様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月　 ５ 日現在

機関番号：１２６０８
研究種目：基盤研究(B)
研究期間：2012～2014
課題番号：２４３０００７１

研究課題名（和文）ロバスト音声合成の深化と多言語音声コミュニケーションへの展開

研究課題名（英文）Research on advanced robust speech synthesis and its applications to multi-lingual speech communication

研究代表者

小林　隆夫（Kobayashi, Takao）

東京工業大学・総合理工学研究科（研究院）・教授

研究者番号：７０１５３６１６

交付決定額（研究期間全体）：（直接経費）　　11,000,000円

研究成果の概要（和文）：多様で表現豊かな音声合成の実現のために，モデル学習用音声データの量や質の変動に頑健で自然性の高い合成音声を生成するロバスト音声合成技術の深化をめざして研究を行った。ロバスト音声合成の基本技術として，学習用音声データのスタイル表出度合に依存しにくいスタイル制御モデルの構築法や韻律モデリング手法を提案し，評価実験を通してその有効性を示した。また，音声資源が乏しい言語へのロバスト音声合成技術の応用や新たなクロスリンガル音声合成手法を提案し，多言語音声コミュニケーションへの展開の検討を行った。

研究成果の概要（英文）：The purpose of the research is to develop advanced techniques that enable us to model acoustic features of prosodic information as well as spectral information with being less dependent on quality and quantity of training speech data for synthesizing natural-sounding and diverse expressive speech. We have proposed several robust techniques such as style control and prosody modeling ones and showed their effectiveness through objective and subjective evaluation tests. We have also applied the proposed techniques to under-resourced languages. Furthermore, we examined a cross-lingual speech synthesis technique for universal speech communication.

研究分野：音声情報処理

キーワード：テキスト音声合成　統計的パラメトリック音声合成　HMM音声合成　表現豊かな音声合成　韻律　クロスリンガル音声合成　音声スタイル制御

１．研究開始当初の背景

任意の文章を入力しそれに対応する音声を出力するテキスト音声合成（以下，単に音声合成と記す）技術は，音声認識・自然言語処理技術と共に，インタラクティブロボット，ナビゲーションシステム，音声検索，電子ブック自動読み上げ，音声対話エージェント，自動音声翻訳システム等を実現する上での最重要基盤技術の一つである。これまでに本研究代表者らは，任意の話者性を反映した多様な声質や発話様式・感情表現を持つ音声を生成・制御可能な音声合成方式を提案し，その基礎を確立した。そして，これらのアイディアを発展させ，「個性及び表現性ロバストな音声インタフェース」の概念を提唱し，その一環として平均声とスタイル制御に基づくロバスト音声合成の検討を行ってきた。

上記研究を進める中で，普段我々の生活で日常的行われている会話音声や講演音声等の自発性の高い音声は，音声特徴とりわけ韻律の多様性が高く，プロのナレータや声優の発声した朗読調や模擬感情音声に対して有用であった従来のアプローチでは合成音声の自然性が十分得られないことが明らかになった。これは，主に多様性を表現する上でのモデル化の不完全さや学習用音声言語資源の不足に起因するものであり，人間の実発声に取って代わる「どこにでも使える音声合成」の実現に大きな障壁となる。同様の問題は，ユニバーサルコミュニケーションの実現に向けた多言語音声の合成でも起きている。多言語音声合成はヨーロッパやアジアなど多言語が使用されている地域において需要が高いが，一般に東南アジア諸国では音声言語資源が十分でなく，「限られた資源で自然性の高い音声を合成できる」技術の開発への期待は非常に大きい。

これに対し本研究は，ロバスト音声合成技術を深化・発展させ，上述の問題を解決するための要素技術を開発することを意図したものである。

２．研究の目的

本研究でテーマとするロバスト音声合成とは「利用可能な音声資源が乏しい条件下においても，人間の持つ個性や感情・スタイル等の多様性を表出可能で自然性の高い音声を合成する」ことと定義される。本研究ではこれを発展させ，ロバスト音声合成に関する以下の項目について，新たな手法の提案と有効性の検討を行うことを目的とする。

