Vol.2014-SLP-100 No.4
2014/1/31

SIG-SLP 第 100 回記念シンポジウム: ショート発表

齋藤 大輔	東京大学大学院情報学環
秋田 祐哉	京都大学
市川 賢	名古屋大学情報科学研究科
柏木 陽佑	東京大学大学院工学系研究科
川渕 将太	名古屋大学情報科学研究科
小林 和弘	奈良先端科学技術大学院大学
白鳥 大樹	山梨大学
鈴木 直人	東北大学大学院
田中 宏	奈良先端科学技術大学院大学
田中 宏季	奈良先端科学技術大学院大学
千葉 祐弥	東北大学
張 聡穎	東京大学大学院工学系研究科
ポンキッティパン ティーラポン	東京大学大学院工学系研究科
中島 陽祐	名古屋工業大学
長野 雄	東北大学 電気通信研究機構
西田 昌史	同志社大学
西村 良太	名古屋工業大学 ながれ領域
橋本 浩弥	東京大学大学院工学系研究科
原　直	岡山大学
Sangeeta Biswas	Department of Computer Science, Tokyo Institute of Technology
増村 亮	NTT メディアインテリジェンス研究所
松山 洋一	早稲田大学
森勢 将雅	山梨大学
吉野 幸一郎	京都大学
Yuan Liang	Department of Computer Science, Tokyo Institute of Technology

あらまし 本セッションでは，音声言語情報処理研究会の第 100 回にあたり,「これまでの SLP」を支えてきた研究者,「SLP の現在」を構成する研究者，そして「これからの SLP」を作っていく若手研究者の，情熱と意欲に満ちた研究を，ショットガン形式（短時間での連続口頭発表）で紹介する．

はじめに

齋藤 大輔（モデレータ）

今回，情報処理学会音声言語情報処理研究会（SIG-SLP）は，記念すべき第 100 回を迎え，その記念シンポジウムとして,「SLP の過去」,「SLP の現在」，そして「SLP の未来」を考える様々な企画が予定されている．本企画は「SLP の現在」に足を据えつつ,「SLP のこれまで」および「SLP のこれから」の要素も盛り込まれたセッションを期待し，ショットガン形式によって数多くの研究を概観していく企画となっている．

ショットガン形式の発表は通常，ポスターセッションの前段に設置されて，ポスターでの発表内容をシンプルにまとめた紹介発表になることが多い．一方，今回はショート発表の時間内で研究のエッセンスを伝える必要があり，発表者の皆さんにとってはなかなか骨の折れるセッションになるかと予想される．しかしながら，本稿提出時点において 24 件の申し込みがあり，参加者にとっては「SLP の現在」における，バラエティーに富んだ研究の数々に触れることができる，よい機会になるのではないかと思う．本セッションを機にディスカッションが白熱する事を切に願っている．

最後に，第 100 回記念の研究会において，本セッションのモデレータという，栄誉ある機会を与えて下さり，様々な助言を頂いた幹事・運営委員の皆様に感謝申し上げたい．

クエリ拡張と音節認識結果を併用した音声ドキュメント検索

市川 賢

音声・画像・ビデオなどの，音声を含むマルチメディアデータが増加している．これらにはメタデータとしてファイル名やタイトルしか付与されていないことが多く，それを対象として検索する従来の検索技術のみでは目的のデータにたどり着くのに限界がある．そこでデータの内容に基づいた検索を可能とする手法として，データ (音声ドキュメント) 中の音声に対して音声認識を行い，音声をテキストデータとして書き起こし，それらに対し，テキストで検索クエリを与え検索する方法がある．このような音声言語情報を対象とした検索技術は「音声ドキュメント検索」とよばれ，マルチメディアコンテンツが増加しているいま，必要不可欠な技術となりつつある．音声ドキュメント検索では，従来の誤りの少ないテキストを対象とした検索とは異なり，検索対象に未知語や音声の認識誤りを含むといった問題がある．そこで，未知語や認識誤りに対して頑健な検索システムを構築する必要があると考えられる．情報検索の代表的な検索モデルに，ベクトル空間モデル，クエリ尤度モデル，適合モデルがある．本研究では，これらの検索モデルに対し，統一的な枠組みで改良を加え，未知語や認識誤りに対処する．まず未知語に対し，新たなクエリ拡張手法を用いる．クエリと，クエリからの Web 検索結果から，検索対象をより精細にモデル化する手法を提案する．また未知語と認識誤りに対し，音節単位の認識に基づく手法を用いる．音節認識結果の音節 3-gram を索引語として用い，ランキングの際のスコアとして単語認識結果のスコアと効果的に組み合わせる手法を提案する．実験の結果，3 つの検索モデルに対し，新たなクエリ拡張手法を用いることで，従来の手法を上回る検索性能を得た．さらに，クエリ拡張と音節認識結果を併用することで，NTCIR-9 SpokenDoc タスクの公式結果と比較し，大幅に上回る結果が得られた．

