Vol.2012-SLP-92 No.4
2012/7/19

クラウド時代の新しい音声研究パラダイム

秋葉 友良[1,a)] 岩野 公司[2,b)] 緒方 淳[3,c)] 小川 哲司[4,d)] 小野 順貴[5,e)] 篠崎 隆宏[6,f)]
篠田 浩一[7,g)] 南條 浩輝[8,h)] 西崎 博光[9,i)] 西田 昌史[10,j)] 西村 竜一[11,k)] 原 直[12,l)]
堀 貴明[13,m)]

概要：個人が複数の携帯情報端末を所有し，そこで得られたあらゆる音声データをクラウドに蓄積することが容易になりつつある．このように音声情報処理の周辺環境・技術が激変していく中で，音声情報処理技術のより一層の高度化が求められている．その期待に応えるためには，クラウド処理を前提とした音声研究プラットフォームの構築と，それを基盤とした新しい音声研究のパラダイムが必要である．本稿では，現在までに培われてきた音声情報処理技術を概観した上で，新しい研究パラダイムの方向性とそこで生じる新たな課題について議論する．

キーワード：音声処理，クラウド，研究パラダイム

New Speech Research Paradigm in the Cloud Era

Tomoyoshi Akiba[1,a)] Koji Iwano[2,b)] Jun Ogata[3,c)] Tetsuji Ogawa[4,d)] Nobutaka Ono[5,e)]
Takahiro Shinozaki[6,f)] Koichi Shinoda[7,g)] Hiroaki Nanjo[8,h)] Hiromitsu Nishizaki[9,i)]
Masafumi Nishida[10,j)] Ryuichi Nishimura[11,k)] Sunao Hara[12,l)] Takaaki Hori[13,m)]

Abstract: Recently most individuals have come to use mobile information devices, and daily upload the information obtained by such devices to Internet Cloud. Accordingly the applications of speech information processing have been changing drastically. We need to create a new paradigm for the research and development of speech information processing to adapt to this change. In this paper, we summarize the state-of-the-art speech technologies, propose how to create a research platform for this new paradigm, and discuss the problems we should solve to realize it.

Keywords: speech processing, Internet, Cloud, research paradigm

[1] 豊橋技術科学大学
1-1, Hibarigaoka, Tempaku-cho, Toyohashi-shi, Aichi
[2] 東京都市大学
3-3-1 Ushikubo-nishi, Tsuzuki-ku, Yokohama
[3] 産業技術総合研究所
1-1-1 Umezono, Tsukuba-shi, Ibaraki
[4] 早稲田大学
1-6-1 Nishi Waseda, Shinjuku-ku, Tokyo
[5] 国立情報学研究所
2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo
[6] 千葉大学
1-33, Yayoimachi, Inage-ku, Chiba
[7] 東京工業大学
2-12-1, Ookayama, Meguro-ku, Tokyo
[8] 龍谷大学
1-5, Yokotani, Seta Oe-cho, Otsu-shi, Shiga
[9] 山梨大学
4-3-11, Takeda, Kofu-shi, Yamanashi
[10] 同志社大学
1-3 Tatara Miyakodani, Kyotanabe-shi, Kyoto
[11] 和歌山大学
930 Sakaedani, Wakayama-shi, Wakayama
[12] 奈良先端科学技術大学院大学
8916-5 Takayama, Ikoma-shi, Nara
[13] 日本電信電話 (株)
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto
a) akiba@cs.tut.ac.jp
b) iwano@tcu.ac.jp
c) jun.ogata@aist.go.jp
d) ogawa@pcl.cs.waseda.ac.jp
e) onono@nii.ac.jp
f) shinot@chiba-u.jp
g) shinoda@cs.titech.ac.jp
h) nanjo@rins.ryukoku.ac.jp

1. はじめに

近年，計算機・通信技術の進歩に伴い，インターネットを介した情報の獲得が容易になってきている．しかしながら，多くの重要な情報は，いまだに人間同士の対話によって交換されることが多く，そこにおける情報獲得の自動化，及び，支援は重要な課題である．一方，現在，1 人が 1 台以上のスマートホンなどの携帯情報端末を所有し，テキスト情報のみならず，獲得した音声などのマルチメディア情報を即座にクラウドに蓄積することが容易になりつつある．このように音声処理の周辺環境・技術が激変していく中で，音声研究においてもその新しいパラダイムが求められている．