(1) 表現性にロバストな音声合成

多様なスタイル（感情表現・発話様式）を含む表現力豊かで自然な音声を合成するために，学習データの量や質の変動に対して頑健な音声合成用モデルの学習法を確立する。

(2) 自発音声・会話音声の合成

朗読調音声や模擬感情音声に比べてより自発性の高い一般話者の日常的な会話音声に対して，自然な韻律をもった合成音声を生成するシステムの構築をめざし，新たなモデル化手法を検討する。

(3) 音声資源が乏しい言語の音声合成

本研究開始時点で音声資源が乏しい環境にあるタイ語やインドネシア語などの東南アジア諸国言語を対象に，これらの母語話者研究者と協力の下，多様な音声合成を実現する手法を検討する。

(4) 多言語の音声合成

統一的なモデルにより２ヵ国以上の音声の合成を可能にする新たな手法を検討し，多言語音声コミュニケーションに向けた音声合成システムの開発をめざす。

３．研究の方法

(1) 表現性にロバストな音声合成

多様な発話様式や感情表現などのいわゆる「スタイル」を伴う合成音声の品質は，学習データに大きく依存する。そこで，学習データに依存しにくい頑健なスタイルモデル学習法として，音声のスタイル表出度合を考慮したモデル学習法を導入する。また，話者性に関する平均声方式の考え方を拡張し，目標音声スタイルの学習データを必要としないスタイル音声合成手法の検討を行う。さらに，音響特徴量のノンパラメトリックモデル化や局所的系列内変動を考慮したパラメータ生成手法など，より自然な音声合成をめざした手法についても検討する。

(2) 自発音声・会話音声の合成

一般の話者による日常生活会話といった自発性の高い音声の合成が難しい理由一つは，韻律の多様性が大きく，正確なモデル化が十分にできないことが挙げられる。ここでは，韻律のモデル化の際に，従来の音韻単位ではなく韻律単位でモデル化する手法や，アクセント正規化学習を導入した手法について検討を行う。

(3) 音声資源が乏しい言語の音声合成

モデル学習に利用可能な音声資源が十分に整備されていない東南アジア諸国の言語として，タイ語とインドネシア語を対象とする。研究代表者らが従来から検討を進めているタイ語に関して，合成音声の了解性と自然に大きな影響を及ぼすトーン（声調）のモデル化手法についてより詳細な検討を行う。さらに，インドネシア語に関して現地研究者の協力の下，音声言語資源の基盤整備やプロトタイプ音声合成システムを構築する。

(4) 多言語の音声合成

平均声方式を話者の多様性ではなく言語の多様性に適用した新たな手法を検討し，日本語・英語間のクロスリンガル音声合成に適用して提案手法の有効性や問題点を明らかにする。

４．研究成果

(1) 表現性にロバストな音声合成

合成音声のスタイルが制御可能な手法として本研究代表者らは重回帰隠れセミマルコフモデル（MRHSMM）に基づくスタイル制御手法を提案している。この手法はスタイルの表出度合を直観的に制御できる点で有用であるが，合成音声の表出度合が学習データの平均的特徴に大きく依存し，学習データによっては意図したスタイルの表出度合が得られないという課題があった。これを解決すると共に品質を改善する手法として，学習データの表出度合の主観評価結果と系列内変動（GV）を考慮した MRHSMM を用いるモデル化及びパラメータ生成手法を提案した〔雑誌論文⑨〕。

従来の MRHSMM に基づくスタイル制御手法では，学習データのスタイルの表出度合（スタイルベクトル）として適当な固定値を仮定してモデルを学習していた。これに対し提案法では，予め学習データのスタイルの表出度合を主観評価により発話毎に求め，個別に得られたスタイルベクトルの値に基づいてモデル学習を行う。これにより，学習データの変動に頑健なモデル化ができるだけでなく，単一のスタイルの学習データのみでモデル構築が可能となる利点が生じる。一方，隠れマルコフモデルに基づく音声合成（HMM 音声合成）では，GV を考慮してパラメータ生成を行うことにより合成音声の聴感上の品質が向上することが確かめられている。そこで，提案法でも GV を考慮した新たなパラメータ生成の枠組みを導入している。