※共著：　柘植覚　北岡教英　武田一哉　北研二

Deep Neural Network を用いたクリーン音声状態識別による雑音環境下音声認識

柏木 陽佑

本発表では，ニューラルネットワークを利用して従来の区分的線形変換による特徴量強調手法を拡張する．雑音環境下における音声認識において，どのように耐雑音性を高めるかが非常に重要な課題である．耐雑音処理には様々なアプローチがあるが，その一つとして特徴量ドメインで雑音の影響を除去する特徴量強調がある．統計的特徴量強調手法の一つである SPLICE (Stereo-based piecewise linear compensation for environments) はノイジー音声の特徴量分布を混合ガウス分布によりモデル化し，それより得られる事後確率を重みとして線形変換によりクリーン音声特徴量を推定する．しかし，雑音により音声特徴量は縮退するためにノイジー音声特徴量の分割が必ずしもクリーン音声特徴量の推定に適しているとは言えない．一方，ニューラルネットワークを利用した特徴強調手法として DAE (Denoising AutoEncoder) がある．これはクリーン音声特徴量をノイジー音声特徴量から直接ニューラルネットワークによって回帰することで推定する．Deep Learning の台頭により，一般に DAE が区分的線形変換法よりも高い性能を示しているが，DAE は過学習しやすいという問題を抱えている．そこで，本発表では DAE と区分的線形変換法の融合により未知雑音に頑健なニューラルネットワークベースの特徴量強調手法を提案する．提案手法では Deep Neural Network (DNN) をクリーン音声状態の事後確率推定にのみ用い，区分的線形変換によってクリーン音声特徴量を推定する．まず，クリーン音声特徴量空間を GMM でモデル化し，クリーン音声状態ラベルを得る．クリーン音声とノイジー音声は時間対応の取れているパラレルデータであるため，このクリーン音声状態ラベルを DNN により観測ノイジー音声特徴量より推定を行い，クリーン音声状態に対する事後確率を得る．その後，観測ノイジー音声特徴量から重み付きの線形変換によりクリーン音声特徴量を推定する．Aurora 2 を用いた連続数字読み上げ認識実験により，本提案手法は従来手法と比較し最も良い性能を得ることができた．

Vol.2014-SLP-100 No.4
2014/1/31

音楽における個人性の信号処理的モデル化

川渕 将太

名古屋大学武田研究室における音楽関連研究について紹介する．現在行っている研究テーマは，音響特徴量を用いた楽曲間主観的類似度の推定と，バネ質量系を用いた合唱における歌声の F0 動特性のモデル化である．

楽曲間主観的類似度とは，人間が感じる楽曲間の類似度のことである．本研究では，楽曲間の類似度を評価したデータを被験者実験により大量に収集し，それを用いて被験者の類似度評価を音響的特徴から予測する手法提案した．楽曲間の主観的類似度を，「楽曲間の音響的類似度」と「被験者がどのような特徴を重視するか」に分けて計算することにより，被験者の個人性を反映した類似度が推定できると考えた．被験者ごとに距離関数を最適化し，その距離関数を類以度計算に用いることによりこれを実現した．実験の結果から，楽曲間の主観的類似度においては，ボーカルの声質の類似度には個人差があり，距離関数を個人毎に変えることが主観的類似度推定に効果的であることなどが示された．

合唱では，複数の歌唱者が同時に歌唱を行うため，他者の歌声から影響を受け，独唱とは異なった歌唱となる．そこで，合唱における歌唱（音高）を特徴づける数学的モデルを構築している．まず，歌声の基本周波数 (F0) に着目し，歌唱における F0 の動特性 (ビブラートやオーバーシュートなど) をばね質量系によってモデル化した．これは，二次系のダイナミクスを分析する簡単なモデルである．この基本的なばね質量系を 1 つの質点と 2 つのばねで構成される結合システムに拡張し，目標音高と随伴歌唱からの影響をモデル化して合唱歌唱の分析を行う．このモデルのパラメータ推定を行うことにより，素人と合唱経験者の歌声の違いが表現できるなど，歌唱の特徴を表現できることを示した．

※共著者：川岸基成，宮島千代美，北岡教英，武田一哉

統計的手法に基づく歌声声質変換

小林 和弘

YouTube やニコニコ動画を代表とするデジタル音楽コンテンツの普及により，プロフェッショナル・アマチュアを問わず，自身の歌声をインターネット上に投稿・公開することが容易となった．歌手は，言語情報である歌詞に対し，メロディーやリズムを与えることで，多様な歌唱表現を生み出すことができる．また，歌手の技量に依るものの声帯や調音器官を巧みに操ることで声質を変化させる事が可能である． しかし，声質に関しては歌手の身体的制約が大きく反映されるため，個々の歌手が表現できる声質は限定される．我々は歌手の持つ声質を自在に変化させ，新たな独自性を追求できるボーカルエフェクターの実現に向けて研究を行っている．

身体的制約を超える歌唱を実現する手法として，統計的手法に基づく歌声声質変換（SVC: Singing Voice Conversion）が提案されている．SVC は，入力歌手の声質を所望の声質を持つ目標歌手へと変換する手法であり，学習処理と変換処理から構成される．学習処理では，入力歌手と目標歌手の同一楽曲の歌唱音声であるパラレルデータを用いて，両歌手の音響特徴量の結合確率密度関数を混合正規分布モデル（GMM: Gaussian Mixture Model）でモデル化する．変換処理では，学習された GMM に基づき，入力歌手の音響特徴量を目標歌手の音響特徴量へと変換する．この手法により，入力歌手の声質を目標歌手の声質へと変換することが可能である．しかし，GMM の学習には入力歌手と目標歌手のパラレルデータが必要であるため，変換可能な歌手はパラレルデータが入手可能な歌手に限定される．また，歌手間の変換モデルを学習しているため，制御可能な要素は歌声の個人性のみに留まっており，歌手が意のままに操作可能なボーカルエフェクターの実現には至っていない．

