

【サーベイ】

ICSLPにおける研究動向－音響モデル・音声合成を中心に－

全 炳河[†] 南角 吉彦[†] 戸田 智基^{††}

† 名古屋工業大学 情報工学専攻 〒466-8555 愛知県名古屋市昭和区御器所町

†† 奈良先端科学技術大学院大学 情報科学研究科 〒630-0192 奈良県生駒市高山町 8916-5

E-mail: †{zen,nankaku}@ics.nitech.ac.jp, ††tomoki@is.naist.jp

あらまし 本稿では、2006年9月17日から21日にアメリカ合衆国ペンシルバニア州ピッツバーグで開催された、ISCAのICSLP2006(Interspeech2006)について、特に音響モデルと音声合成に関する報告を行う。

キーワード 音声情報処理、音声認識、音声合成、音響モデル

ICSLP 2006 Summary – Acoustic Modeling and Speech Synthesis –

Heiga ZEN[†], Yoshihiko NANKAKU[†], and Tomoki TODA^{††}

† Department of Computer Science and Engineering, Nagoya Institute of Technology
Gokiso-cho, Showa-ku, Nagoya-shi, Aichi, 466-8555 Japan

†† Graduate School of Information Science, Nara Institute of Science and Technology
8916-5, Takayama-cho, Ikoma-shi, Nara, 630-0192 Japan

E-mail: †{zen,nankaku}@ics.nitech.ac.jp, ††tomoki@is.naist.jp

Abstract This paper summarizes the acoustic modeling and speech synthesis-related topics in ISCA ICSLP2006 (Interspeech2006) held at Pittsburgh, PA, USA, from Sept. 17 to Sept. 21, 2006.

Key words Speech processing, Speech recognition, Speech synthesis, Acoustic modeling

1. 音響モデル

音響モデル関連では、Acoustic Modeling I-V, Discriminative Training, Speaker Characterization and Recognition I-IV, Robust ASR, Robustness and Adaptation for ASRと題したセッションがあり、計106本の論文発表が行われた。本節では、これらの発表の中から幾つかトピックを抽出し、関連する発表を取り上げその概要を述べる。

1.1 話者照合

話者照合システムにおいて、学習・評価時の雑音環境の違いは照合性能に大きな影響を与える。ヤコビ適応は、学習・評価間の雑音環境の差異に適応する有効な手法の一つであるが、学習・評価時の雑音音声が必要という問題がある。これまでには、文頭の無音区間より推定するなどして得られた雑音を用いて適応していたが、実環境では文頭に雑音推定に十分な無音区間があるとは限らない。そこでAnguitaら(Technical Univ. of Catalonia)は、発話区間検出と再帰的雑音推定法を組み合わせて学習・認識時の雑音を推定し、これを用いてヤコビ適応を行った(pp. 925-928)。話者照合実験において評価した結果、文頭の無音区間のみを用いて雑音推定した場合と比較して、様々なSN比において改善が得られた。

話者照合手法の一つに、cohortモデルを用いたものがある。これは、話者照合のスコアとして広く用いられる対数尤度比を計算する際、背景モデル(UBM)の代わりに評価データ周辺に分布する話者モデルを用いることで、詐称者の音響スコアを近似する方法である。cohortモデルでは、評価データと話者モデル(cohortモデルの候補)間の類似度を定義する必要がある。Liuら(Univ. of Illinois)は、UBM-MAPを発話変換とみなし、変換された固定長の発話データ間の類似度の計算・cohortモデル候補の選択を繰り返すことで、各評価データに対して適切なcohortセットを動的に選択する手法を提案した(pp. 929-932)。NIST話者照合タスクで評価した結果、従来のcohortモデルと比較して性能が改善した。

1.2 話者・環境適応

話者適応の代表的手法である最尤線形回帰(MLLR)や固有声(Eigenvoice)は環境適応にも広く利用されている。Tsaoら(Georgia Tech.)は、EigenvoiceとEigenspace-based MLLRを雑音環境適応に適用し、Aurora 2タスクで評価した(pp. 785-788)。適応データが比較的小量(2-10発話)の場合について、様々なSN比においてMLLRを用いた雑音環境適応と比較したところ、本手法では教師あり・教師なしいずれの場合においても優位な改善が得られた。また、EigenvoiceとEigenspace-based MLLR