図１に提案法による合成音声の自然性の主観評価結果の一例を示す。図に示す通り，提案法は従来法に比べ，有意に主観評価値が向上する結果が得られた。

ところで，任意話者のあるスタイル音声を合成するには，目標話者の目標スタイルの音声データを多少なりとも利用できることが前提となっている。しかし，任意の話者に対し様々なスタイルによる音声を収録しラベル付けを行うには高いコストが必要である

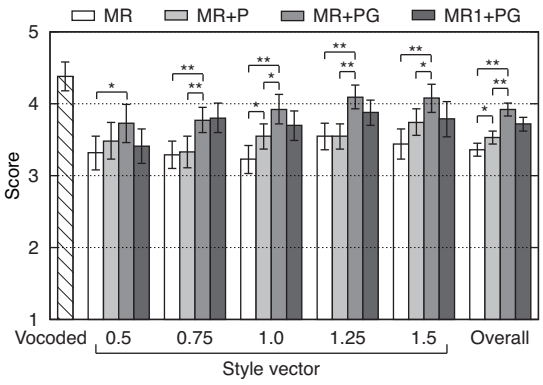

図1 「楽しい」スタイル合成音声の MOS 値による自然性評価結果（MR:従来法，MR+P:提案法/GV なし，MR+PG:提案法，MR1+PG:提案法/目標スタイルのみで学習）〔雑誌論文⑨〕

ため，目標話者の読上げ音声のみからその話者の多様なスタイルの音声を合成できることが望ましい。これに対し本研究代表者らは，先に平均声方式に基づく不特定話者スタイル変換法を提案した。この手法は，モデル学習に目標話者の目標スタイル音声を必要としない利点があるが，合成スタイル音声の品質は改善の余地があった。そこで，提案手法の品質改善を目的として話者適応学習を導入した新たなモデル構築法を提案した〔雑誌論文⑧〕。

提案手法では，あらかじめ複数話者の読上げおよび目標スタイルによる音声を用意し，各スタイルの平均声モデルを学習しておく。そして，スタイル適応の枠組みを用い，読上げスタイルから目標スタイルへの話者非依存の変換行列を求める。変換行列を推定する際には話者適応学習（SAT）を導入し，正規化された特徴量を用いてスタイル変換行列の再推定することにより，個々の話者の音響的差異に依存しにくいスタイル変換行列を求める工夫をしている。そして，得られた変換行列を目標話者の読上げスタイルのモデルに適用することでその話者のスタイルを目標スタイルへと変換している。

図２に４名の話者によりスタイル変換行列を学習し，目標話者の読上げスタイルモデルから宣伝口調スタイルに変換したモデルを用いて合成した音声の主観評価結果を示す。モデル学習に SAT を導入することにより，主観評価スコアが有意に改善する結果が得られた。

スタイル音声に限らず，統計的パラメトリック音声合成の枠組みにおいて，合成音声の品質改善を目的とした音声のモデル化及びパラメータ生成手法についても二つの新たな手法を提案した。一つは音声特徴量の局所的系列内変動のモデル化とパラメータ生成手法であり，もう一つはガウス回帰過程に基づいたモデル化とパラメータ生成手法である。

前述の通り，HMM 音声合成において系列内変動（GV）を考慮することにより合成音声の品質が改善することが知られている。しかし，GV は発話毎の静的特徴量系列の分散であるため，発話内における音素などに依存した分散の変化をモデル化することができない。これに対し，GV の概念をフレーム単位に拡張し

