と他のメディアとの敷居は低くなる．音声だけセンサがあるという状況はもはやありえない．Deep Learning の進展に伴い，特徴量を求める問題は構造を決める問題へと置き換わる．パターン認識一般における音声認識のみの課題はほとんどなくなるかもしれない．

1.3 研究コミュニティのあり方について

筆者が音声認識の研究を始めた頃，HMM を用いた統計的アプローチが世の中に出てきた．HMM を用いた研究を発表したら，周りから「なんで HMM なんて使うんだ．計算資源を食い尽くす数学のオモチャではないか」，と言われた．それから時が過ぎ，米国で n-gram を用いた特定話者大語彙音声認識がディクテーションソフトとして実用化された．我々もやるべきだと提案したが，「日本語は英語とは違う．n-gram は役に立たないよ」と言われた．念のため断っておくが，これらの発言をした研究者は多くの業績をもつ高名な方々であり，今でも尊敬している．このように誰もが「バカの壁」を築きがちで，それが進歩の障害となる．我々が気づいている将来の芽のうち，歓迎しないもののほうが進歩するというのが，これらから得られる教訓である．

最近，高名な研究者から，「最近の大学の研究の多くは，修士号や博士号をとるためであって，本当に世の中に役に立つものは少ない.」と言われた．痛い指摘だが我が身を振り返ってもその通りと感じた．忸怩たる思いである．未来を見据えた骨太の研究を進めていきたい．

2. 一を聞いて十を知る音声認識 (堀貴明)

音声認識は今後どうなるであろうか．昨今のスマートフォンを中心とする実用化の流れの中で自然に思い出されるキーワードは，ユビキタス (ubiquitous)，ウェアラブル (wearable)，パーソナライズ (personalize) の３つであろうか．どれも聞き飽きた言葉かもしれないが，これらの概念は音声インタフェースとの関連が深い [2][3].

ユビキタスは，「いつでも，どこでも，だれでも」が利用できるインタフェースを意味する．人にとって音声は正にユビキタスな情報伝達の手段であり，音声インタフェースがユビキタスを目指すのはむしろ自然な流れである．そして，ユビキタスな音声インタフェースにはメガネや腕時計に内蔵されたウェアラブルなマイクが良いであろうし，ウェアラブルであればシステムを個々のユーザーにパーソナライズして認識性能を大きく改善できる．

では，今後ユビキタスな環境が整って行くと仮定した場合，どのような音声認識タスク，どのような方法論があり得るかを雑駁に考えてみたい．

2.1 日々の会話を記録する

音声による機器の操作は，未だ十分ではないものの，おおよそ可能になってきたと言ってよい．今後，ユーザーが常にマイクを装着している状況を考慮すると，音声認識の次なるターゲットは，聞こえる音をすべて認識することになるだろう．つまり，ユーザーが話したこと，聞いたことの全てを書き起こし，蓄積することを目指す．そして，個人の生活における仕事およびプライベートな活動において「いつ，どこで，誰と，何を話したか」を記録し，外部記憶のようにいつでも思い出せる仕組みが構築されるであろう．これはライフログの一種と考えても良い．更に，会話に参加した人がログを共有して議事録にしても良いし，会話中に関連する情報や問題解決のヒントをタイムリーに検索するなど，より高度な応用につながる可能性もある．

しかし，このような音声認識は簡単ではない．ウェアラブルマイクへの遠隔発話や，会話における同時発話，背景雑音・残響など，音声認識の精度を劣化させる様々な要因

が含まれている [4]．従って，ウェアラブルなマイクロフォンアレイやデバイス間通信による雑音/残響除去，音源分離などは要素技術として非常に重要である．また，多人数の自由会話音声認識も高い精度を得るのは難しい．少なくとも，パーソナライズによる音響モデルや言語モデルの長期に渡る話者適応により，認識精度を改善していくことが必要であろう．

2.2 多入力・多出力系

遠隔マイクで日常的な会話を収録した場合，話者や雑音などの音源から発せられる信号は重なり合っている．音源同士は独立であってもマイクロフォンアレイのチャネル間には収録環境に依存した未知の相互作用がある．会話であれば，共通の話題，質問とその回答，もしくは意見の同意や衝突といった発話間の相互作用がある．更には，作業しながら発話する場合に，種々の音イベントと音声が連動することもあり得る．