本発表ではより柔軟な声質制御を実現するべく提案された，任意の入力歌手から任意の目標歌手への声質変換が可能な多対多固有声 GMM に基づく SVC および声質を直感的に表す声質表現語の一つである知覚年齢に基づく声質制御を実現する多対多重回帰 GMM に基づく SVC について発表を行う．

Vol.2014-SLP-100 No.4
2014/1/31

危機的状況を瞬時伝達する警報音の合成を目指した音響パラメータ制御の検討

白鳥 大樹

内閣府によると，地震を代表とする自然災害が，世界各地で増加傾向にあると報告されている．災害による被害を最小限に留めるため，緊急地震速報や津波警報など，危機的状況を瞬時に伝達することが望まれており，警報音を適切に設計することが注目されている．警報音は，聴取して即，危機的状況であることを想起することが望ましく，警報音によっては状況が察知できずに避難が遅れる可能性もある．そこで，本研究では，危機的状況を瞬時に伝達する次世代の警報音の設計に取り組んでいる．ここでは，言語情報を含み様々な危機的状況が報知可能な警報音として叫び声に着目し，叫び声に含まれる危機感に相当する音響パラメータの強調によりこの課題の解決を図る．

叫び声を叫び声足らしめる特徴量については，先行研究から基本周波数 (F0)，第一フォルマントの増加，スペクトル傾斜の減少が報告されている．したがって，これらの音響パラメータを強調する変換を施すことで，より危機的状況を想起できる可能性があるといえる．ここでは，叫び声と平静音声を収録した叫び声データベースを解析し，先行研究と同様の傾向があるかについて確認する．また，本研究の目的は，叫び声に含まれる危機感に相当するパラメータの強調であることから，分析結果に基づいて叫び声加工を行う変換規則の構築にも取り組むため，さらにいくつかの音響パラメータについても解析を行うこととした．

本実験では，叫び声データベースを解析し，F0，第一フォルマント，第二フォルマント，スペクトル傾斜，1/3 オクターブバンド分析によるスペクトルの大局構造の違いについて解析した．F0 とスペクトル包絡を TANDEM-STRAIGHT で分析し，得られたスペクトル包絡を対象に，次数 8 での最尤スペクトル推定によりフォルマントの推定を行った．スペクトル傾斜については，対数スペクトル包絡を対象として最小二乗法により得られた直線の傾きとした．これらの結果から，F0，第一フォルマント，スペクトル包絡の大局構造により，叫び声の特徴を説明できることが明らかとなった．この結果は，変換規則の構築について，スペクトル傾斜のような単純なパラメータではなく，1/3 オクターブバンド分析結果程度の大局的構造を補償するフィルタの設計が重要であることを示唆する．

スペクトル補正及び統計的音源生成に基づくハイブリッド電気音声強調法

田中 宏

喉頭摘出者のための代用発声法の一つとして，電気式人工喉頭を用いた発声法がある． 外部から機械的に生成される音源信号を用いて発声を行う方法であり，習得が容易で，かつ，比較的聞き取りやすい音声（電気音声）を生成できるという利点がある．一方で，自然な音源信号を機械的に生成するのは困難であり，特に発話内容に応じた自然な基本周波数パターンを生成するのは本質的に極めて困難な処理となる． 結果として，電気音声の自然性は大きく劣化する．また，電気式人工喉頭から生成される音源信号自体が外部に漏れるため，雑音として電気音声に混入し，その品質を劣化させる．これらの問題に対処するため，本稿では，電気音声の聞き取りやすさを保持しながら自然性を大幅に改善する音声強調法として，従来の強調法である雑音抑圧に基づくスペクトル補正処理と統計的声質変換に基づく音源特徴量生成処理を組み合わせたハイブリッド法を提案する．提案法では，統計的手法によるスペクトルおよび有声無声情報への変換処理を回避することで，変換誤差により聞き取りやすさが劣化する事態を回避する．また，統計的手法により，通常音声の基本周波数パターンを予測することで，自然性を大幅に改善する．実験的評価結果から，本手法の有効性を示す．