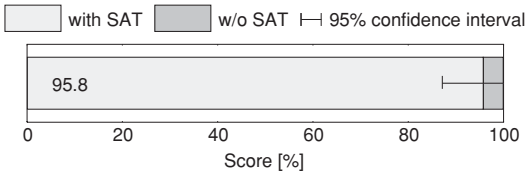

図 2 目標話者の目標スタイル学習データを用いずに変換したモデルからの合成音声の対比較評価結果（読上げから宣伝口調への変換，with SAT:SAT 有，w/o SAT:SAT 無）〔雑誌論文⑧〕

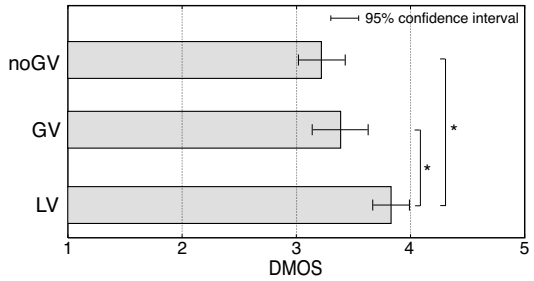

図 3 局所的系列内変動(LV)を考慮した HMM 音声合成に置ける合成音声の再現性の評価結果〔雑誌論文⑦〕

た局所的系列内変動（LV）は，GV に比べ一発話内の音響特徴量の変動をより良くモデル化できると考えられる。そこで，LV 系列を従来の音響特徴量系列と同様に HMM によりモデル化し，LV を考慮した HMM からのパラメータ生成手法を提案した〔雑誌論文⑦〕。

図 3 に LV を考慮したモデルから合成した音声の再現性に関する主観評価結果を示す。GV を考慮しない場合（noGV）や GV を考慮した場合と比べて，有意に良いスコアが得られた。

一方，HMM 音声合成に代わる新たな統計的音声合成の枠組みとして，ガウス過程回帰に基づくフレームレベルのモデルを用いる合成手法を提案した〔雑誌論文⑥〕。ガウス過程回帰はノンパラメトリックベイズモデルとして知られており，学習データ量に応じて柔軟にモデル化を行うことが期待できる。提案手法では，連続音声の合成を行うために，スパース行列を用いた近似と畳み込みカーネルを導入することにより計算量を削減しつつ，さらに音素境界における共分散を滑らかにするために，フレームコンテキストを拡張し，カーネルとして畳み込みカーネルを用いている。

図 4 に，提案手法をスペクトル特徴量のモデル化とパラメータ生成に適用し，合成音声の主観評価を行った結果を示す。従来の HMM 音声合成手法（GPR-MGE）と異なり，提案手法（GPR-LS，GPR-PE）は動的特徴量を使用していないにも関わらず，同等かそれ以上の主観評価スコアが得られている。また，系列内変動を考慮することにより，さらに合成音声の品質が改善することも確かめられている

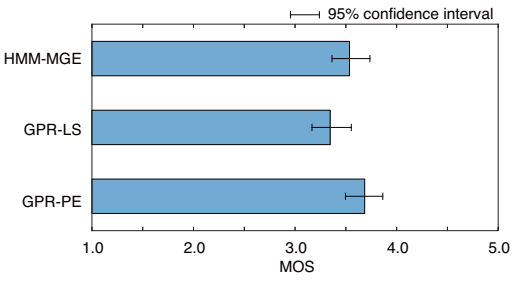

図 4 ガウス過程回帰に基づくスペクトル特徴量のモデル化とパラメータ生成による合成音声の主観評価結果(MOS 値)〔雑誌論文⑥〕

〔雑誌論文⑤〕。

この他にも，学習データ量の変化に対してロバストなモデル化手法として，新たな音素コンテキストセットの利用〔学会発表④〕と基本周波数に関する正規化学習の導入〔学会発表②〕を提案し，評価実験を通してその有効性を示した。

