氏名	タカ キ シン ジ **高 木 信 二**
学位の種類	博士（工学）
学位記番号	博第958号
学位授与の日付	平成26年3月31日
学位授与の条件	学位規則第4条第1項該当　課程博士
学位論文題目	CONTEXTUAL ADDITIVE STRUCTURES IN HMM-BASED SPEECH SYNTHESIS （HMM音声合成のためのコンテキストの加算構造）
論文審査委員	主査　教授　徳田恵一 教授　北村正 教授　松尾啓志 准教授　李晃伸

論文内容の要旨

Speech is the most important ways for human communication, and a number of research topic for human-machine communication have been proposed. Automatic speech recognition (ASR) and text-to-speech synthesis (TTS) are fundamental technologies for human-machine communication. In recent years, they are used in many application such as car navigation system, information retrieval over the telephone, voice mail, speech-to-speech translation (S2ST) system, and so on. The goal of ASR and TTS systems is perfect speech recognition and speech synthesis with natural human voice characteristics.

Most state-of-art speech synthesis systems are based on large amounts of speech data. This type of approach is generally called corpus-based systems. This approach makes it possible to dramatically improve the performance compared with early systems such as rule-based one. In these days statistical approaches based on hidden Markov models (HMMs) have been dominant in TTS, due to their ease of implementation and modeling flexibility. In this approach, the HMMs are used for modeling sequences of speech spectra. In this paper, improved techniques for acoustic modeling are proposed for HMM-based speech synthesis.

It is well known that spectral features are affected by contextual factors, e.g., phoneme identities, accent, parts-of-speech, etc., and extracting the context dependencies is a critical problem for acoustic modeling. One of the major difficulties in the context

dependent modeling is finding a good balance between model complexity and availability of training data. In this paper, a novel acoustic modeling is proposed for representing complicated context dependencies.

First, an acoustic modeling with contextual additive structures in HMM-based speech synthesis is proposed. To represent more moderate dependencies between contextual factors and acoustic features, an additive structure of acoustic feature components that have different context dependencies has been proposed for HMM-based speech recognition. Contextual additive structure models can represent complicated dependencies between acoustic features and context labels using multiple decision trees. However, the computational complexity of the context clustering is too high for the full context labels of speech synthesis. To overcome this problem, this paper proposes two approaches; covariance parameter tying and a likelihood calculation algorithm using the matrix inversion lemma. Additive structure models can be applied to HMM-based speech synthesis using these techniques and speech quality would significantly be improved. Experimental results show that the proposed method outperforms the conventional one in subjective listening tests.

Next, a technique for constructing independent parameter tying structures of mean and variance using additive structure models for HMM-based speech synthesis is proposed. Conventionally, an HMM stream-level tying structure is constructed in HMM-based speech synthesis, i.e., mean vectors and variance matrices have exactly the same parameter tying structure. However, it has been reported that a clustering technique of mean vectors while tying all variance matrices improves the quality of synthesized speech. This indicates that mean and variance parameters should have different optimal tying structures. In the proposed technique, the decision trees for mean and variance parameters are simultaneously grown by taking into account the dependency on mean and variance parameters. Experimental results show that the proposed technique outperforms the conventional one.

Finally, I proposed a spectral modeling technique based on a contextual partial additive structure which provides an efficient representation of context dependencies to acoustic features for HMM-based speech synthesis. The contextual additive structure models assume that the observation vectors are generated from the sum of additive components with tree regression structures and they can be regarded as an intermediate structure between linear regression and tree regression. However, the additive structure models still have a limitation that the number of additive components is fixed for all output probability distributions. The proposed technique is a generalization of the additive structure models which have variable number of additive components dependently on contextual sub-spaces, and the clustering algorithm for extracting partial additive structure is provided. Experimental results show that the proposed technique outperformed the technique of extracting only standard additive structures in a subjective test.