Vol.2014-SLP-100 No.4
2014/1/31

自閉症スペクトラム児と定型発達児のナレーティブ発話分析

田中 宏季

自閉症スペクトラム障害とは，社会性とコミュニケーションに困難がある発達障害であり，言語や非言語の理解・表出に影響を及ぼすと報告されている．特にこれまで，自閉症者では感情，社会性，知覚について言及する発話の表出が少ないことが報告されており，また単調な韻律となることもレオ・カナーによる 1940 年代の自閉症スペクトラム研究報告に見られる．しかしこれらを定量的および包括的に調査した報告はこれまで存在しない．これらを定量的に測定することは，スクリーニングツールの開発，あるいは自閉症を理解するための助けとなる．本研究は，知能指数と年齢のマッチした自閉症スペクトラム児 4 名，定型発達児 2 名による予備実験として行われた．我々は，Linguistic Inquiry and Word Count (LIWC) 辞書を用いて言語情報，韻律情報を抽出し，ナレーティブ発話からの自閉症スペクトラム児と定型発達児の特性の違いを分析した．本研究でのナレーティブとは，主人公が発話者自身である，個人的に印象に残った体験についての説明である．まず t 検定を用いた特徴量ごとの優位差を検証し，両グループで，社会性，知覚，認知に関する発話頻度，フィラーの使用頻度，基本周波数 F0 の標準偏差と Coefficient of variation に優位差がある事を確認した．さらに主成分分析，因子分析，決定木などにより重要な特徴量について分析を行った．最終的には，スクリーニングへの応用を目指した識別を行い，SVM と Naive Bayes を用いた交差検定で 66 %の正解率を得た．Leave-one-speaker-out の結果でも，同程度の正解率を得る事ができたが，個人差が見られることも確認された．今後はスクリーニングツールの開発に向け，個人差を考慮したカットオフ値，他の言語・非言語特徴量，ナレーティブ以外のタスク検討などを進めていく．またインタラクションの観点から，対話中の発話応答時間や応答内容についても調査する必要がある．

Using Phonetic Context for Continuous Speech Recognition with Invariant Structure

張 聡穎

Continuous speech recognition is facing the problem that the recognition accuracy is still not satisfying. It could possibly be solved by applying more effective modelling approaches or new features. One new feature proposed in recent years is the invariant structure. An invariant structure is one of the long-span acoustic representations, where acoustic variations caused by non-linguistic factors are effectively removed from speech. It consists of all f-divergences between each acoustic events pairs. Here one f-divergence is named as an edge. And the model trained for all edges is named as the statistical edge model (SEM). In previous research, the invariant structures are leveraged as features of discriminative reranking for the hypotheses from the automatic speech recognition (ASR) system. First, the Gaussian Mixture Models (GMM) are built for all edges between monophone model pairs. Then the discriminative SEM is built on the log possibilities of all edges. After the discriminative model is built, for each hypothesis, the invariant structure score is calculated. The new score will be the score of combining the invariant structure score and the ASR score. And the hypotheses are reranked according to the new score. However, in previous research, the phonetic context is not considered, and representing the validation for invariant structure by the log possibilities of edges are still need to be discussed. In this research, first, the phonetic context is considered. The f-divergences are calculated between each triphone pairs in order to represent the phonetic context. Second, the discriminative SEM is trained by only considering the validity of appearance of edges. The reranking step remains the same as previous research. The proposed approach is tested in continuous digits speech recognition task and large vocabulary continuous speech recognition task and both results showed recognition accuracy improvment.

日本人英語音声を対象とした単語了解度の自動予測

ポンキッティパン ティーラポン

本研究では，日本人が英文を読み上げた場合に日本語訛りによって聞き取り難くなってしまう単語を自動的に予測する手法を検討する．我々の先行研究 [1] では，日本人による 800 の読み上げ文音声を 173 名の母語話者に呈示して書き起こさせ，発声中の単語毎に聞き取り率を求めている．本研究ではこの実験結果を用いて「日本語訛りによって聞き取り難くなる単語発声」を定義し，その単語発声を自動的に予測することを考える．意図された文とその読み上げ音声から，言語的素性，語彙的素性のみを使って CART (Classification And Regression Tree) による予測を試みた．次に，英語と日本語の音韻体系の違い，音素配列の違いを考慮して新たな素性を導入し，更には，入力音声と当該文の母語話者発声に対する IPA 書き起こしに基づく素性も導入した．言語的素性及び語彙素性のみを用いた手法に対し，新しく導入した二素性は予測率の向上に大きく貢献することが分った．最終的に，提案手法は「非常に聞き取り難くなる単語」と「やや聞き取り難くなる単語」を，F1 スコア 69.59%及び，78.36%で予測可能であることが分った．