我々は，その問題意識のもと，2012 年 5 月に情報処理学会音声言語処理 (SLP) 研究会のワーキンググループとして，「音声・音響クラウド WG」を立ち上げた．このワーキンググループは，対話・ミーティングなどで，その参加者各々の携帯情報端末から得られる情報を統合して利用することで，そこにおける有益な情報を獲得する手段を提供することを目的としている．より具体的には，多数の機器で収音された，ミーティング，学会講演の質疑，TV における対話音声，会議音声，音楽やその他の背景音下の音声などを対象とし，そこから， Transcription, 言語, 話者, 話者位置などの情報や，対話の情報 (会話が弾んでいるかなどの雰囲気など) を獲得するための技術を開発する．アプリケーションとして，マルチメディア議事録の自動作成，会話からの情報マイニングなどを想定している．特に関連する音声処理分野として，以下の 5 分野があげられる．

- 音声ドキュメント処理
- 音響イベント認識，耐雑音処理
- 話者 Diarization
- 複数マイクによる音響信号処理
- 協調的音声・音響アノテーション

音声・音響 WG では，この目的の達成のために，多くの研究機関の様々な分野の研究者が結集し，新しい音声研究パラダイムの構築，及び，そこにおける新たな課題の策定とその解決を図る．当面の活動として，共通評価基盤 (データベース，ソフトウエア）の整備を行う．様々な環境でのミーティング音声，携帯端末音声の収録を行い，評価のためのテストセットコレクションを作成する．

以下，各分野ごとに，クラウド時代における音声研究とその展望について述べる．(文責 篠田)

i) hnishi@yamanashi.ac.jp
j) mnishida@mail.doshisha.ac.jp
k) nisimura@sys.wakayama-u.ac.jp
l) hara@is.naist.jp
m) hori.t@lab.ntt.co.jp

2. 音声ドキュメント処理

音声認識や音声ドキュメント検索などの音声ドキュメント処理技術は，クラウドを活用するための基盤技術と捉えることができる [1]．せっかくインターネットを介して膨大な音声・映像データにアクセスできるにもかかわらず，それらを効率良く検索・分類する技術がなければ，クラウドを活用した音声情報処理は難しい．

本章では，クラウド時代の音声研究を支える音声認識技術ならびに音声ドキュメント検索技術について解説する．

2.1 音声認識

音声認識は，音声研究の根幹となる技術である．例えば，音声対話・言語理解や音声ドキュメント検索を行う際には，音声認識性能がそれらのパフォーマンスに大きく影響することから，音声認識技術の更なる発展を目指す必要性があることは，疑う余地はない．

音声認識技術については，音響・言語モデリングやデコーディング技術など様々な要素技術から成り立っている．現在までに，国内外の多くの会議や論文誌において，これらの技術報告がなされている．ここ近年では，ICT の発展により大量のデータを扱えるようになり，また，それに伴い効率の良い学習アルゴリズムが数多く開発されている．

では，これからの時代の音声認識はどうなるだろうか．テレビや雑誌では，ビッグデータというキーワードが注目されている．音声認識の分野でも，数年前から，WEB 上の大量のテキストデータを利用することで，言語モデルを学習・適応化する研究も始まっている [2]．また，Google は語るに及ばないが，日本でも例えば NTT ドコモは，通訳電話サービスにおいて，不特定多数の利用者から集めた大量のデータやログを，多様な声質・アクセント・速さの音声をより高精度に音声認識するために利用している．[*1]

今後は，様々な人間が，あらゆる場所で発話した音声を，あらゆる種類のマイクロフォンで取得し，それを高精度で認識する必要がでてくる．これらに対応するために必要な事前処理やモデル学習等を行うため，多様な音声情報を含む大規模な音声データベースの構築は必要不可欠である [3].

2.2 音声ドキュメント検索

1 つのミーティングや音声メモ，テレビ番組，講義等，音声を含むデータを音声ドキュメントと呼び，複数あるいは大量の音声ドキュメントがある中で，クエリ (検索要求) に関連する音声ドキュメントを特定することを音声ドキュメント検索 (SDR) と呼ぶ．また，特定の単語やフレーズが音声ドキュメント群中のどのドキュメントのどの位置に含ま

[*1] 2012 年 6 月 4 日付日本経済新聞に関連記事が掲載.