間には顕著な差は見られなかった。

MLLR を用いて話者・環境適応する際、回帰木を用いてガウス分布をクラスタリングし、クラス毎に変換行列を共有することがよく行われる。Mandal ら (Washington Univ.) は、話者を幾つかのクラスに分類し、各クラスにおいて異なる回帰木を構築する手法を提案した (pp. 1133-1136)。まず、不特定話者モデルから学習話者へ MLLR を施し、変換行列を推定する。この変換行列に話者性が含まれると考え、これをベクトル化した後 k-means 法でクラスタリングし、話者を複数のクラスに分類する。適応時は、各クラスの回帰木を用いて推定された変換行列を重み付き線形結合して用いる。このときの重みは、適応データに対する尤度を最大にするよう推定される。CTS タスク (大語彙連続音声認識)において、話者非依存の回帰木を用いて MLLR 適応した場合と比較したところ、顕著な改善は見られなかった。しかしながら、話者非依存の回帰木を用いた場合単語誤り率が悪化した話者が約 15% 存在したが、提案法を用いて適応したところ、悪化した話者のうちの 70% について認識性能の向上が見られた。

1.3 全共分散モデル

近年、数千時間規模の音声コーパスが整備され、全共分散ガウス分布を隠れマルコフモデル (HMM) の出力確率分布に利用することが可能となりつつある。しかしながら、現在音声認識において広く利用されている技術には、対角共分散を仮定して導出されているものが幾つかある。MLLR はその代表例である。全共分散ガウス分布に対する MLLR はこれまでにいくつか提案されているが、計算量が多い・粗い近似を用いるなどの問題があった。Povey ら (IBM) は、全共分散ガウス分布のためのモデル空間・特徴量空間 MLLR を効率的に行うため、2 次項が対角共分散の場合と同一であると仮定し、勾配法を用いて変換行列を更新する手法を提案した (pp. 1145-1148)。EARS タスクにおいて評価したところ、これまでの手法と比較して若干の改善が得られた。また Povey は、全共分散を良く近似する手法の一つである SPAM [1] の基底とその重みを、効率良く繰り返し更新する手法を提案した (pp. 1159-1162)。SPAM の基底の更新は計算量が多いため、これまでには初期値のまま固定されることが多かったが、提案する手法では少ない計算量で基底を更新することが可能となった。TC-STAR タスク (大語彙連続音声認識) で評価したところ、総分布数が比較的少ない場合、基底の更新が有効であることが示された。また、SPAM 及び全共分散ガウス分布を用いたシステムの認識性能を比較したところ、SPAM は全共分散行列には若干及ばないものの、少ないパラメータ数で全共分散に近い性能が得られた。

1.4 ベイズ学習

これまで、ベイズ基準に基づくコンテキストクラスタリングが提案されているが、モデル構造に関する事前分布は一様分布を仮定していた。Hu ら (Univ. of Missouri) は、音素決定木に基づく状態共有において、モデル構造の事前分布を設定する手法 (BTIC) を提案した (pp. 1738-1741)。ここでは、モデルパラメータに関しては、ベイズ情報量基準 (BIC) と同様にラプラス近似による点推定を用いているものの、モデル構造の事前分布

を設定することによって、小量のデータで決定木を構築する際の精度の改善を図っている。本研究では、あらかじめ与えられた大量の不特定話者データに基づいて、質問の事前確率を決定し、小量の特定話者データでモデルの構造を推定している。質問の事前確率は、決定木内のノードの位置に依存すると考えられるが、あらかじめ全ての木構造と、その各ノードにおける事前確率を計算しておくことは不可能である。そこで、本研究では、事前確率を与えるための決定木と、少量のデータで構築するターゲットの決定木を同時に生成することによって、この問題を解決している。大語彙連続音声認識実験において評価した結果、尤度最大化 (ML) 基準を用いた場合と比較して、少ない分布数で高い認識率が得られた。

1.5 トポロジーの自動決定

HMM の状態数・混合分布数や状態遷移などといったモデルトポロジーを、対象とするタスクに応じて適切に設定することにより性能が改善することが知られている。Suk ら (AIST) は、逐次状態分割法 (SSS) では状態方向とコンテキスト方向に分割を繰り返すところを、コンテキスト方向でなく混合分布方向に分割を行う SSMS を提案した (pp. 629-632)。本手法を用いたコンテキスト依存モデルの構築は難しいが、組み込み機器向けの音響モデルにコンテキスト依存モデルを使うことは計算量・メモリの観点から現実的でないため、混合数・状態数が自動的に決定できる本手法は有効であろう。状態数や混合数が一定のモデルと比較して、少ない総分布数で同等の認識性能が得られることが報告された。