クラウドソーシングを用いたインタラクティブな音声対話システムのための大規模主観評価プラットフォームの構築

中島 陽祐

音声認識技術の進展により，音声対話の分野においても単純なタスクならばタスク指向型対話システムはかなりの精度で達成できるようになった．それにともない，雑談対話システムではさらに自然で多様なコミュニケーションが求められてきており，現在では Siri やしゃべってコンシェルといったより会話調のインタフェースも登場してきている．音声対話システムにおいて，システム効率・精度，応答の良し悪しは対話時間等による客観評価が行われてきた．一方，ユーザビリティやインタラクティブ性の客観評価は応答遅延等により行われてきた．主観評価はユーザアンケートや感性評価により行われるが，これにはマシンスペックやエージェントの身体性といったインタラクション環境が評価に大きく影響するため，画一的な評価は非常に難しいと言われている．また，これをラボラトリ環境で大規模に行うには，実験者がインタラクション環境を統一または把握するのに多大な人的・時間的コストを要し，さらに評価方法が確立されておらず，実験者がタスクやシステムに合わせて様々に設定するため，十分に信頼性のある評価結果を得ることは困難である．一方，音声対話システム評価において，クラウドソーシングを用いて不特定多数のユーザからより信頼性のある主観評価を得ようという試みがある．クラウドソーシングを用いた主観評価実験は Blizzard Challege を始め，音声合成の分野では既に行われており，音声対話の分野では録音された対話の良し悪しをユーザが評価する主観実験も行われている．そこで本研究では，クラウドソーシングを用いてインタラクティブな音声対話システムの大規模評価を行う枠組みについて提案する．これに MMDAgent を用いることで，身体性を伴った細やかなインタラクションまで可搬性高く表現できる．この枠組みでは，クラウドソーシングプラットフォームを通してリクルートされた被験者が実験セットを自分のマシンへダウンロードし，実行ファイルを起動することにより自動で実験が進行する仕組みとなっている．この際，MMDAgent が収集したインタラクションログ，フレームレート，SN 比，認識遅延，応答遅延，またユーザが Web アンケートで入力するモニタサイズ，解像度，ユーザプロパティ，周辺情報，音声入出力デバイス情報を収集できる．この時生じるバッドデータはインタラクションログ，フレームレート，SN 比等から判断し，棄却する．また，エージェントの身体性をスクリーンサイズから逆算して後段の評価に役立てる．このような仕組みを用意することによって，様々なマシンで実験を行うことが可能となり，大規模な評価収集が可能となる．本発表ではシステムおよび評価結果について報告する．

最近の研究内容について

西田 昌史

本発表では，最近取り組んでいる話者認識，音声認識，マルチモーダル会話分析に関する研究内容について紹介します．まず，話者認識では会議や討論などの多人数会話を想定した話者分類について取り組んでおり，従来の階層的なクラスタリングに比べて高速で高精度な非負値行列因子分解に基づく手法を提案し有効性を示しました．音声認識では，単語の重要度を考慮したベイズリスク最小化音声認識を音声クエリーによる音声ドキュメント検索に適用し，従来の尤度最大化音声認識に比べてクエリーならびに検索対象のいずれに対しても認識精度の改善が得られ，検索精度の改善も得られました．また，日本人の英語発話を対象として文法的な誤りを指摘する対話型 CALL システムを構築しています．本システムでは，日本人が英語を発話する際に誤りやすいと考えられる音素を決定木によりクラスタリングすることで，認識精度の改善を図っています．マルチモーダル会話分析では，多人数会話において言語が異なる際に参加者の振る舞いがどのように影響を受けるかについて分析を行っています．日本人の 3 名が同一テーマで日本語と英語で会話を行い，ビデオカメラ，マイク，視線追跡装置を用いて会話を収録しています．これらの会話から日本語と英語の言語の違いや話題の内容に応じて話し手と聞き手の視点で，視線や体の動きがどのように変化しているか，あるいは音声や非言語情報を用いた話者交替の予測などについても取り組んでいます．

楽しい音声対話システムを作りたい！

西村 良太

私は，音声対話システムを専門に研究をしており，音声対話システムを使いやすくするために，また，楽しく音声対話をするために，どのようにすれば良いかについて，考えている．これまでに行ってきた研究としては，大きく 3 つに分けられる．

一つ目は，「対話のリズム」に着目し，人間同士の対話を分析し，それに基づいたモデルを構築し，音声対話システムへの組み込みを行った．具体的には，応答タイミングや，韻律的な同調と対話の盛り上がりとの関連を分析し，このモデルを音声対話システムに組み込んだ．このシステムは，リアルタイムに応答タイミングを検出し，種々の雑談現象を扱い応答する．そして，応答を出力する際の韻律情報を，ユーザに同調していくように制御する．このシステムにて被験者実験を行ったところ，システムへの評価が有意に向上した．このシステムを発展させ，一人のユーザと 2 つのシステムの三者対話を行うシステムを，現在，豊橋技科大中川研の学生と共同で研究・開発中である．

二つ目の研究は，データベースの検索に音声対話システムを用いる研究を行った．具体的には，レストラン検索システムである．音声対話システムでは，音声認識誤りは避けられず，ユーザの意図と異なるシステム挙動になる．そこで，音声認識誤りの誤受理防止，また対話状態に応じたシステム発話生成を行うため，対話状態の推定を行う．このために，対話データの収集を行い，機械学習によるモデル構築を行った．このモデルを対話システムに組み込んだ結果，対話の状況推定結果を用いて，システムの誤りを修正して対話を行うことができた．

三つ目の最近主に取り組んでいる研究は，音声対話システムの対話シナリオを構築する環境の開発を行った．研究全体としては，音声対話システムの一般普及を目指しており，その中で，対話コンテンツやシナリオを簡単に作れる環境づくりを行った．具体的には，web ブラウザで動作する対話シナリオエディタを構築した．これにより，OS などのシステム環境に依存しない，見やすく編集しやすいシナリオエディタができ，音声対話システムの一般普及へと近づいた．この研究は現在進行中であり，更に改善していくものである．