れているのかを特定する研究を音声中の検索語検出 (STD) と呼んでいる．

SDR の研究は，TREC[4] において，1996 年に SDR トラックによって取り上げられた．これを機に，音声ドキュメント検索の研究が推進・活性化している．STD の分野では，2008 年，アメリカ国立標準技術研究所 (NIST) の主導の下テストコレクションが制定され，競争型ワークショップが開催された [5]．特に未知語を検出する研究を中心に盛んに研究が行われている．

SDR・STD 技術では，クラウド上に蓄積されたデータから意味のある情報を取り出す音声データ (テキスト) マイニング技術に利用可能である．特に，STD 技術は，未知語の自動獲得 [6] 等に利用でき，話題依存言語モデリングにも応用されつつある．さらに，SDR 技術は，発話を意味概念や話題毎にクラスタリングするのに応用できると考えられる．また，話者検索を利用することで，音響モデルの話者適応化に応用できそうである．

一方，日本でも，2006 年に情報処理学会音声言語情報処理研究会音声ドキュメント処理 WG が設立され，日本語を対象とした SDR・STD のテストコレクションの整備が行われた．これらのテストコレクションでは，日本語話し言葉コーパス (CSJ) を検索対象の音声ドキュメントとして利用している．また，同 WG が主体となって，情報検索技術を評価する競争型ワークショップである NTCIR-9 に，SDR と STD タスクを提案し，採択されている [7]（2013 年開催予定の NTCIR-10 にも採択されている）．

このように，この 10 年の間に，SDR や STD 技術が数多く提案され，大きな研究成果が挙げられている．今後は大量の音声データの中から意味ある情報を抽出する音声データ (テキスト) マイニングを行うことを考えると [8]，発話をその意味概念毎や話題毎にカテゴリ化し，カテゴリ毎に自動的に概念や話題のメタデータを付与する技術も必要となる [9]．

以上のことから，今後の音声ドキュメント検索に必要な技術として：

- クエリに内容が似ているドキュメントの検索技術 (これまでの SDR の延長)
- 特定の単語やフレーズが含まれている音声を特定する技術 (STD)
- 多様な発話を意味概念や話題毎にクラスタリングする技術 (SDR・STD の技術が利用できる)

などが挙げられよう．(文責 秋葉，南條，西崎)

3. 音響イベント認識・耐雑音処理

ユーザー個人がもつ携帯情報端末を利用して対話や会議音声を即座にクラウドに蓄積できる環境では，多くの場所で点在して繰り広げられている対話や会議の情報を大量に集約し，それを各ユーザーが相互に有効活用することが可能になる．そこでは議決を目的とする統制のとれた会議だけでなく，ワークショップやイベント会場における数名の立ち話のような小規模で即興的な意見交換など，様々な目的レベルの対話・会議が処理対象となってくる．例えば，会議議事録や自分がいつどこでだれと対話を行ったかの履歴情報を自動作成できれば便利であろう．さらには，様々な場所でどのようなキーワードやトピックがどのような状況で話し合われ拡散しつつあるかを地図上で表示する新しいコミュニケーション視覚化システムなども考えられる．これらを効果的に実現するためには，以下に示す様々なイベントを音響情報から抽出することが必要となる．

対話グループ検出　ユーザー同士が直接会話し携帯端末を傍聴的なマイクロホンとして利用する場合，物理的に近接するユーザーのうち 2 名以上のどの端末の持ち主同士が実際に対話を行っているのかという情報は自明ではない．音声のクロストークや発話内容を基にして対話グループを同定する技術が求められる．

キーワード・トピック検出　対話・会議の内容を素早く把握，検索するための情報として，発言キーワードやトピック情報を検出する．

盛り上がり検出　従来までに議論の白熱した部分と韻律情報やクロストークの生起が深い関係性を有していることが報告されている [10]．また，笑い声やあいづちが活発に生起する部分を検出する手法 [11] や，議論の「発散／収束」を非言語情報を利用して判別する手法 [12] などが提案されており，これらの統合的な利用が有用であろう．関連して，発話者の感情をイベントとしてメタデータ化することも有益である可能性があるため，感情認識（[13] など）についても検討を要する．

目的検出　それぞれの対話・会議がどのような目的（議決・意見交換・報告など）で行われているかを表すマクロな情報も，対話・会議の検索などに有益であると考える．

これら各種イベントを高精度に検出するためには耐雑音性の高い音声認識が必要であり，5 章で後述する複数マイクを用いた音声信号処理技術の応用が期待される．場合によっては携帯端末で撮影される動画像情報を利用したマルチモーダル音声認識 [14] の活用も検討事項となる．また，収集された音声を利用した語彙や言語モデルの教師なし適応や，そのためのトピックのクラスタリング技術も重要と考えられる．(文責 岩野，篠崎)

4. 話者認識・Diarization

話者 diarization とは，ミーティングのように複数人が発話している音声から「いつ，誰が発話したか」(“Who spoke When?”) を推定する問題であり，推定された情報は，特定話者の発話の検索や，音響モデルの話者適応への利用が期待できる．また，話者 diarization は，1) 発話区間の検