以上の成果は，表現性にロバストな音声合成に向けた頑健なモデル学習・パラメータ生成に有用と考えられ，今後のさらなる研究展開や応用が期待できる。

(2) 自発音声・会話音声の合成

自発性の高い音声では韻律の多様性が大きく，自然性の高い合成音声を生成するための新たなモデル化手法が望まれている。そこで，従来の HMM 音声合成で用いられている音韻単位ではなく，韻律イベントに基づいたモデル化単位を導入し，さらに基本周波数（F0）の観測値の不連続性を考慮した HMM によりモデル化する手法を提案した〔雑誌論文⑩〕。

図 5 に韻律イベントに基づいた HMM のモデル化単位の概念を示す。モデル化対象である F0 の値は，有声音区間で実数値，その他の区間では値を持たないことから，不連続な観測事象となる。提案手法では，有声の単一空間確率分布からなる F0 モデルを用い，有声音以外の区間を隠れ変数として有声区間の観測系列に対して尤度を最大化することによってモデルを学習している。

日本語話し言葉コーパス（CSJ）を用いて F0 のモデル化を行った結果，提案手法は従来手法に比べ，より少ないモデルパラメータ数で同等のモデル化性能が得られた。

ところで，自発性の高い音声を合成する際，合成音声の品質を改善するためには，音声が持つ多様性に見合った量の学習データが必要である。一方で自然な韻律生成のためのモデル学習には韻律情報ラベリングが必須であり，大量の音声の正確なラベリングは高コストになりがちである。

この問題の解決に向けて，アクセント型・アクセント句境界を同時推定する手法を提案した〔雑誌論文③〕。提案手法では，言語モデルに加えて HMM により表現された F0 の音響モデルを利用し，両方のモデルを考慮した尤度が最大となるように，アクセント型・アクセント句境界を決定している。プロナレータの読上げスタイル音声を対象として行

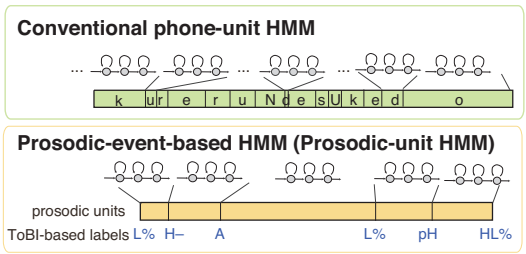

図 5 HMM の音韻単位モデル化と韻律イベントに基づいたモデル化の概念

った詳細な評価実験の結果，提案手法は従来手法に比べ手作業によるラベリングにより近い結果が得られることを示した〔学会発表①〕。

この他にも，音韻・韻律コンテキストバランスを考慮した音声コーパス構築手法に基づいて，インターネットのツィート文等，自発性の高い音声データベース整備とそれを学習データに用いたモデル化の検討を行った〔学会発表③⑥〕。

(3) 音声資源が乏しい言語の音声合成

タイ語は声調（トーン）言語の一つであり，トーンの再現性が合成音声の了解性・自然性に大きく影響する。タイ語のトーンは５種類に分類されるが，個人性，発話内容，発話環境等に影響されて多様に変化するため，その正確なモデル化は容易ではなく，タイ語合成音声においてトーンの再現性・自然性を向上させることが課題となっていた。そこで，従来提案されているタイ語 HMM 音声合成において，韻律モデル化の際に考慮されていなかった変動要因（コンテキスト）として，ストレスの有無の情報が重要であることを示し，ストレス情報を考慮したトーンモデル化手法を提案した〔雑誌論文④〕。さらに，このストレス情報を自動ラベリングする手法も提案した〔雑誌論文①〕。

インドネシア語音声合成に関しては，音声資源の基盤整備から始め，インドネシア・スラバヤ工科大学の研究者の協力を得て，まず男女１名の話者について約 1500 文章の読上げスタイル音声を収録した。これを音声データベースとして整備し，プロトタイプの音声合成システムを構築した。今後は研究代表者らが開発した手法の適用や問題点の明確化等，引続き協力して研究を進める予定である。

(4) 多言語の音声合成

多言語対応の自動音声翻訳システムを実現する上では，多言語に対応した音声合成システムが不可欠であり，マルチリンガルやポリグロット音声合成の研究が進められている。クロスリンガル音声合成もその一つであり，母国語しか話せないユーザに対し，そのユーザの声の特徴を保ちつつ外国語音声を合成出力する技術である。