発表内容は，これまでの研究の成果として，音声対話システムの一例を示したい．

日本語アクセントに基づく基本周波数パターンの区分線形回帰と HMM 音声合成への適応

橋本 浩弥

HMM 音声合成は，TTS (Text-to-Speech) システムの 1 種であり，波形接続方式と比較して柔軟な音声合成を可能にするため，近年注目されている．しかし，HMM 音声合成は音声分析合成技術に基づいて音声を特徴量で表現するが，フレーム単位で扱うため，より長時間にまたがってあらわれる韻律的特徴のモデル化が困難であるという問題がある．韻律を担う重要な特徴量である基本周波数（F0）の時系列パターンを表現するために様々なアプローチが試みられてきた．1 つに基本周波数パターン生成過程モデルのような物理的・生理的に基づくモデルがある．しかし，観測される F0 パターンは，無声区間や microprosy 等に起因して，モデルパラメータの抽出が極めて困難なケースがある．2 つ目に，HMM において，フレーム単位，音素単位，シラブル単位，単語単位などの各階層ごとに分離して足し合わせる手法がある．しかし，各階層が持つ F0 の物理的な意味を解釈することが難しい．3 つ目に，波形接続方式における韻律の生成を目的として，母音の重心点の F0 を設定し，その間を直線補間で接続することによってその F0 パターンを表現する点ピッチモデルが提案されている．本研究で提案する F0 パターンの区分線形回帰は，この手法を土台とする．提案手法では日本語のアクセント型に基づき，区分回帰する．そして母音だけでなく，有声子音の F0 も考慮し，境界位置を初期値から更新しながら最適化する．これにより，聴覚上重要なアクセントの構造を捉えた F0 パターンを少数のパラメータで表現することができる．この回帰による概形 F0 パターンを学習用音声から抽出した元の F0 パターンの代わりに HMM の学習に用いる．さらに，F0 パターンと概形 F0 パターンの差分を用いて，別途 HMM を学習して合成時に足し合わせる．合成時の継続長は自然音声のものを用いた．提案手法の有効性を確認するため，学習時に抽出された F0 パターンをそのまま用いる場合を従来手法として合成音声の比較実験を行った．ATR 日本語音声データベースから話者 MHT を選び，全 503 文のうち，サブセット A から I までの 450 文で HMM を学習し，サブセット J の 53 文を合成した．自然音声に対する合成音声の対数 F0 平均二乗誤差は，従来手法は 0.18，提案手法は 0.15 となり，従来手法に比べて誤差が減少した．今後の課題として，回帰パラメータを直接取り扱うような統合モデルの実現を目指す．

地理情報を活用したモバイル音声対話システムに関する研究

原　直

本研究では，現在位置の地理情報を活用して情報を検索するための音声対話システムに関する研究を行っている．スマートフォンなどの携帯端末で音声対話システムを利用する場面を想定すると，利用者はその音声対話システムとの対話だけを通して場所に適した情報を求めていると考えられる．しかし，現在の携帯端末向けの音声対話システムは詳細な情報を提供する場合には WWW での検索結果を返すことが多い．これは開発者の持つ「情報」に限りがあるためである．音声対話システムの開発の敷居が下がって，インターネット上の WWW ページを作るのと同程度の知識で開発可能になれば，例えば店舗や施設の管理者による音声対話システム提供が可能となる．WWW ページに代わる音声対話システムとその環境が構築されれば，現在の WWW 検索に頼った音声対話システムとは一線を画したシステムになることが期待される．本研究では個人運用の WWW ページと同様の手軽さで音声対話エージェントを構築するための手法と，それらのエージェントを複数連携しながら携帯端末利用者に有益な情報を提供するための手法を検討している．基盤システムとしては音声対話システム「たけまるくん」を利用している．「たけまるくん」は奈良先端科学技術大学院大学で開発されたシステムで，「音声対話」機能と「音声検索」機能を兼ね備えており，利用者からの声に従って自動判別と適切な応答を返すように設計されている．「たけまるくん」の大きな特徴は「一問一答対話方式」を採用していることである．この対話方式は，開発者があらかじめ質問されそうな文章とそれに対応する応答文を大量に登録するだけで良いため，音声に関する専門知識をあまり必要としていない．そこで「たけまるくん」を基盤として，まずは「たけまるくん」と同等の音声対話システムをネットワーク上で運用するためのプロトタイプシステムを実装した．現在はこのプロトタイプシステムを基盤として，多数の対話システム構築が容易に行える環境を整備している．

Vol.2014-SLP-100 No.4
2014/1/31

Clustering i-Vectors for Training PLDA Models in Speaker Verification

Sangeeta Biswas

Recently, systems combining i-vector and probabilistic linear discriminant analysis (PLDA) have become the state-of-the-art method in speaker verification. An i-vector system maps utterances into a low dimensional space, known as the total variability space (TVS). The coordinate vectors in the TVS are known as i-vectors. Each i-vector contains most of the information related to speaker identity, as well as irrelevant factors such as the transmission channels or the speaker's emotion. The PLDA model separates speaker factors from irrelevant factors. In order to train a good PLDA model, two conditions need to be fulfilled. First, training data should be plentiful. Second, the training data should be suitable; the training data should have similar properties as the evaluation data. There is a trade-off between these two conditions. Using gender-dependent clusters is one good compromise for this trade-off. Obviously, speakers' acoustic properties depend not only on gender but also on the physical characteristics of the vocal tract, dialect, age etc. In addition, channel factors such as transmission type or background noise are known to greatly affect the acoustic properties of a recording. It seems therefore natural to group the training data based on more factors than gender. Therefore, we go beyond gender-dependent clusters in PLDA-based speaker verification. Since we do not know what factors are important to consider, we adopt an unsupervised approach. We propose to cluster i-vectors used for PLDA training by an agglomerative hierarchical clustering (AHC) algorithm. We also compare some popular linkage methods and distance metrics. Our proposed method obtained significant performance improvements on the male trials of the core condition of the NIST 2006 SRE and 2008 SRE (tel-tel) dataset.