出（Voice Activity Detection: VAD），2) 得られた発話に対する話者交替の検出（話者セグメンテーション），3) クラスタリング（話者クラスタリング）という一連の処理からなる．

近年の ICASSP (2007 年から 2012 年まで）において発表された話者 diarization に関連する 27 件の論文を対象に，評価データ，評価尺度，手法の観点で近年の傾向について述べる．評価データについては，ミーティングを対象としたものが 14 件でそのうち NIST Rich Transcription (RT) Evaluation を用いたものが 11 件，Augmented Multi-Party Interaction (AMI) Corpus を用いたものが 3 件，ニュース番組を対象としたものが 5 件，電話対話を対象とした NIST Speaker Recognition Evaluation (SRE) を用いたものが 3 件であり，大規模かつ公開されたデータを用いた評価が求められていると言える．評価尺度については，話者 diarization では非音声を音声または音声を非音声と誤った検出と話者の誤りを時間長に基づいて算出した Diarization Error Rate (DER)，話者クラスタリングでは DER，発話の分類精度に着目した Purity や Rand Index など，話者セグメンテーションでは話者交替の検出精度として False Acceptance Rate (FAR) と Missed Detection Rate (MDR) が用いられている．手法については，大まかに分類するとモデル学習・選択に焦点を当てたものが 12 件，特徴抽出が 8 件，システム統合が 4 件，距離尺度に関するものが 2 件であった．例えば，近年話者照合において盛んに検討されている，話者やチャネル情報を含んだベクトルから因子分析に基づき話者情報のみを抽出した特徴表現が話者 diarization でも用いられている [15]．音響情報以外の情報を用いる研究としては，音響情報に加えて話者交代のパターンを陽にモデル化する方式 [16] が提案されており，Multiple distant microphones (MDM) 条件においては，複数のマイクによる音声の到来時間差 (Time Differences Of Arrival: TDOA) が特徴として一般的に用いられている [17]．また，オンライン処理を扱うものとして，ニュース番組を対象として任意の時間長ごとに音素クラスごとの BIC (Bayesian Information Criterion) 値を用いて話者セグメンテーションならびにクラスタリングを行う手法 [18] や，エルゴディック HMM の状態を話者モデルとみなし，状態の増加をデータに応じて変分ベイズ基準で決定する方式 [19] などが提案されている．さらに，話者 diarization システムを用いて音響単位の自動生成を行うことで，zero resource speech recognition に寄与しようという研究も見られる [20].

今後は，話者 diarization の結果に対する特定話者の検索，複数話者の発話が重畳したオーバーラップの検出ならびに話者の分離に関する研究が進むと考えられる．方式開発という観点からは，本 WG が対象とするような大規模かつ実環境音声データを扱う場合，雑音，話者，発話スタイルといった変動に対して頑健であることが必要要件となる．そのためには，データに応じて効率的にシステムを修正可能な方式の開発，BIC と i-vecter を組み合わせた 2 段階のクラスタリング [21] や，映像情報と音響情報の統合 [22] など，相補的な特徴量やシステムを統合する方式の開発が効果的と考えられる．しかしながら，音響的な特徴表現においても，従来用いられている MFCC に代わる，真に話者を表す特徴の表現・抽出方式に関する研究が依然として重要と考えられる．また，MDM 条件においては空間情報の利用が可能であるため，高精度な音源位置推定，音源分離方式との統合が重要である．（文責 小川，西田）

5. 複数マイクによる音響信号処理

マイクロホンアレイに基づく音響信号処理は，遠隔発話音声認識の有効なフロントエンドとして期待されている．遠隔マイクが使用可能であれば，音声認識を利用したアプリケーションの利便性は大きく向上する．例えば，ロボットとの対話や自動書き起こしを前提とした会話において，マイクを装着せずにいつでも会話を始められる即時性が生まれ，かつマイクの前で話さなければならないといった発話場所の制約も少なくなる．しかし，遠隔マイクの使用は，複数の話者や雑音源が混在する実環境において，S/N 比の低下や音声信号のオーバラップによる認識精度の著しい低下を招く．更に，ヘッドセットマイクやピンマイクのように各話者との明確な対応付けがなされないことから，話し手が分からなくなる問題もある．複数人の会話を扱うアプリケーションでは，音声認識だけでなく，話し手も同時に認識することが求められるため，話者ダイアライゼーション等の技術が必要になる．