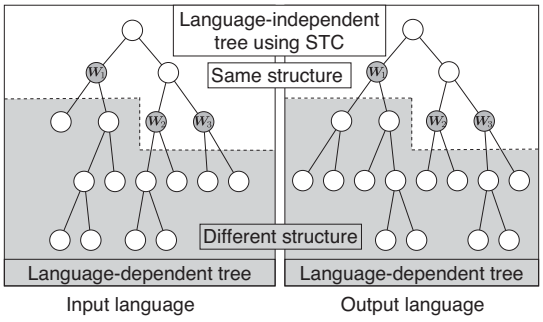

図 6 クロスリンガル話者適応における共有決定木コンテキストクラスタリング（STC）に基づく言語非依存木の概念

クロスリンガル音声合成としてこれまでに，声質変換に基づく手法や状態マッピングに基づく手法が提案されているが，未だ品質は十分とは言えない。そこで，合成音声の自然性改善を目的として，共有決定木コンテキストクラスタリング（STC）を用いた平均声からのクロスリンガル話者適応に基づく手法を提案した〔学会発表⑧〕〔雑誌論文②〕。

この手法では，目標話者の母国語言語（入力言語）と目標言語（出力言語）のそれぞれの平均声モデルをあらかじめバイリンガル音声データを用いて学習しておく。そして，目標話者の音声と入力言語の平均声モデルの間で話者適応により求めた変換行列を出力言語の平均声モデルに適用することにより，出力言語の目標話者モデルを求める。変換行列を求める際には，STC に基づいた２つの言語に非依存な木とそれぞれの言語に依存した木の２段階構造とする（図 6 参照）ことにより，平均声モデル間の分布パラメータだけでなく，言語間で共通する調音情報などもコンテキストとして考慮することができる。提案手法を英語／日本語のクロスリンガル音声合成に適用して合成音声の評価を行い，合成音声の自然性が向上する事を確認している。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 42 件）

① Decha Moungsri, Tomoki Koriyama, Takao Kobayashi, HMM-based Thai speech synthesis using unsupervised stress context labeling, 査読有, Proc. 2014 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2014, pp.1-4, DOI: 10.1109/APSIPA.2014.7041599, 2014.

② Daiki Nagahama, Takashi Nose, Tomoki Koriyama, Takao Kobayashi, Transform mapping using shared decision tree context clustering for HMM-based cross-lingual speech synthesis, 査読有, Proc. 15th Annual Conference of the International Speech Communication Association, INTERSPEECH 2014, pp.770-774, http://www.isca-speech.org/archive/archive_papers/interspeech_2014/i14_0770.pdf, 2014.

③ Tomoki Koriyama, Hiroshi Suzuki, Takashi Nose, Takahiro Shinozaki, Takao Kobayashi, Accent type and phrase boundary estimation using acoustic and language models for automatic prosodic labeling, 査読有, Proc. 15th Annual Conference of the International Speech Communication Association, INTERSPEECH 2014, pp.2337-2341, http://www.isca-speech.org/archive/archive_papers