Liu ら (Illinoi Univ.) は、テキスト依存型話者照合のための話者モデル学習において、混合ガウス分布 (GMM) 型 UBM の各ガウス分布を HMM の状態とみなして状態遷移行列のみ学習し、これを初期モデルとして MAP 適応し話者モデルを学習する手法を提案した。数字単語を用いたテキスト依存話者認識に適用したところ、HMM 型 UBM から MAP 推定して学習した話者モデルと比較して、エンロールデータが少量のとき優れた照合性能が得られた。

1.6 セグメントモデル

HMM は音響モデルとして広く利用されているが、状態内で統計量が一定・状態出力確率の独立性を仮定・状態継続長のモデル化能力が不足、といった問題を有している。これらの問題を克服するため、セグメントモデルと総称される様々な統計モデルが提案されている。セグメントモデルは HMM より柔軟な表現能力を有するものの、認識時の探索空間が大幅に広くなる、という問題がある。このため、HMM を用いて複数の仮説候補を出力し、セグメントモデルを用いてこれを再評価する、という枠組みが用いられることが多い。しかしながら、この枠組みではセグメントモデルの性能を見極めることが困難である。この問題に対し、Li ら (Microsoft Research) は、セグメントモデルの一種である隠れトrajエクトリモデル [2] のためのフレーム同期ビーム探索を用いたデコーダを提案した (pp. 609-612)。TIMIT 連続音素認識実験において評価した結果、計算量は未だ HMM と比較して大きいものの、若干高い認識性能が得られた。Zhang ら (Edinburgh Univ.) は、こちらもセグメントモデルの一

種であるトراجエクトリ HMM [3] のためのフレーム同期ビーム探索に基づくデコーダを設計した (pp. 589–592)。TIMIT 連続音素認識実験で評価したが、HMM と比較して高い性能は得られなかった。

1.7 識別モデル

近年、サポートベクターマシン (SVM) や条件付確率場 (CRF) といった識別モデルを、音声認識に適用する様々な試みが行われている。以前は SVM を時系列や多クラスへ拡張する研究が盛んだったが、最近は CRF を用いたものが目立つ。

Keshet らは、HMM を用いないカーネルベースの音素認識手法を提案した (pp. 593–596)。本手法では、音響特微量系列、音素列及び音素開始時間を特微量として、正解音素系列とそれ以外の音素系列を分類する巨大なカーネルマシンを構築している。学習は、カーネルマシンより予測された音素系列と正解音素系列の Levenshtein 距離を最小にする基準で行なわれる。まず、カーネルマシンから音素系列を予測し、次に正解音素系列と予測された誤り音素系列のマージンが最大になるように学習を行なう。これを繰り返すことにいより、予測系列と正解系列の距離が縮まり、その間に音素列を正解・不正解に分類する超平面が作られる。また、本手法では適切なカーネル関数を定義することが重要となるが、ここでは、音響特微量の差位を表すカーネル、音素継続長の違いを表すカーネル、音素の遷移を表すカーネルの 3 種類を定義し、これらを組み合わせて用いている。認識実験においては、通常の HMM に基づく手法に比べ、改善は得られていないが、HMM を用いず、音声認識を巨大な 2 クラスの識別問題として取り扱っている点が非常に興味深い。

Longworth ら (Cambridge Univ.) は、音声認識において有効であった Augmented model を話者照合に適用した (pp. 1467–1470)。これは、確率場モデルにおける変数間の依存関係をシステムティックに決定する手法である。本手法では、UBM-MAP 適応した話者 GMM に Augmented model を適用し、話者照合のための話者モデルとして用いている。また、パラメータの推定法について、話者照合の問題を 2 クラス (受理と棄却) の分類問題とみなし、SVM と同様の手法でパラメータ推定を行なっている。NIST 話者照合タスクにおいて評価したところ、高い改善率が得られた。

1.8 識別学習

近年大語彙連続音声認識システムの音響モデルは、ML 基準ではなく相互情報量最大化 (MMI)・最小分類誤り (MCE)・単語誤り最小化 (MWE)・音素誤り最小化 (MPE) といった、識別的基本基準に基づき学習されることが多い。識別学習では通常、直接・間接的に学習データの誤り率を最小化するようモデルが学習される。識別学習は ML 学習と比較して高い認識性能を示すことが知られているが、学習データへ過適応するなどの問題があり、より良い汎化能力を持つ学習基準の模索が行われている。近年機械学習の分野で注目されるマージン最大化学習 (LME) はその一つであり、2 件の発表があった。

Yu ら (Microsoft Research) は、誤り最小化学習 (MCE) に LME のコンセプトを導入した (pp. 2421–2418)。通常の MCE では、シグモイド関数の中心は 0 に設定されており、識別境界付近