マイクロホンアレイを用いれば，ビームフォーミングなどのアレイ信号処理により，目的の音源 (話者) の信号を強調し，背景音 (雑音や他の話者の音声) を抑圧することで，各話者の音声信号を比較的高い精度で推定することができる [23]．近年，話者の位置情報が先見的にわからなくても混合音から個々の話者の音声信号を分離するブラインド音源分離の技術が発展してきた．話者ごとに分離された信号を音声認識すれば，多少のオーバラップがあっても話者ごとの認識結果を得ることができる [24][25]. ステレオマイクの使用を前提に，生活環境雑音下で音声認識性能を競う国際コンテストも試みられている [26]．更に，話者ダイアライゼーションでは，参照マイクと各マイクの信号間の到来時間差 (TDOA) や信号の到来方向 (Direction of Arrival : DOA) を特徴量とすることで同定精度が大きく向上することが知られている [27]．このような音源の空間情報を単一のマイクで取得することは難しい．

このように，マイクロホンアレイに基づく音響信号処理は効果的であるが，ヘッドセットマイク使用時の音声認識精度には未だ及ばない．ヘッドセットマイクに匹敵する認

識精度を得るには，音声強調処理の高度化に加えて，後段の音声認識との密接な連携が重要になる．例えば，強調音声と音声認識用音響モデルとのミスマッチを軽減する環境適応や，強調音声の劣化度合に応じた分布の動的制御を行う Uncertainty Decoding [28] などのアプローチが提案されている．更に今後は，音声強調と音声認識を一体で考え，認識誤りを最小にするような全体最適化の枠組が必要であろう．

一方，通常のマイクロホンアレイ信号処理では，マイクロホンを規則的に配置し，これらを厳密に同期した多チャンネル A/D 変換器に接続することによってシステムが構成されるが，性能向上のためにマイクロホン数を増やしたり，マイクロホンを広範囲に配置するには，大きなコストを生じてしまう．よって，ユーザが持っているスマートフォンやタブレット端末といった個別の録音機器を利用してアレイ信号処理を行うことは，近年の魅力的なシナリオの 1 つであり，distributed microphone array, もしくは ad hoc microphone array などと呼ばれている．ここで問題となるのは，録音チャンネル間の同期性が保証されないことである．マイクロホンアレイ信号処理では，各マイクロホンで録音される信号間の微小な時間差（例えば，3.4cm の伝播経路差に対し 100μs．16kHz サンプリングの 1.6 サンプルに相当）が音源の空間情報の主要な手がかりとなっているが，個別の録音機器を用いた場合には，録音開始時刻やサンプリング周波数を同期させることは難しい．また，話者位置を推定するためには，録音機器自体の位置推定も必要となる．こうした問題に対して，無線を用いて同期信号を供給する手法 [29]，みかけの到来時間差 (TDOA) に基づき，音源位置，マイクロホン位置，録音開始時刻を同時に推定する手法 [30][31]，信号の観測エネルギーに基づき，厳密な時間同期を必要とせずに音源位置とマイクロホン位置を推定する手法 [32][33]，また，非同期アレイを用いたブラインド音源分離 [34][35] やビームフォーミング [36] などが試みられている．

こうした非同期アレイ信号処理技術が発展すれば，将来的には，会議の参加者が会議の始まりと同時に各自の携帯端末で録音を開始し，会議が終了したら各自が録音した音声ファイルをサーバにアップすると，複数非同期録音に基づき，音源分離，雑音除去，残響除去，音声認識を経て，会議の議事録が会議の参加者に自動的に送信されるといった，自動議事録作成システムが実現できるかもしれない．(文責 小野，堀)

6. 協調的音声・音響アノテーション

ウェブの高度化とともに始まったクラウド時代，キーボードとマウスを代表とする伝統的なインタフェースの一部が，これまでは次世代の技術と考えられてきた音声認識・音声合成や対話の音声インタフェースに置き換わろうとしている．クラウド環境は，計算機リソースの制限からの解放を我々にもたらしたが，音声言語資源へのメタデータの付与（アノテーション）には依然として大きな人的コストを要し，むしろ，ビックデータ化に伴って，アノテーションの対象も爆発的に増加，その取扱いが難しくなっている側面もある．また，音声言語が潜在的に含んでいる話者や収録環境，言語等に関する多様性は極めて広く，クラウドの莫大なリソースを利用しても，そのすべてを容易に包括することはできない．クラウド時代のアノテーションにおいては，これまでの話しの中心でもあった**計算機リソース**活用としての「クラウド (**Cloud**)」だけではなく，大規模な**人的リソース**を有効活用するという意味での「クラウド (**Crowd**)」が重要な観点となる．以上のような問題を踏まえ，「協調的音声・音響アノテーション」グループの役割は，現代のウェブやソーシャルネットワーク等で注目されている，協調的アノテーションの方法を採用することで，従来から引き継がれるデータベースやコーパスの構築手法ではカバーすることが難しかった，様々な多様性に対応する知識の集積方法を確立することにある．