論文審査結果の要旨

音声は人間同士のコミュニケーションにおいて，最も重要なコミュニケーションツールの一つであり，近年，音声を用いた機会との情報伝達が注目を集め，盛んに研究が行われている．このような流れの中で，ユーザへの情報提供，ユーザとのコミュニケーションを取ることを目的とした音声対話システムの実用化に向けた試みが数多くされ，その中で音声合成は，音声認識とともに音声対話を実現するための重要な技術となっている．音声合成における代表的な枠組みである，隠れマルコフモデル(Hidden Markov Model; HMM)に基づく音声合成手法ではスペクトル，基本周波数および音韻継続長を同時にモデル化し，得られたモデルから動的特徴量を考慮してパラメータ生成を行うことにより，音声を合成する手法である．HMMは時間とともに変動する観測系列を統計的な枠組で扱うことができるため，音声のモデル化に適していると考えられる．HMM音声合成手法は統計モデルに基づく手法であり，学習データから自動的にモデルを学習できることやパラメータを変換することで様々な声質に変換できるなどの特徴がある．このHMMでの音声のモデル化においては，同一の音素でも文脈的な要因(コンテキスト)を考慮することで，より精度の高いモデルを構築することができることが知られている．コンテキスト依存性の抽出には決定木に基づくコンテキストクラスタリング手法が広く用いられているが，コンテキスト依存性の表現は合成音声に多大な影響を与えることから，より適切なコンテキスト依存性の表現が期待されている．本論文では，より高性能なHMM音声システムの構築のため，適切なコンテキスト依存性の表現が可能な音響モデル構築手法を提案している．

まず，HMM音声合成のためのコンテキストの加算的構造に基づく音響モデルを提案している．従来のコンテキスト依存性の表現は決定木に基づくコンテキストクラスタリング手法が広く用いられているが，この手法では各リーフノードの分布が直接音響特徴量に対応付いている．しかし，このような直接的なモデル化が必ずしも適切とは言えない．この問題に対して，コンテキストの加算的構造を考慮した音響モデリングを用いて解決している．加算構造モデルでは音響特徴量が複数の加算因子の和として表現され，各加算因子は異なるコンテキスト依存性を持つと仮定する．加算構造モデルでは複数の決定木に基づくコンテキストクラスタリング手法により，尤度を用いて学習データから自動的に加算構造の抽出を行うが，HMM音声合成ではHMM音声認識と比べ非常に多くのコンテキストを考慮するため，加算構造モデルで用いるコンテキストクラスタリングの計算量が実現不可能なほど膨大になる．本論文では，コンテキストクラスタリング時のパラメータ推定の計算量削減のため，共分散共有と逆行列の補題に基づく尤度計算手法を提案している．これにより，これまで適用が困難であったHMM音声合成への加算構造モデルの適用を可能としている．また，客観評価実験，主観評価実験により提案法の有効性を確認している．

また，平均，分散パラメータの共有構造の同時最適化クラスタリングアルゴリズムを提案している．この手法では加算構造モデルの枠組みを用い，平均，分散パラメータの共有構造の同時構築を行う．従来のHMM音声合成では平均，分散パラメータが同じ共有構造を持つことを仮定しているが，提案法では平均，分散パラメータそれぞれに個別に適切な共有構造の構築が期待できる．これにより高精度な音響モデルが構築できると考えられるため，合成音声の品質の向上が期待できる．主観評価実験により提案法の有効性を確認している．

最後に，コンテキストの部分的な加算構造の抽出手法を提案している．加算構造モデルでは全ての出力分布において固定数の加算因子が用いられるが，適切な因子数はコンテキストに依存して異なることが考えられる．この問題を対し，決定木の中間ノードにおける因子の抽出を考え，学習データから尤度を用いて自動的に従来の加算構造に加え部分的な加算構造を抽出する手法を提案している．主観評価実験により提案法の有効性を示している．

以上のように，本論文ではHMM音声合成システムの性能向上を目的とした音響モデルの高精度化が提案されており，その有効性を示している．また，本論文の内容は国際学会にて公表されている．よって，本研究は情報工学の分野において寄与するところが多大であり，博士論文として十分価値あるものと認める．