言語モデリングにおける学習データの課題を解決するための 2 つのアプローチ

増村 亮

音声認識のための言語モデルとして，n-gram 言語モデルは未だに最も基本的かつ不可欠な役割を担っている．言語予測において，直近の単語履歴の重要性は疑う余地がなく，n-gram 言語モデルは今後も言語予測の中心的な役割を担い続けると考えられ，その高度化は欠かせない．実際に，Recurrent Neural Network 言語モデルや Model M 等が近年注目されているが，いずれも n-gram 言語モデルと併用されることが一般的である．

n-gram 言語モデルの高度化の鍵は，やはり学習データの課題であろう．n-gram 言語モデルは膨大なパラメータを持つため，頑健な学習には膨大な学習データが必要となるが，そのパラメータ推定は学習データに強く依存するため，想定するドメインに適した学習データの存在が重要となる．しかしながら，想定するドメインに適した学習データを十分に準備することは容易ではない．そこで本発表では，この学習データの課題を解決するための 2 つのアプローチを紹介する．

まず 1 つ目は，外部言語資源から学習データを集めるアプローチである．言語資源であれば，Web 上の豊富な言語資源を活かす手段が考えられる．そこで，話し言葉において代表的なコーパスである CSJ に注目し，Web 上の言語資源の有用性について検討を行った．具体的には，CSJ をシードデータにして，Web 上から話し言葉のテキストのみを選択し，適宜整形して利用する枠組みを提案した．本アプローチにより，Web データのみから CSJ と同等の性能の n-gram 言語モデルを構築できることを確認している．

次に 2 つ目は，学習データ自体を自動で生成するアプローチである．具体的には，データ生成のためのモデルを最初に構築し，生成したデータから，n-gram 言語モデルを学習する枠組みである．このデータ生成のためのモデルとして，柔軟なモデル構造を持つ Latent Words Langauge Model を利用し，その確率過程に基づくデータ生成の枠組みを提案した．本アプローチにより，限られた学習データから性能の高い n-gram 言語モデルを構築でき，同時に，学習データのドメインと異なるドメインでも頑健に動作することを確認している．

多人数会話ファシリテーションロボット

松山 洋一

多人数で構成されるグループに参加し，ファシリテーションできる会話ロボットシステムを提案する．本発表では特に，(1) 4 者会話グループを調和させるファシリテーション戦略と，(2) 意外性のある意見文自動生成手法について紹介する．

(1) 4 者会話グループを調和させるファシリテーション戦略

3 者が参加する会話で発生する「発話機会の不均衡」を解消するために，ファシリテーターとしてのロボット（第 4 番目の参加者）は「置いてけぼり」状態になっている参加者を検出し，その人に発話機会を提供するために，主導的に会話を進めている他の参加者の状態も見ながら適切に場をコントロールする手続きを発動させる．この手続きのルールは，部分観測マルコフ決定過程（POMDP）によってモデル化する．

(2) 意外性のある意見文自動生成手法

「自己目的的に楽しまれる会話」を実現するために，ロボットの発話コンテンツの自動生成について検討する．自己目的的に楽しまれる会話における「魅力的な対話相手」というのは，会話を展開するために，聞かれたことに単に応答するだけではなくて，それに関連した有用な情報や自分の意見などを付け加えて発話してくれるものだろうという仮説のもと，a）客観的事実に関する発話（Wikipedia などの情報に由来）と，b）意見・感想発話を，文脈や状況に合わせて組み合わせる仕組みを提案する．後者の意見・感想発話は，特定の対象に関する Web 上の不特定多数のレビュー文から意見文を抽出し，「発話文の長さ」，「文脈との整合性」，「意外性」などの基準からランキングされ出力される．

これらの対話システムは，会話ロボットプラットフォーム SCHEMA（シェーマ）上で実装されている．SCHEMA は，会話ロボットとしての親和的な外見も加味しながら，座位の会話相手の目線に合わせて身長はおよそ 120cm とし，人間との会話のプロトコル（物理層）を合わせるための必要条件の検討の結果，合計 22 自由度を有する．

Deep neural network による音声認識に適した特徴量抽出の検討

森勢 将雅

本研究では，音声認識の識別器に Deep neural network (DNN) を用いることを前提とした特徴量抽出について検討している．従来の音声認識では音声のスペクトル情報を少ない次元数で表現するための特徴量抽出処理が必要不可欠であり，そのためのアイディアとして Mel-frequency cepstrum coefficients (MFCC) が利用されてきた．これまで筆者らは話者識別を対象とした特徴量抽出について検討し，DNN を識別器として用いた場合は対数パワースペクトルをそのまま入力することが最も高い性能を示すことを確認している．