このように複雑に絡み合った混合音を分離して，それぞれの認識結果やその関係を出力させるのは容易ではない．恐らく，複数のチャネルや音源の相互作用を捉えた多入力・多出力系を考える必要があるだろう．しかし，実際に個々の音源とその相互作用をモデル化するのは非常に難しく，もしできたとしても考慮する組合せの多さから，学習や認識が困難になる場合が多い．

但し，もう少し単純に考えれば，このような多入力・多出力系は身近なところにも存在する．ニューラルネットワークである．つまり，複数の信号を一度に入力し，複数の出力が得られるように学習すれば，信号間の相互作用が何らかの形で捉えられ，より良い出力が得られるものと期待できる．つまり，ディープニューラルネットワーク音響モデル [5] やリカレントニューラルネットワーク言語モデル [6] などを多入力・多出力系に拡張していくことも一つの可能性として考えられる．

2.3 一を聞いて十を知る

ユビキタス環境において音声認識システムがパーソナライズされ，ユーザーの発話履歴が蓄積されていく状況では，個々人の履歴をどのように活用するかが重要であろう．そして，この履歴情報も音声認識の精度向上に役立つものと考えられる．単純にユーザーの言語モデルを適応しても良いし，言語モデルを会話の状況や相手によって使い分けても良い．

しかし，最も有効と考えられるのは，ユーザーがコンピュータに話すときに必要以上のことを話さなくて済むことが挙げられる．つまり，システムはユーザーのすべての履歴（移動経路や購買記録なども含む）を知っているので，何か話せば，足りない情報は履歴から補完して解釈することが可能になる．これは，「一を聞いて十を知る」音声認識と言えるかも知れない．

人同士のコミュニケーションの中で生まれる自然さ，手軽さはどこから来るのかを考えたとき，それは恐らく，適当に話しても分かってくれる，という安心感と思われる．そして，人同士の会話ではお互いの知識やコンテキスト (履歴) を共有しているとき，曖昧な表現や省略が許される場合が多い．つまり，聞き手は，社会通念や話し手と共有したコンテキストに基づいて，話し手の伝えたいことを推測・補完して理解している．この観点から，今後の音声認識は，ユーザーの伝えたいことを履歴情報から推測，補完，(場合によっては修正まで) できる能力を持つことが期待される．

しかし，ユーザーの履歴情報を利用してどのように発話の意図を推測・補完すれば良いだろうか．ユーザーの発話を意味フレーム等の表現に落とし込んで，空のスロットを履歴を手掛かりに埋めれば良いかもしれないが，このようなアプローチには一般に多くの知識やルールが必要であり，汎用的なシステムを構築することは難しい．明確な知識表現を考えるよりも，リカレントニューラルネットワーク (RNN) のように履歴を連続空間内のパターンとして記憶し，入力した発話によって想起されるパターンから足りない情報を予測する方が簡単かも知れない．

3. 今後の音声認識研究について (堀智織)

現在，スマートフォンでクラウドコンピューティングにより超大語彙音声認識が実現されており，多くのエンドユーザが，テキスト入力に代わる簡便な入力手段として，音声入力の利便性を実感している．2003 年に実験室レベルで達成されていた技術 [7] がこのように急速に社会に普及した理由として，スマートフォンというマイクを備えた個人用デバイスと高速ネットワークの普及が挙げられるが，加えてスマートフォン上での無料アプリ公開による実証実験が行えた事も大きな理由である．このような社会的背景により，音声認識技術のより広い分野への応用に対する期待が高まっている．今後，研究開発分野では，現状の技術で「できる事」,「できない事」が明確になり，基礎研究と開発研究の分離が進むことが予想される．「できる事」は開発案件としてタスクの多様化が進み，音声認識技術を用いたシステムを構築するために多くの開発者が携わる事となるであろう．一方，基礎研究では，今まで着手してきた課題よりもさらに困難な課題に取り組まなければならない．音声認識技術の研究が約 50 年間弛まず成果を挙げ続けてきたように，今後も問題設定の難度を段階的に上げ，研究リソースの集中により，要素技術の性能向上を目指すことが肝要である．