PodCastle[37], [38], [39] は，ウェブ上の動画等音声を含むコンテンツ（音声コンテンツ）の検索を可能にしたウェブサービスである．アノテーション技術の立場から特に注目すべきは，音声認識の誤り訂正機能にある．PodCastle では，ウェブ上の音声コンテンツの書き起こしを音声認識により自動生成し，さらには認識性能の向上をはかるために，ユーザからのアノテーションを通じた積極的協力を得る仕組みを提供している．図 1 に示すように，訂正インタフェースは，競合候補のリストをユーザに提示し，ユーザは認識誤りを見つけたときに，「候補選択」「タイプ入力」のどちらかの手段で訂正を行うことができる．特に人気のあるチャンネルでは，より多くの訂正が行われるため，チャンネル別の音響・言語モデルの学習効果も高まる．その結果，精度の高い音響・言語モデルが構築され，チャンネル特有の語彙やフレーズ等の認識が可能になっている [39]．

Amazon Mechanical Turk (MTurk)[40] のように，ウェブを通じた不特定多数の人々に対する業務委託（クラウドソーシング）を利用した試みは増えている．NAACL-2010 では Creating Speech and Language Data With Amazon's Mechanical Turk というワークショップが開かれている [41]．また，Interspeech2011 においてもスペシャルセッション Crowdsourcing for Speech Processing で，10 件の関連発表がなされている [42]．クラウドソーシングの適用事例は，(1) 音声データの収集，(2) 音声データラベリング・書き起こし，(3) システム性能評価に分類できる．日本国内向けのデータ収集手段としても，同様に，安価なウェブサービスを利用することができる [43]．これらの有償のクラウドソーシングは，協調的アノテーションによる知識を集積する上で有効な手段であるが，一方で，PodCastle は，上記

図 1 PodCastle の音声訂正インタフェース（全文モード及び詳細モード），文献 [38] より引用.

(2) の音声データ書き起こしをボランティアベースで成功させている．また，MusicNavi2 は，楽曲検索のための音声認識インタフェースを利用者に提供することで，ネットワークを介した (1) 音声データの収集を実現している [44].

ここで，無償での協力を利用者から得るユーザ参加型サービスの成否は，利用者のモチベーションの維持が鍵となる．サービスやシステムそのもの利用者にとって魅力的であり，「使ってみよう」という社会的な気運を高めたうえで，それらを維持することが重要である．

Social IME[45] は，ユーザ参加型の日本語入力システムであり，インターネットを介して辞書を共有し，世界中の利用者が単語を登録することで辞書が自律的に成長する．これにより，専門用語や流行語などの，従来型の日本語入力システム（IME）では変換できない単語を変換することができるようになる．日常的なキータイプ作業の中で性能向上を実感し，利用者間で共有することができる点が利用継続の動機となっている．

協調的映像アノテーションを学術的に取り上げた Synvie[46] は，付与されたタグの集合をタグクラウドの形式で表示し，共有することで，利用者のアノテーション意欲の向上を得ている [47]．このことは，ニコニコ動画や YouTube などの動画共有サイトにおける，タグやコメント等の付与にユーザ参加型サービスを導入した成功事例からも理解できる．動画の内容への関心に加え，付与されたタグやコメントが新たなる演出効果や利用者間のコミュニケーションを促すことで，サービス全体の魅力を高めているためである．

また，非日常的な作業として「医療分野向け用例の評価」を対象とし，そのモチベーション維持支援を試みた「用例の森」[48] 等，ユーザ参加型サービスを活用するためのモチベーション維持に関する研究がはじまっている．

以上のように，利用者のモチベーションを維持するための工夫は，これまでの音声・音響分野では研究対象として見なされることの少なかった観点であり，「協調的音声・音響アノテーション」の方法確立に向けた重要な課題となる．(文責 緒方，西村，原)

7. おわりに

以上，各分野におけるサーベイと将来の展望を述べた．音声・音響クラウド WG は，次世代の音声研究を産み出すインキュベーションセンターとしての役割を負う．多くの新しい音声研究課題とその解決のための新しいアプローチを創出したい．なお，本 WG に参加を希望される方はぜひメンバに声をおかけ頂きたい．

参考文献

[1] 中川聖一, “音声ディクテーションから音声ドキュメント処理へ”, 日本音響学会 2007 年秋季研究発表会講演論文集, 1-3-1, pp.1–4, 2007.