のサンプルはモデル学習に強い影響を及ぼすが、学習が進むにつれ境界から遠ざかったサンプルは、影響力が弱くなり、やがて学習は収束する。これに対し提案法では、シグモイド関数の中心を正解クラス側にシフトすることにより、各サンプルの識別境界からのマージンを大きくとるように学習を行なう。さらに、学習が進むに従って徐々にシフト量を増加させることで、よりマージンが大きくなることを期待している。Li ら (Georgia Tech.) は、HMM に基づく連続音声認識において、SVM のソフトマージンの考え方を取り入れたソフトマージン最大化学習 (SME) を提案した (pp. 2422–2425)。SME では、識別境界までの距離がマージンより小さいサンプルに注目し、学習を行なう。ここでは、HMM の学習にソフトマージンを適用するため、正解のラベル列と最も尤度が高かった誤りラベルを用意し、学習データの全フレームのうち正解ラベルと誤りラベルが異なっていたフレームのみを集め、その対数尤度の差を基準として用いる。さらに、この識別尺度にマージンを適用し、最終的な目的関数が構成される。この目的関数を用いて GPD によりモデルパラメータを更新する。Du ら (Microsoft Research Asia) は、識別学習における正解度を正解ラベル列に対応する HMM と認識仮説に対応する HMM の KL 距離に定義する手法を提案した (pp. 2410–2413)。MWE や MPE では、正解ラベルと認識ラベルのマッチ度に基づき正解度を定義しているが、これらは言語モデルや単語辞書・音素セットにより大きく影響を受ける。また、ラベル間の適合度によってのみ正解度が定義されており、音響的類似度が考慮されていない。そこで本手法では、正解ラベル列に対応する HMM と認識仮説に対応する HMM の KL 距離より正解度を計算し、MWE や MPE と同様にして Extended Baum-Welch 学習によりモデルを推定している。これら 3 つの手法 (マージン最大化 MCE, SME, KL 距離最小化) は、全て TIDIGIT タスク^(注1)において評価された。それぞれの単語 (文章) 誤り率 (%) は 0.19 (0.72), 0.22 (0.67), 0.49 (1.47) であった。いずれの場合においても、ML 学習と比較して改善が得られた。上記の SME とマージン最大化 MCE の誤り率は、TIDigits タスクにおいて、本稿執筆時において最も良いものだと思われる。今後大規模なタスクでどのような性能が得られるか、興味が引かれる。

1.9 SPLICE に基づく音声認識とその周辺

Deng ら (Microsoft Research) により提案された SPLICE [4] は、音声認識のための特徴抽出において注目されている。SPLICE は、ロバスト音声認識のための雑音補償法として考案されたものであり、雑音が含まれる音声 y を入力としてクリーン音声の推定値 \hat{x} を出力する。 x と y の関係は制約付き GMM を用いて以下のようにモデル化される。

$$p(x, y) = \sum_m p(x | y, m)p(y | m)p(m) \quad (1)$$

$$p(y | m) = \mathcal{N}(y | \mu_m, \Sigma_m) \quad (2)$$

$$p(x | y, m) = \mathcal{N}(x | A_m y + b_m, \Gamma_m) \quad (3)$$

(注1) : Yu ら及び Li らは男性・女性データのみ使用しているが、Du らは加えて男子・女子のデータを使用している。

但し, m は GMM 中のガウス分布のインデックスである。上式は, $p(\mathbf{y}) = \sum_m p(m)p(\mathbf{y} | m)$ を用いて音響空間を幾つかの領域に分割し, 各領域で異なる線形変換を行う, 区分線形変換とみなすことができる。この制約付き GMM は, クリーン音声 \mathbf{x} と雑音が含まれる音声 \mathbf{y} のステレオデータより, ML 基準で学習される。最小二乗誤差基準を用いると, 雜音が含まれる観測 \mathbf{y} に対するクリーン音声の推定値 $\hat{\mathbf{x}}$ は,

$$\hat{\mathbf{x}} = \sum_m p(m | \mathbf{y}) \cdot (\mathbf{A}_m \mathbf{y} + \mathbf{b}_m) \quad (4)$$