本報告では，同様の話者識別実験をさらに様々な条件で系統的に実施することで，DNN を使うことを前提とした音声のパターン認識における特徴量のあり方の仮説を示す．実験には，日本語 5 母音かつ 1 オクターブの範囲で持続的な発話を行った男女各 2 名の歌声データベースを用いた．この条件の音声のうち半分を学習に利用し，もう半分をテストに利用する．前報ではパワースペクトルをそのまま用いることが最適であることを示したが，学習に必要な時間が膨大になることから，性能を落とさない範囲での次元削減として，MFCC の次元数を 8 次元〜102 次元までについて，コサイン変換前のメル周波数スペクトルから平均を取り除いた特徴量 (MFSP) とともに比較評価することとした．また，同様の条件での実験を異なる初期値を用いて 100 回繰り返すことで，初期値の違いによる統計的な分布についても評価した．本実験から確認された結果は以下の通りである．(1) エラー率は隠れ層の数が 2 で概ね収束する．(2) MFCC と MFSP との比較の場合，同じ次数でも MFCC の場合 0 次を利用しないことから 1 次元分高性能である．0 次を利用した MFCC では誤差は同様となる．

これらの結果から得られた特徴量抽出の仮説について述べる．入力層から隠れ層へは，各層のユニット数の積から構成される数の係数が存在する．この係数は入力ユニットの値に乗算されるため，この係数を適切に設定することで，特定の帯域の選択，平滑化や先鋭化，ダウンサンプリング，アップサンプリング，コサイン変換など様々な処理が可能になる．したがって，特徴量抽出における次元削減では，明らかに不要な帯域の除去とパワースペクトルの周波数分解能の設定が重要になる可能性が示唆される．

Vol.2014-SLP-100 No.4
2014/1/31

ユーザの焦点に適応的な音声によるニュース案内システム

吉野 幸一郎

ユーザの多様な要求に対し一問一答を行うような対話システムが，これまでに多く研究・開発されてきた．これに対し，ユーザの複雑で曖昧な情報要求に対して，対象ドメインの知識を利用しながら複数ターンにわたって対話を行うシステムが求められている．これは単純なキーワードベースの検索ではなく，観光地やレストラン，ニュースの内容などについてより詳細な情報の案内を行うようなものである．このようなアプリケーションは，対象とするドメインの知識を記述した文書の情報を抽出・検索することによって実現することができる．

こうしたシステムを実現するため，日々動的に更新される Web 上のニュース記事を対象として，音声によるニュースの案内を行うシステムを提案する．このシステムでは，従来から扱われていたシステムに対するユーザの要求に加えて，ユーザがどの情報に興味があるかという焦点情報に着目する．ここで言う焦点とは,「ユーザの興味状態に沿った情報案内を行う上で不可欠な対象」である．これにより，ユーザとの対話を通じて，曖昧で具体化されていない情報要求に応えることを目標とする．また，ニュースから情報を抽出・利用するために，述語項構造および述語項構造を用いて自動構築された述語項構造テンプレートを用いる．対話のための情報構造・テンプレートを自動で定義することにより，ニュース記事の様々なドメインに対してこの枠組みを適用することができる．

音声対話システムは，音声認識結果やユーザの意図理解結果，加えて焦点の解析結果などの誤りを想定しなければならない．そこで部分観測マルコフ決定過程（POMDP）を用いて，誤りに頑健でユーザの要求と焦点に適応的な統計的対話制御を行う．具体的には，ユーザの要求と焦点に対する信念状態を定義・更新し，ニュースの案内を行うために最適なモジュールの選択に利用する．

Error Correction Interface for Speech Recognition

Yuan Liang

In recent years, speech input interface has become popular in smart phone applications. In this interface, speech recognition errors are unavoidable. When high quality transcriptions are needed, users are required to verify and correct the transcriptions obtained by speech recognition. In most speech interfaces, when a user finds an error word in the recognition result, he/she first marks it and then either selects the correct word from a candidate list provided by the interface, or input the correct word by speech, handwriting, or virtual keyboard. The error correction process is time-consuming. Therefore, efficient error correction interfaces have been strongly demanded.

The goal of our research topic is to realize efficient error correction for speech recognition. The main problem need to be solved is how to use the information generated in the human-machine interaction process to reduce the users' effort. Rodriguez [1] proposed a computer assisted transcription of speech approach, in which every time the user corrects a word, this correction is immediately taken into account to re-evaluate the transcription of words following it. They proposed a nature assumption: when user corrects an error word, all the previous words and this new corrected word are correct or already be corrected, they called this information as user validated prefix. Another research [2] used user validated prefix, higher-order N-gram language model (LM), and caching LM to reorder the confusion network. Its results are very promising. There may be many ways to use user validate prefix in the error correction procedure, and how to use the information generated in the human-machine interaction process is still need to be studied. So studies in these directions are also promising.

References

[1] Luis Rodriguez, et al., "Computer Assisted Transcription of Speech", IbPRIA, 2007
[2] Antoine Laurent, et al., "Computer-assisted transcription of speech based on confusion network reordering", ICASSP, 2011

以下の 4 件も当日発表予定である.

音声言語処理技術を用いた講義・講演の字幕付与

秋田 祐哉

AR キャラクタとの英会話練習時における交替潜時のタイムプレッシャーによる制御

鈴木 直人

マルチモーダル情報を使った音声対話システムのユーザ状態推定

千葉 祐弥

省リソースな計算機のための音声認識における演算量の削減

長野 雄