[2] R. Masumura, S. Hahm, A. Ito, “Training a language model using webdata for large vocabulary Japanese spontaneous speech recognition,” *Proc. INTERSPEECH2011*, pp. 1465–1468, 2011.

[3] 古井貞熙, “第 4 世代の音声認識を目指して,” 電子情報通信学会誌, vol. 95, no. 5, pp. 422–426, 2012.

[4] J. S. Garofolo et al., “The TREC spoken document retrieval track: a success story,” *Proc. Text Retrieval Conference (TREC) 8*, pp. 16–19, 2000.

[5] NIST, “The Spoken Term Detection (STD) 2006 evaluation plan”. `http://www.itl.nist.gov/iad/mig/tests/std/2006/docs/std06-evalplan-v10.pdf`

[6] P. Carolina et al., “A spoken term detection framework for recovering out-of-vocabulary words using the Web,” *Proc. INTERSPEECH2010*, pp. 1269–1272, 2010.

[7] T. Akiba et al., “Overview of the IR for spoken documents task in NTCIR-9 Workshop,” *Proc. 9th NTCIR Workshop Meeting*, pp. 223–235, 2011.

[8] 那須川哲哉, “テキストマイニング：テキスト化された音声データによる価値創出の可能性,” 第 6 回音声ドキュメント処理ワークショップ論文集, 2012.

[9] W. Qu and K. Shirai, “Sounds of Speech Based Spoken Document Categorization: A Subword Representation Method,” *IEICE Trans. Info. Systems*, vol. E87-D, no. 5, pp. 1175–1184, 2004.

[10] O. Cetin, and E. Shriberg, “Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: insights for automatic speech recognition,” *Proc. INTERSPEECH2006*, pp. 293–296, 2006.

[11] 河原達也, 須見康平, 緒方淳, 後藤真孝, “音声会話コンテンツにおける聴衆の反応に基づく音響イベントとホットスポットの検出, ” 情報処理学会論文誌, vol. 52, No. 12, pp. 3363–3373, 2011.