のように求まる。こうして得られた $\hat{\mathbf{x}}$ を直接認識器の入力に使用することも可能であるが, 雜音の不確定性を考慮したデコーディング手法に組み込むことも可能である [5]。今回の会議においても, 関連する発表が幾つか見られた。Liao ら (Cambridge Univ.) は, SPLICE 等 Front-End 領域における変換手法に共通する問題点を指摘し, その改善法を幾つか提案している (pp. 1121–1124)。その問題点とは, 変換用の GMM において低 SN 比に対応する混合要素が生成された場合, 変換前後の特徴量の相関が極端に低くなり, 変換後の特徴量の分布が雑音の分布のみに依存するため, 言語モデルなどの制約がない場合, 大量の挿入誤りが発生する可能性があるというものである。また, HMM の状態に雑音・クリーン音声の同時確率分布を対応させたモデルベース法では, 前述した問題が生じないことが述べられている。Aurora 2 タスクにおいて評価した結果, モデルベース法が高い性能を得られた。また Huo ら (Univ. of Hong Kong) は, 式(3)において $\forall \mathbf{A}_m = \mathbf{A}$ として, 変換行列とバイアスを ML 基準で学習する手法を導出した (pp. 1129–1132)。Droppa らによる SPLICE の実装では, $\forall_m \mathbf{A}_m = \mathbf{I}$ として, \mathbf{b}_m のみが推定されるが, 本手法ではグローバルな線形変換がさらに施される。 \mathbf{A} と \mathbf{b}_m は, 制約付き MLLR(CMLLR) と SPLICE のバイアス推定を繰り返すような形で最適化される。

一方, 識別的基準を用いて SPLICE の制約付き GMM を学習する手法も提案されている [6]。興味深いことに, 近年活発に研究が行われている識別的特徴抽出法の多くは, この枠組みで記述できる。Droppa らによる MMI-SPLICE では, $\forall_m \mathbf{A}_m = \mathbf{I}$ として, \mathbf{b}_m のみ MMI 基準で最適化しているが, Povey ら (IBM) によって提案された fMPE [7] は, $\forall_m \mathbf{A}_m = \mathbf{I}$ として, \mathbf{b}_m のみ MPE 基準で最適化しているものとみなせる。また, Zhang ら (BBN) によって提案された RDT [8] は, $\mathbf{b}_m = \mathbf{0}$ として \mathbf{A}_m のみ MPE 基準で最適化するものである。今回の会議においても, 識別的 SPLICE に関連した発表が見受けられた。Zhang ら (BBN) は, RDT と話者適応学習 (SAT) を組み合わせる手法を提案した (pp. 1495–1498)。通常 SAT の変換行列は ML 基準で推定されるため, 識別的に重要な情報が失われる可能性がある。そこで本手法では, まず RDT を適用し識別的な特徴量を抽出し, この特徴量を用いて各話者毎に SAT の変換行列を推定する。これにより, 識別的な情報をなるべく残した状態で適応学習のための変換行列が推定される。こうして話者正規化されたデータに対して再び RDT を施し, 最終的な特徴量が得られる。CTS タスク (大語彙連続音声認識) で評価したところ, RDT なしの

モデルや話者非依存 RDT と比較して, それぞれ 11% と 7% の誤り削減率が得られた。

(全炳河, 南角吉彦)

2. 音声合成

音声合成関連では, **Modeling Prosodic Features (pp. 297–320)**, **Corpus-Based Synthesis (pp. 1742–1765)**, **Speech Synthesis (pp. 2430–2453)** と題した 3 つのオーラルセッションと, **TTS I (pp. 1296–1355)**, **TTS II (pp. 2026–2077)**, **Voice Morphing (pp. 2250–2301)**, **Prosody (pp. 2358–2405)** と題した 4 つのポスターセッションにて, 計 71 本の論文発表が行われた。大規模音声コーパスの使用により高品質な読み上げ音声の合成が現実味を帯びた事から, 多様性に富む音声の合成が注目を集めている印象を受けた。特に, 統計的手法に基づく合成・変換に関する研究は大幅に増加しており, 音声変換に関しては 1 つのポスターセッションが行われるまでに至った。なお, 本稿では取り扱わないが, 2005 年に引き続き, 2006 年もテキスト音声合成 (Text-to-Speech: TTS) に関するコンペティション **Blizzard Challenge** がサテライトワークショップとして開催された。昨年を大きく上回る研究機関の参加により, 大盛況のうちに終わった。発表論文は <http://festvox.org/blizzard/blizzard2006.html> にて公開されているので, ゼビ参考にされたい。

以下では, TTS に関する論文を素片選択に基づく合成方式及び統計モデルに基づく合成方式といった 2 つの技術に大別して紹介する。次いで, 音声変換及び音声合成の応用に関する論文について紹介する。スペースの都合上, 紹介できない論文が多くあることをご容赦願いたい。