/interspeech_2014/i14_2337.pdf, 2014.
④ Decha Moungsri, Tomoki Koriyama, Takashi Nose, Takao Kobayashi, Tone modeling using stress information for HMM-based Thai speech synthesis, 査読有, Proc. 7th International Conference on Speech Prosody, SPEECHPROSODY 7, pp.1057-1061, 2014.
⑤ Tomoki Koriyama, Takashi Nose, Takao Kobayashi, Parametric speech synthesis based on Gaussian process regression using global variance and hyperparameter optimization, 査読有, Proc. 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2014, pp.3862-3866, DOI: 10.1109/ICASSP.2014.6854319, 2014.
⑥ Tomoki Koriyama, Takashi Nose, Takao Kobayashi, Statistical parametric speech synthesis based on Gaussian process regression, 査読有, IEEE Journal of Selected Topics in Signal Processing, Vol.8, pp.173-183, DOI: 10.1109/JSTSP.2013.2283461, 2014.
⑦ Takashi Nose, Vataya Chunwijitra, Takao Kobayashi, A parameter generation algorithm using local variance for HMM-based speech synthesis, 査読有, IEEE Journal of Selected Topics in Signal Processing, Vol.8, pp.221-228, DOI: 10.1109/JSTSP.2013.2283459, 2014.
⑧ Hiroki Kanagawa, Takashi Nose, Takao Kobayashi, Speaker-independent style conversion for HMM-based expressive speech synthesis, 査読有, Proc. 2013 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2013, pp.7864-7868, DOI: 10.1109/ICASSP.2013.6639195, 2013.
⑨ Takashi Nose, Takao Kobayashi, An intuitive style control technique in HMM-based expressive speech synthesis using subjective style intensity and multiple-regression global variance model, 査読有, Speech Communication, Vol.55, pp.347-357, DOI: 10.1016/j.specom.2012.09.003, 2013.
⑩ Tomoki Koriyama, Takashi Nose, Takao Kobayashi, Discontinuous observation HMM for prosodic-event-based F0 generation, 査読有, Proc. 13th Annual Conference of the International Speech Communication Association, INTERSPEECH 2012, pp.462-465, http://www.isca-speech.org/archive/archive_papers/interspeech_2012/i12_0462.pdf, 2012.

〔学会発表〕(計 39 件)

① 増子理菜, 言語モデルと音響モデルを用いた自動韻律ラベリングの評価, 日本音響学会 2015 年春季研究発表会, 2015 年 3 月 16 日, 中央大学(東京都文京区).
② 大西浩之, HMM 音声合成における正規化学習を用いたアクセント誤り削減の検討, 日本音響学会 2014 年春季研究発表会, 2014 年 3 月 10 日, 日本大学(東京都千代田区).
③ 荒生侑介, 音声合成のための音韻・韻律コンテキストを考慮した文選択アルゴリズムの評価, 日本音響学会 2014 年春季研究発表会, 2014 年 3 月 10 日, 日本大学(東京都千代田区).
④ 舘野英樹, HMM 音声合成のための音節出現頻度にロバストな音素セットの検討, 日本音響学会 2014 年春季研究発表会, 2014 年 3 月 10 日, 日本大学(東京都千代田区).
⑤ 小林隆夫, 多様な音声合成に向けた取組みと課題, 第 15 回音声言語シンポジウム, 2013 年 12 月 20 日, 筑波大学(東京都文京区).
⑥ 荒生侑介, 対話音声合成のための音韻・韻律コンテキストを考慮した音声コーパス構築法の検討, 日本音響学会 2013 年春季研究発表会, 2013 年 3 月 15 日, 東京工科大学(東京都八王子市).
⑦ 能勢 隆, 統計モデルに基づく音声合成における話者・スタイルの多様化, 電子情報通信学会・日本音響学会音声研究会, 2013 年 1 月 31 日, 同志社大学(京都府京田辺市).
⑧ 能勢 隆, 共有決定木を利用した話者適応に基づくクロスリンガル音声合成の検討, 日本音響学会 2012 年秋季研究発表会, 2012 年 9 月 20 日, 信州大学(長野県長野市).

〔その他〕
ホームページ等
http://www.kbys.ip.titech.ac.jp/

６．研究組織

(1)研究代表者
小林　隆夫 (KOBAYASHI, Takao)
東京工業大学・大学院総合理工学研究科・教授
研究者番号：70153616

(2)研究分担者
能勢　隆 (NOSE, Takashi)
東北大学・大学院工学研究科・講師
研究者番号：90550591

(3)連携研究者

(4)研究協力者
郡山　知樹 (KORIYAMA, Tomoki)
東京工業大学・大学院総合理工学研究科・助教
研究者番号：50749124

ARIFIANTO, Dhany
スラバヤ工科大学・工学物理学科・講師