3.1 利便性のための音声認識

メールなどの文章の音声入力，エージェントシステムと

の音声対話，異なる言語を話す話者間の対話を補助する音声翻訳など，スマートフォン上で利用可能な音声認識技術の普及が進んでいる．しかしながら，「いつでも」，「どこでも」利用可能な音声認識技術の実現からすると，音声インタフェースを用いることで利便性が向上すると期待されるアプリケーションであっても，いまだ音声インタフェースが実現されていないのが現状である．オープンソースで開示された音声認識システムや音声認識サーバへの API が公開されているにも関わらず，音声認識システムの導入を阻む問題が潜在的に存在している．第一にタスク毎に語彙が異なるという問題，第二に音声を書き起こしただけではユーザの意図を理解しシステムを駆動することはできないという問題，第三に既存のシステムとの統合が困難であるという問題がある．語彙の問題を解決するため，音声認識システムの既存の言語モデルを変更することなく，新語彙を追加できる仕組みが必要である [8]．音声認識結果からユーザの発話意図を推定し，その発話意図に基づいて処理を決定する対話制御を実装するための簡便なツールが必要である [9]．音声認識, 合成, 翻訳, その他様々な既存のサーバを接続するための通信ライブラリが必要である [10]．多言語音声翻訳により言語の壁の無い世界を実現するため，NICT は 23 ヶ国 26 研究機関が加盟する国際研究コンソーシアム U-STAR（Universal Speech Translation Advanced Research Consortium，http://www.ustar-consortium.com/）を主導し，音声認識，合成，翻訳およびネットワーク型音声翻訳の通信ライブラリの研究開発を行っている．U-STAR は，通信ライブラリのオープンソース化，音声翻訳および音声対話システムの開発キットとして，スマートフォン上のサンプルアプリと音声対話システムの設計と駆動が可能な WEB 版対話制御ビルダー，音声翻訳および音声対話サーバを 2014 年 3 月に公開する予定である．本公開により，一般の開発者が U-STAR が提供する多言語の音声翻訳や音声対話を用いて試験的なシステムを構築・運用することが可能となり，実用化が加速されることが期待される．

3.2 世界中の情報を収集するための音声認識

世界中で起きている様々な出来事の中で，我々が知り得ている情報はわずかに過ぎない．インターネット上の情報はその一部で，さらに我々が収集可能な情報は主に日本語に限られる．公に報道されている情報は，報道者のフィルターを介しているという点において，すでに偏った情報である．このような僅かで偏った情報源に基づいて，我々は世界中の様々な人間と良好な関係を築き，その関係を永く保つ必要がある．一つの出来事を異なる立場の人間がとらえれば，事実としては一つであるにも関わらず，感じ方，反応の仕方は異なる．一つの事実に対して世界の反応の動向を知り，世界の中でバランスの取れたコンセンサスを形成するため，個々人が世界中の人の意見を直接知る機会は，島国日本において必要不可欠である．音声認識技術は，世界中で配信されている音声情報に実時間で自動索引付与することで，その一助となる．世界中のニュース音声に対して音声翻訳が可能になれば，日本語による情報検索が可能となる．また，話者認識技術を用いて特定の話者の発話検索が可能になれば，テキストで報道された言葉であっても，WEB 上にその動画ないし音声が存在すれば，生の音声で聞くことができ，より直接的にメッセージの真意を知ることができる．さらに，人間の音声に留まらず，音響イベントを認識する事で，個人が掲載する動画サイトから暴動，銃撃戦などの事件，事故，災害などに関する情報が抽出でき，より早い対応が可能となる．このような技術を実現しようとした場合，一人の話者が接話マイクに向かって明瞭に話した音声を認識するスマートフォンのための音声認識技術では十分ではない．「実世界音声」と呼ばれる音声にはスタジオで話されたニュースキャスターの音声から，聴取者を意識した講演音声など条件が安定している対象が多く含まれるが，難しい問題として残されているのは，1 チャンネルで収録された複数の音響環境，複数の話者の発話や複数の音響イベントが重なった音源から情報を抽出することである．我々はこのような音声を「実世界音声」から切り出して「現場音声」と名付け，研究対象として注力している．全ての言語の「現場音声」の認識が可能となれば，言語が全く分からない外国で危険を感じた際，自分の周囲の音声や放送音声を日本語に翻訳することができ，安全を確保するのに役立つであろう．