2.1 素片選択に基づく合成方式

素片選択に基づく TTS では, 言語情報から韻律ターゲットを予測し, それに基づき最適な素片系列を選択する方式が広く用いられる。しかし, 実際には同一文に対して複数の韻律パターンが許容される場合が数多く存在する。そこで, 韵律ターゲットと素片系列の両者を同時に探索する手法が提案された (pp. 1312–1315)。合理的な処理である一方で, 探索空間は膨大なものとなる。これまでにも複数の韻律ターゲットを考慮する枠組み [9, 10] は提案されているが, 任意のテキストを対象とした合成方式においては, 計算量の問題が未だ解決されていない。今後の研究成果に期待がかかる。

コーパスサイズが拡大し素片数が増えるにつれ, 考慮する素片接続の組み合わせは指数的に増加する。接続コストを 0/1 の 2 値に量子化することで, ビット演算により高速に素片系列の候補を絞り込む手法が提案された (pp. 2074–2077)。現状における物理量からの聴感上の接続ひずみ予測精度を考えると, ひずみが生じるか否かの判別のみに接続コストを用いるというアイデアは, 大胆ではあるが妥当といえる。ビット演算を活かした応用例が今後期待される。

知覚的に対応のとれた素片選択尺度 (コスト) を求める研究は古くから行われているが, 未だ十分な精度は得られていない。純粹に物理的側面のみに着目し, 韵律及びスペクトルパラ

メータをモデル化した HMM の尤度に基づき、フレーム単位で素片を選択する手法が提案され、極めて良好な結果が報告された (pp. 2034–2037)。単一の物理的尺度に基づく素片選択の実現可能性を示すものであり、極めて有意義な論文といえる。この他にも新規性の高い方法として、テキスト解析結果から得られる言語情報等に対して個々の素片が割り当てられる条件付確率を CRF で直接モデル化し、素片選択に利用する手法が提案された (pp. 2026–2029)。従来の枠組みで用いられているターゲットコスト (pp. 2038–2041) 及び接続コスト (pp. 1742–1746, pp. 1746–1749) に関する研究の進展も報告された。コストを出来る限り数学的に的確に記述した上で、極少量のパラメータを知覚的に最適化する枠組みに、今後大きな期待が寄せられる。

統計モデルに基づく合成方式が得意とする複数話者の音声データを有効利用するという枠組みが、素片選択に対しても導入され始めている。複数話者の音声素片を簡易な信号処理を施して使用する手法が提案された (pp. 2062–2065)。

素片選択手法及び素片接続手法に依存するものの、音素セグメンテーション精度は合成音声の品質に影響を与える。大規模音声コーパスに対する手動セグメンテーションは膨大な労力及び時間を要するため、高精度な自動セグメンテーション法の実現が望まれる。音響モデルを用いた強制アライメントに基づく自動セグメンテーションにおいては、音素毎にモデルの複雑さやトポロジーを最適化することで精度改善が得られる [11]。この知見に基づき、複数のセッティングで音響モデルを学習しておき、手動セグメンテーションデータに基づき、音素境界の種類に応じて最適なモデルを予測する回帰木を自動学習するという手法が提案され、その有効性が報告された (pp. 2066–2069)。音響モデルを用いない方式の一つである Latent semantic mapping (LSM) に基づく接続境界学習は、合成音声の自然性改善に有効である事が示された (pp. 1320–1323)。本手法では、まず、考慮すべき全接続境界周辺のピッチ波形に対して特異値分解を行うことで、ピッチ波形から低次空間への写像を定義する。接続コストは写像空間上でのコサイン距離に基づき定義され、最小の接続コストをもたらす接続境界がオフライン処理にて決定される。スペクトルのみでなく位相や基本周波数も同時に考慮する事ができ、かつ接続候補となる全素片を考慮して接続ひずみを定義できるという点で非常に興味深い手法である。論文では、ダイフォン接続のみを扱っているため同一音素素片のみを接続候補として扱えば良いが、音素間接続に適用する際には組み合わせ数が爆発的に増加するため、計算量を大幅に削減する必要があると予想される。

2.2 統計モデルに基づく合成方式

ここ数年、HMM に基づく音声合成方式 [12] の研究は急速な広がりを見せており、本方式は、木構造を用いて、コンテキストに依存した音声パラメータの確率密度分布をモデル化する。合成時には静的・動的特徴量の対応関係を考慮した上で、最尤基準により HMM から直接音声パラメータを生成する [13]。合成時と学習時の評価基準を統一化する試みとして、HMM パラメータの尤度基準に基づく学習法 [3] や最小誤差基準に基づく学習法 [14] が提案されている。この度さらなる進展として、木

構造を決定する際にも最小誤差基準を適用する手法が提案された (pp. 2046–2049)。膨大な計算量を削減するために多大な近似を必要とするものの、その有効性が確認された。

