（ 続紙 1 ）

京都大学	博士（情報学）	氏名	坂井 信輔
論文題目	A Probabilistic Approach to Concatenative Speech Synthesis		

（論文内容の要旨）

In recent years, corpus-based concatenative methods for speech synthesis by unit selection have received increasing attention because of their ability to synthesize naturally sounding speech. A speech synthesis system based on this approach is equipped with a large inventory of phone- or subphone-sized synthesis units. At run time, this inventory (called the unit database) is searched for the best sequence of synthesis units for the input text. This task is usually formulated as a cost minimization problem, in which each unit is assigned the target costs that measures how close its features are to the model values and the concatenation costs that measures how naturally it is connected with the surrounding units. Thus, the sequence of units that minimizes the overall sum of these costs is sought in the unit database.

This thesis adopts a probabilistic approach to target and concatenation modeling, where models can be trained from the corpus based on statistical learning techniques. To improve fundamental frequency (or F0) target modeling, an additive modeling framework is adopted in which the whole F0 contour is regarded as the sum of F0 component functions from several different levels. The concatenation cost is formulated through probabilistic models that captures how likely it is to observe the spectral shape of the current unit given the spectral shape of the previous unit in the current linguistic context. The issue of the large amount of computation in the unit selection search is also addressed. Two novel schemes for search efficiency are proposed by using the prior knowledge about the closed acoustic space of the unit database.

Chapter 1 and Chapter 2 describe an overview of the thesis and the concatenative speech synthesis system, respectively.

Chapter 3 presents a novel super-positional approach to F0 modeling based on a statistical learning technique called additive models. A two-layer additive F0 contour model is defined, consisting of intonational phrase-level and accentual phrase-level components, along with a least-square error criterion that includes a regularization term. A backfitting algorithm derived from this error criterion estimates the both components simultaneously by iteratively applying cubic spline smoothers. The F0 model was trained using a 7,000-utterance Japanese speech corpus and the proposed method is shown to be effective through RMS errors and correlation coefficients.

In Chapter 4, a derivation of the backfitting training algorithms for generic p-layer additive F0 models is presented for arbitrary positive integer p. This is the generalization of the two-layer additive model in the previous chapter. The additive F0 model has smoothing parameters that establish a trade-off between the fit to the training data and the smoothness of the fitted curves, which were all set to unity in the previous chapter. In this chapter, an optimization method for these parameters is developed using cross validation. By utilizing these methods, a three-layer additive F0 model for English is developed, consisting of intonational phrase, word-level, and pitch accent components. The model was trained using the Boston University Radio News Corpus and the effectiveness of the proposed method was shown by RMS errors and correlation coefficients.

Chapter 5 presents a probabilistic approach to concatenation modeling in which the goodness of concatenation is measured by the conditional probability of observing the spectral

shape of the current unit given the previous unit and the current phonetic context. This conditional probability is modeled by a conditional Gaussian density whose mean vector has a form of affine transform of the past spectral shape. Decision tree-based parameter tying is performed to achieve robust training that balances the model complexity and the amount of training data. The concatenation model was implemented in the speech synthesizer described in Chapter 2 and shown to be effective by objective closeness tests as well as subjective listening tests. The proposed method is a generalization of some popular conventional methods.

For reducing the amount of computation in unit selection, Chapter 6 proposes two early stopping schemes for Viterbi beam search with which we can stop early in the local Viterbi minimization for each unit as well as in the exploration of candidate units for a given target. They take advantage of the fact that the space of the acoustic parameters of the database units is closed and certain lower bounds of the concatenation costs can be pre-computed. The proposed method for early stopping is admissible in that it does not change the result of the Viterbi beam search as long as the lower bounds are correct. Experimental results show that the proposed methods effectively reduce the amount of computation while keeping its result unchanged.

Chapter 7 wraps up and concludes this thesis.

（続紙 ２ ）

（論文審査の結果の要旨）

本論文は、現在最も一般的に用いられている単位接続型の音声合成システムを改善するための方法に関する研究をまとめたものであり、得られた主な成果は次の通りである。

1. 音声合成の目標となる基本周波数(F0)のモデルに関して、アクセントとイントネーションのように複数の要因を加算的にモデル化する統計的かつ一般的な枠組み、及びこれを効率的に最適化する方法を考案した。日本語については２層加算モデルで、英語については３層加算モデルで、各々実装・評価し、良好な結果が得られることを示した。

2. 音声合成の単位を接続するモデルに関して、確率的な定式化、具体的には条件付きガウス分布によるモデル化を考案した。本手法は、従来の距離に基づくモデル化の一般化になっており、客観評価と主観評価の両方において高くなることを示した。

3. 単位接続型音声合成においては単位の探索・選択に多大な計算量を要しているが、合成単位データベースの音響特徴量がすべて事前に既知であることを利用して、適格な枝刈りを行う方法を考案し、実際に計算量が効果的に削減できることを示した。

以上のように本論文は、音声合成システムにおいて従来ヒューリスティックであった部分について確率統計的な枠組みによるモデル化を行い、音声品質の向上と処理時間の削減につながる方法を示したもので、学術上・実際上寄与するところが少なくない。よって、本論文は博士（情報学）の学位論文として価値あるものと認める。
また、平成23年12月27日実施した論文内容とそれに関連した試問の結果合格と認めた。

Webでの即日公開を希望しない場合は、以下に公開可能とする日付を記入すること。
要旨公開可能日：　　年　　月　　日以降