3.3 音声認識技術は水道のようなもの

人間が社会生活を営む上で，言語によるコミュニケーションは不可欠であり，音声言語の担う役割は大きい．言語獲得，情報収集のための手段として，あるいは，言語の壁を越えたコミュニケーション手段として，音声認識技術は無くてはならないインフラである．全ての言語の音声認識を達成し音声翻訳を実現できれば，世界中の人間の英知を集約する事ができる．現在，音声認識技術の研究開発は，経済的に発展した国々の言語を対象として盛んに行われているが，教育の普及や経済発展の加速を必要とする国々の言語にも早急に取り組む必要がある． U-STAR では，少資源言語の音声認識技術の研究開発に取り組んでいるが，未だ十分とは言えない．NICT はアジアパシフィック標準化会議 ASTAP を通して，各国政府への音声認識研究への支援を強く訴えており，今後さらに音声認識技術の多言語化を推進していく予定である [11].

4. 今後の音声認識研究について (篠崎隆宏)

音声認識研究は 1950 年代頃の小規模孤立単語認識から連続読み上げ音声，自然発話音声へとタスクを高度化させながら発展してきました．そして近年のスマートフォン向

けアプリケーションの普及などにより，実用化という点で一定の目的を達成しました．しかし，では音声認識技術は十分な普及段階に入ったかというと，まだそのような状況でもありません．その一方で研究テーマとしてこれまでのようにタスクを高度化しつつ認識率を向上させることの重要性を世の中にアピールすることは，次第に難しくなっていくのではないかと思われます．

大学で研究する立場としては危機感を覚える状況ですが，同時に関連分野の発展と合わせ音声認識研究を新しいステージに前進させるための準備が整ったチャンスの時代にも思われます．すなわち，音声認識の研究対象を基礎研究としても応用研究としてもこれまでよりもより多様化した，認識率だけではないものへと深化させていく必要があるのではないかと思います．

4.1 タスクについて

認識タスクに適合した書き起こしラベルつき音声データを大量に収集し既存の技術を丁寧に作り込んでいけば改善の余地はあるにせよ実用レベルの認識性能を実現できるということは既に実証されたと言えます．応用面から見た問題点は，そのようなシステムを実現するためにはデータの収集やシステムのチューニングに膨大な手間と費用がかかることです．また一度そのようなシステムを構築したとしても言語は常に変化するため，性能を維持するためには継続的なメンテナンスが必要となります．

この問題は，現在の音声認識システムの学習技術が大量の人手による書き起こしテキストに依存していることが原因です．他方，人間の言語獲得過程を見てみればそのようなデータは必要とされておらず，音声対話等を通じて高度に自律的な学習が行われ，また環境変化にも柔軟に対応できています．少ない費用でより高度な認識性能を達成するという応用面での要請とともに，知的で自律的なシステムを構築するという技術的な観点からも，教師無し学習や半教師つき学習，あるいはそれをサポートする仕組みの発展が期待されます．

4.2 環境・デバイスについて

従来から言われていることですが，キーボードと音声入力が両方使える状況では多くの場合キーボードの方にアドバンテージがあります．この理由としては音声認識では誤認識が避けられずその修正にはキーボードを使いたくなることと，音声とキーボードの両方を交互に使用するのはメンタル的な負荷が大きいことが理由ではないかと思います．外国の友人を居酒屋に案内して日本語と英語を交互に話すとやたらに疲れますが，それと同じことの様に思います．他方テレビのスイッチを押すために 2m 歩くことが許せないという要請からリモコンが広く普及しましたが，キーボードにはテレビのスイッチと同様な不便さがあり，音声認識にはリモコンと同様な利便性を提供できる可能性があるのではないかと思います．

スマートフォンではデバイスサイズからキーボード操作に制約があり，また近年の認識性能の向上から音声認識が部分的としても実際に使われ始めていますが，今後も小型デバイスは音声認識技術の普及の上で重要な足掛かりになるのではないかと思います．腕時計型スマートデバイスなどでは，音声入力の重要性がさらに増加すると思われます．その際には，音声だけで必要な操作を完結させられるインタフェースを実現することが重要と思われます．その他，MEMS マイクと組み合わせてワンチップ実装した音声認識器を実現しセンサネットワークと組み合わせられれば，面白いのではないかと思います．特に，チップの消費電力をミリワット以下のオーダーにまで抑えられれば環境発電を電源とすることが可能となり，これまでに無い応用が開けるのではないかと期待します．