HMM 音声合成方式はモデル適応を行う事で、様々な話者の音声を合成する事ができる。制約付き SMAPLR により、適応精度はさらなる向上を見せた (pp. 2286–2289)。LSP の確率密度分布に対して MLLR を行う際には、帶行列を回帰行列として用いる事で、LSP 係数の disorder 問題を低減できる事が報告された (pp. 2250–2253)。また、450 文程度のデータを用いて特定話者モデルを学習するよりも、予め多数の話者のデータを用いて学習された平均声モデルを SMAPLR と MAP により適応する方が、より高い性能が得られるという結果が報告された (pp. 1328–1331)。これは、他の話者の音声データを話者正規化を施して有効利用する事で、特定話者データのみから学習されるモデルよりも、詳細なコンテキスト情報を用いた複雑なモデルを学習できるためである。

柔軟性に優れた音声合成の実現に向けて、HMM 音声合成方式に基づく合成音声の制御法について、積極的に研究が展開されている。操作性に優れる合成音声制御を実現する枠組みとして、知覚的に対応のとれた部分空間を最尤推定する方法が提案されており [15]、話者性制御 (pp. 2438–2441) や発話スタイル制御 (pp. 1324–1327) においてその有効性が報告された。

回帰木に代表される非線形モデルのみでなく、線形モデルにより音声パラメータを記述する枠組も研究されている。属性間の相互関係も考慮を入れた一般化線形モデルにより、基本周波数パターン (pp. 313–316) や音素時間長 (pp. 2374–2377) をモデル化する手法が報告された。ベイズ情報量基準 (BIC) 及び F 検定に基づき使用属性を選択する事で、完全自動学習が実現されている。

世界的に広く用いられている素片選択型テキスト音声合成プログラム Festival においても、統計的手法に基づく合成方式が新たに実装された (pp. 1762–1765)。また、素片選択型合成方式と統計的手法に基づく合成方式を結びつけようとする試みも見られた (pp. 1758–1761)。

2.3 音声変換

ここでは、入力された音声に対して言語情報を保持したまま話者性等の非言語情報を意図的に制御する音声変換技術について取り上げる。代表的な応用例である話者変換においては、言語情報を必要としない GMM に基づく変換法 [16] が主流である。

一般に GMM 学習時には、同一文発声の入力音声と出力音声のペアからなるパラレルデータが必要となるが、この学習時の制約を取り除く研究が行われている (例えば [17])。素片選択の枠組みに基づき、入力・出力フレーム間距離と隣接出力フレーム間距離を考慮して、入力フレームに対応する出力フレームを選択する事で擬似的にパラレルデータを構築する手法が提案されており、異なる言語間における声質変換に適用された (pp. 2262–2265)。また、入力話者からある出力話者への変換関数を予めパラレルデータを用いて学習しておき、目標出力話者の音声データのみを用いて変換関数のパラメータを MAP 推定する

手法も提案された(pp. 2254-2257)。本手法は確率的に問題解決に取り組む枠組みではあるものの、変換関数に対して大幅な近似が施されるといった問題点がある。これとは別に、音声認識における話者適応技術として知られる固有声を音声変換に導入した手法が提案された(pp. 2446-2449)。入力話者と多数の事前収録出力話者間のパラレルデータを用いて、固有声 GMM の事前学習が行われる。具体的には、音響空間内にて音韻性を維持したまま出力側の話者性のみを制御可能とする部分空間を構成する事で、少量パラメータによる出力平均ベクトル制御を実現する。2文程度の目標出力話者の音声データを用いて、部分空間上の少量のパラメータを最尤推定する事で、固有声 GMM の教師なし適応が可能となる。また、パラメータの手動制御により、変換音声の声質を自在に変化させる事もできる。ある入力話者から任意の出力話者への変換（一対多の変換）においてその高い有効性が示されている。

極少量の学習データから頑健に変換モデルを学習する手法として、混合因子分析を用いる手法が提案された(pp. 2278-2281)。GMM を用いない枠組みとしては、予め出力空間をクラスタリングしておき、変換時に得られる様々な音響特徴量からクラスターを選択するモデルを学習するという手法が提案された(pp. 2258-2261)。また、変換音声の品質を改善するために、音源モデルとして混合励振源を導入した変換法が提案され、その有効性が報告された(pp. 2266-2269)。

2.4 音声合成の応用

TTS の要素技術に関する研究は数多く存在するが、実際のアプリケーションに適用した際の評価は十分になされていない。その中で、音声翻訳における評価(pp. 2434-2437)や、対話システムにおける評価(pp. 2450-2453)についての論文発表が行われた。音声翻訳において、機械翻訳から得られる誤りを含む文を音声出力すると、聞き手は自身の言語的知識を反映せしる事で、異なる文として認識する傾向がある。その結果、自然な文と比較し機械翻訳出力文は合成音声の明瞭性が劣化するという結果が示された。対話システムにおいては、出力文が長くなるにつれて、意味が伝わりづらくなるという結果が報告された。この種の評価に関する研究は今後増えていくものと予想される。