[12] 市野順子, 田野俊一, “発言の時系列的パターンを用いた会議における発散／収束の判別の可能性,” 人工知能学会論文誌, vol. 25, no. 3, pp. 504–513, 2010.
[13] J. H. Jeon, R. Xia, and Y. Liu, “Sentence level emotion recognition based on decisions from subsentence segments,” *Proc. ICASSP2011*, pp. 4940–4943, 2011.
[14] 吉川正祥, 篠崎隆宏, 岩野公司, 古井貞熙, “軽量な画像特徴量を用いたマルチモーダル音声認識, ” 電子情報通信学会論文誌 D, vol. J95-D, no. 3, pp. 618–627, 2012.
[15] H. Aronowitz, “Speaker diarization based on GMM supervectors and unsupervised intra-speaker variability modeling,” *Proc. ICASSP2010*, pp. 4402–4405, 2010.
[16] F. Valente et al., “Speaker diarization of meetings based on speaker role N-gram models,” *Proc. ICASSP2011*, pp. 4416–4419, 2011.
[17] D. Vijayasenan et al., “Speaker diarization of meetings based on large TDOA feature vectors,” *Proc. ICASSP2012*, pp. 4173–4176, March 2012.
[18] T. Oku et al., “Low-latency speaker diarization based on Bayesian information criterion with multiple phoneme classes,” *Proc. ICASSP2012*, pp. 4189–4192, 2012.
[19] T. Koshinaka et al., “Online speaker clustering using incremental learning of an ergodic hidden Markov model,” *Proc. ICASSP2009*, pp.4093–4096, 2009.
[20] M. Huijbregts et al., “Unsupervised acoustic sub-word unit detection for query-by-example spoken term detection,” *Proc. ICASSP2011*, pp. 4436–4439, 2011.
[21] J. Silovsky et al., “Speaker diarization of broadcast streams using two-stage clustering based on i-vectors and cosine distance scoring,” *Proc. ICASSP2012*, pp. 4193–4196, 2012.
[22] G. Friedland et al., “Multi-modal speaker diarization of real-world meetings using compressed-domain video features,” *Proc. ICASSP2009*, pp.4069–4072, 2009.
[23] X. Anguera, C. Wooters, and J. Hernando, “Acoustic beamforming for speaker diarization of meetings,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, pp. 2011–2022, 2007.
[24] H. K. Maganti, D. Gatica-Perez, and I. McCowan, “Speech enhancement and recognition in meetings with an audio-visual sensor array,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, pp. 2257–2269, 2007.
[25] T. Hori, S. Araki, T. Yoshioka, M. Fujimoto, S. Watanabe, T. Oba, A. Ogawa, K. Otsuka, D. Mikami, K. Kinoshita, T. Nakatani, A. Nakamura, and J. Yamato, “Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, No. 2, pp. 499–513, 2011.
[26] http://spandh.dcs.shef.ac.uk/projects/chime/challenge.html
[27] C. Wooters and M. Huijbregts, “The ICSI RT07s speaker diarization system,” *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007*, pp. 509–519, 2008.
[28] J. Droppo and A. Acero, “Uncertainty decoding with SPLICE for noise robust speech recognition,” *Proc. ICASSP2002*, vol. I, pp. 57–60, 2002.
[29] R. Lienhart, I. Kozintsev, S. Wehr, and M. Yeung, “On the importance of exact synchronization for distributed audio processing,” *Proc. ICASSP2003*, pp. 840–843, 2003.
[30] V. C. Raykar, I. V. Kozintsev, and R. Lienhart, “Position calibration of microphones and loudspeakers in distributed computing platforms,” *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 1, pp. 70–83, 2005.
[31] N. Ono, H. Kohno, N. Ito, and S. Sagayama, “Blind alignment of asynchronously recorded signals for distributed microphone array,” *Proc. WASPAA*, pp.161–164, 2009.
[32] Z. Liu, Z. Zhang, L. -W. He, and P. Chou, “Energy-based sound Source Localization and Gain Normalization for Ad Hoc Microphone Arrays,” *Proc. ICASSP2007*, pp. 761–764, 2007.
[33] M. Chen, Z. Liu, L. -W. He, P. Chou, and Z. Zhang, “Energy-based position estimation of microphones and speakers for ad hoc microphone arrays,” *Proc. WASPAA*, pp. 22–25, 2007.
[34] Z. Liu, “Sound source separation with distributed microphone arrays in the presence of clock synchronization errors,” *Proc. IWAENC*, 2008.
[35] T. Ono, S. Miyabe, N. Ono, and S. Sagayama, “Blind source separation with distributed microphone pairs using permutation correction by intra-pair TDOA clustering,” *Proc. IWAENC*, 2010.
[36] I. Himawan, I. McCowan, and S. Sridharan, “Clustered blind beamforming from ad-hoc microphone arrays,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 4, 2011.
[37] 緒方淳, 後藤真孝, 江渡浩一郎, “PodCastle: ポッドキャストをテキストで検索, 閲覧, 編集できるソーシャルアノテーションシステム,” *Proc. WISS2006*, pp. 53–58, 2006.
[38] 緒方淳, 後藤真孝, “Web で生きる／活きる音声認識,” 日本音響学会 2012 年春季研究発表講演論文集, pp. 49–52, 2012.
[39] J. Ogata and M. Goto, “PodCastle: collaborative training of acoustic models on the basis of wisdom of crowds for podcast transcription,” *Proc. INTERSPEECH2009*, pp. 1491–494, 2009.
[40] https://www.mturk.com/mturk/welcome .
[41] C. Callison-Burch and M. Dredze, “Creating speech and language data with Amazon’s Mechanical Turk, *Proc. NAACL HLT 2010 Workshop*, pp.1–12, 2010.
[42] G. Parent and M. Eskenazi, “Speaking to the Crowd: looking at past achievements in using crowdsourcing for speech and predicting future challenges,” *Proc. INTERSPEECH2011*, pp. 3037–3040, 2011.
[43] 西村竜一, 宮森翔子, 鈴田健太郎, 河原英紀, 入野俊夫, “安心ウェブの実現に向けた大人・子ども発話のネット収集実験,” 情報処理学会研究報告, 2009-SLP-77-19, 2009.
[44] 原直, 宮島千代美, 伊藤克亘, 武田一哉, “多様な音響環境下における音声認識システム利用時のデータ収集システム,” 電子情報通信学会論文誌, vol. J90-D, no. 10, pp. 2807–2816, 2007.
[45] http://www.social-ime.com/ .
[46] 山本大介, 増田智樹, 大平茂輝, 長尾確, “映像を話題としたコミュニティ活動支援に基づくアノテーションシステム,” 情報処理学会論文誌, vol. 48, no. 12, pp. 3624–3636, 2007.
[47] 山本大介, 増田智樹, 大平茂輝, 長尾確, “タグクラウド共有に基づく協調的映像アノテーション,” 人工知能学会論文誌, vol.25, no.2, pp. 243–251, 2010.
[48] 狩野翔, 福島拓, 吉野孝, “用例評価のモチベーション維持支援システム「用例の森」の開発と評価,” 情報処理学会論文誌, vol.53, no.1, pp. 138–148, 2012.