4.3 方法論について

これまでの音声認識研究は人間の音声認識能力を目標としつつも脳の中身はブラックボックスとして扱い，入力としての音声信号から出力としての単語列を推定するという機能にのみ着目してきました．しかし近年ブレインマシンインタフェース技術の進展に伴い，脳による特徴抽出の結果を計算機による識別器に繋いだ実験などが画像認識などで行われるようになりつつあります．

生体電極や赤外線などを用いた非侵襲的なニューロン活動のモニタリングは空間および時間分解能が低い問題があります．しかし動物実験において神経細胞に直接電極をコンタクトさせる手法については，動物の行動を制限することなく数万本オーダーの電極を 10kHz オーダーの周波数でサンプリングするようなことがそう遠くない将来において可能ではないかと思われます．すると，音声知覚の学習過程を長時間かつ詳細にモニタリングしたり，あるいは計算機からの信号でそれに干渉したりすることが可能となります．これにより脳の高レベルな機能をリバースエンジニアリングする新しい分野が発展し，脳機能の解明やその工学への応用が大いに進展するのではないかと期待します．またその際は，理学と工学でコラボレーションする機会も増えるのではないかと思います．

4.4 分野交流について

他分野との交流は，音声認識の発展のために重要と思います．HMM は音声認識からバイオインフォマティクスに輸出された技術ですが，HMM とニューラルネットを組み合わせたパターン認識器の有効性は長距離依存性のモデル化のためにバイオインフォマティクス分野で先行して広く認められていました [12]．今から思うと昨今の DNN ブームの前に素早くそれを逆輸入して自分で真似してみなかっ

たのは残念です．今後に向けたトピックとしては，蛋白質の分子設計などは面白そうに思います．自由エネルギー最小化問題としてグラフィカルモデルを用いて定式化できるので，適当な近似のもと音声認識における大規模探索技術や学習技術が応用できる可能性があるのではないかと思われます．

参考文献

[1] 秋葉友良, 岩野公司, 緒方淳, 小川哲司, 小野順貴, 篠崎隆宏, 篠田浩一, 南條浩輝, 西崎博光, 西田昌史, 西村竜一, 原直, 堀貴明，クラウド時代の新しい音声研究パラダイム情報処理学会研究報告, vol. 2012-SLP-92, no. 4, 2012.
[2] D. Abowd, et al., “Context-awareness in wearable and ubiquitous computing”, Virtual Reality”, vol. 3, no. 3, pp. 200–211, 1998.
[3] S. Furui, “Speech recognition technology in the ubiquitous/wearable computing environment”, in Proc. ICASSP, pp. 3735–3738, 2001.
[4] T. Hori, et al., “Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera”, IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 2, pp. 499–513, 2012.
[5] G. Hinton, et al., “Deep neural networks for acoustic modeling in speech recognition”, IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82–97, 2012.
[6] T. Mikolov, et al., “Recurrent neural network based language model”, in Proc. Interspeech, pp. 1045–1048, 2010.
[7] T. Hori, C. Hori, et al., “Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition”, IEEE Transactions on Audio, Speech and Language Processing, vol. 15, pp. 1352–1365, 2007.
[8] P.-R. Dixon, C. Hori, et al., “A Specialized WFST Approach for Class Models and Dynamic Vocabulary”, in Proc. Interspeech, 2012.
[9] E. Mizukami, C. Hori, et al., ”WFST-Based Spoken Dialogue System on Smartphones–Its Development and Implementation for Field Use,” in Proc. IEEE 14th International Conference on Mobile Data Management (MDM), 2013.
[10] http://www.ustar-consortium.com/standardization.html
[11] http://www.apt.int/
[12] P. Baldi and Y. Chauvin, “Hybrid modeling, HMM/NN architectures, and protein applications”, Neural Computation, vol. 8, no. 7, pp. 1541–1565, 1996.