Microsoft Research Asia から 2 つの興味深い研究発表が行われた。一つは人名の発音推定を目的とするもので、Web からの検索情報と文字 N-gram を併用する事で、人名から言語識別を行なう手法が提案された(pp. 1352-1355)。もう一つは、本を朗読する TTS の構築を目指すもので、ナレータが果たす個々の役割の音声を効率的に合成するための TTS データベースを、ナレータによる朗読音声データからの自動選択により構築する手法が提案された(pp. 1750-1753)。新しい事に挑戦する姿勢が感じられ、今後の研究展開が大いに注目される。

歌声という発話様式も興味深い研究ターゲットである。HMM 音声合成の枠組みにて、テンポのずれを明示的にモデル化する手法が提案された(pp. 2274-2277)。歌声音声データに対する音素セグメンテーション法として、スペクトル情報のみでなく基本周波数情報を考慮する手法が提案された(pp. 2294-2297)。また、歌声を音程の面から評価するアプリケーションに関する

発表も見られた(pp. 2298-2301)。

統計的音声変換技術の応用例としては、非可聴つぶやき(NAM) [18] をより自然な音声へと変換する事で、声を出さなくても電話ができる「無音声電話」の実現が期待される。NAM からささやき声への変換法が提案され、その有効性が報告された(pp. 2270-2273)。

(戸田智基)

文 献

- [1] K.C. Sim and M. Gales, "Precision matrix modeling for large vocabulary continuous speech recognition," Tech. Rep. CUED/F-INFENG/TR.485, Cambridge University, 2004.
- [2] J.L. Zhou, F. Seide, and L. Deng, "Coarticulation modeling by embedding a target-directed hidden trajectory model into HMM – model and training," Proc. of ICASSP, pp.744-747, 2003.
- [3] H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic features," Computer Speech & Language, vol.21, no.1, pp.153-173, 2007.
- [4] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," Proc. of ICSLP, pp.806-809, 2000.
- [5] J. Droppo, A. Acero, and L. Deng, "Uncertainty decoding with SPLICE for noise robust speech recognition," Proc. of ICASSP, pp.57-60, 2002.
- [6] J. Droppo and A. Acero, "Maximum mutual information SPLICE transform for seen and unseen conditions," Proc. of Interspeech (Europespeech), pp.989-992, 2005.
- [7] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltan, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," Proc. of ICASSP, pp.961-964, 2005.
- [8] B. Zhang, S. Matsoukas, and S. Schwartz, "Discriminatively trained region dependent transforms for speech recognition," Proc. of ICASSP, 2006.
- [9] I. Bulyko and M. Ostendorf, "Efficient integrated response generation from multiple targets using weighted finite state transducers," Computer Speech & Language, vol.16, no.3-4, pp.533-550, 2002.
- [10] T. Hirai, S. Tenpaku, and K. Shikano, "Speech unit selection based on target values driven by speech data in concatenative speech synthesis," Proc. of IEEE 2002 Workshop on Speech Synthesis, 2002.
- [11] H. Kawai and T. Toda, "An evaluation of automatic phone segmentation for concatenative speech synthesis," Proc. of ICASSP, pp.677-680, 2004.
- [12] K. Tokuda, H. Zen, and A. Black, "An HMM-based speech synthesis system applied to English," Proc. of IEEE Speech Synthesis Workshop, 2002. CD-ROM Proceeding.
- [13] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," Proc. of ICASSP, pp.1315-1318, 2000.
- [14] Y.J. Wu and R.H. Wang, "Minimum generation error training for HMM-based speech synthesis," Proc. of ICASSP, pp.89-92, 2006.
- [15] K. Miyanaga, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based speech synthesis," Proc. of Interspeech (ICSLP), 2004.
- [16] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," IEEE Trans. Speech and Audio Processing, vol.6, no.2, pp.131-142, 1998.
- [17] A. Mouchtaris, J. der Spiegel, and P. Mueller, "Non-parallel training for voice conversion by maximum likelihood constrained adaptation," Proc. of ICASSP, pp.1-4, 2004.
- [18] 中島淑貴, 柏岡秀紀, N. Campbell, 鹿野清宏, "非可聴つぶやき認識", 信学論(D-II), vol.J87-D-II, no.9, pp.1757-1764, 2004.