A Study on VLSI Design for Real-Time Large Vocabulary Continuous Speech Recognition

He, Guangji

(Degree)
博士（工学）

(Date of Degree)
2014-03-25

(Date of Publication)
2015-03-01

(Resource Type)
doctoral thesis

(Report Number)
甲第6102号

(URL)
https://hdl.handle.net/20.500.14094/D1006102

(別紙様式 3)

論文内容の要旨

氏　名　何 光霽

専　攻　情報科学専攻

論文題目（外国語の場合は，その和訳を併記すること。）

A Study on VLSI Design for Real-Time Large Vocabulary Continuous Speech Recognition
(実時間大語彙連続音声認識プロセッサ VLSI 設計技術に関する研究)

指導教員　吉本 雅彦

Speech recognition has been used recently in various applications such as automatic transcript, website indexing, navigation, mobile systems, ubiquitous systems, and robotics as a human interface. Large vocabulary continuous speech recognition (LVCSR) with acoustic and language models is too resource-hungry and power-sensitive for software applications. Hardware implementation by very large scale integrated circuit (VLSI) or field programmable gate array (FPGA) is demanded especially for use in mobile equipment and intelligent robots because of advantageous high processing speed and low power consumption.

This dissertation reports VLSI designs for large vocabulary real-time continuous speech recognition based on context-dependent Hidden Markov Model (HMM). As the background of this research area, the objective of this study and an overview of this dissertation are presented in Chapter 1. Then issues related to the VLSI implementation for real-time continuous speech recognition are noted in Chapter 2. The main issues are explained as four parts: 1) large-vocabulary are needed to support a higher word-cover rate. 2) low-power is required for longer working time and lower temperature, 3) high-speed is demanded for a more comfortable human interface, and 4) small area which means less recourse would be cost. Algorithm optimization and specialized architectures are proposed to solve these problems.

In Chapter 3, some algorithm optimizations for large vocabulary are presented: beam pruning using a dynamic threshold to cut the workspace and memory bandwidth for sort processing; the modified unigram language model which eliminated the update process for word-internal transitions; two-stage language model searching to reduce cross-word transitions which reduce the memory access greatly. The accuracy degradation of the important parameters in Viterbi computation is strictly discussed. These optimizations are utilized in all the three chips introduced in the following chapters.

Chapter 4 describes the first chip (HMM1) we implemented for 60-kWord real-time continuous speech recognition. It includes a cache architecture using locality of speech recognition, as some of the data that has been used in the current frame may be reused in the following frames, on chip caches are introduced for keeping some of them to reduce the external memory access, the specialized cache architecture achieve a hit-rate of 75%; a highly parallel Gaussian Mixture Model (GMM) architecture based on the mixture level; a variable-50-frame look-ahead scheme; and elastic pipeline operation between the Viterbi transition and GMM processing. Results show that our implementation achieves 95% bandwidth reduction (70.86 MB/s) and 78% required frequency reduction (126.5 MHz) comparing to the referential Julius system. The test chip, fabricated using 40 nm CMOS technology, contains 1.9 M transistors for logic and 7.8 Mbit on-chip memory. It dissipates 144

mW at 126.5 MHz and 1.1 V for 60 kWord real-time continuous speech recognition.

Chapter 5 describes the second chip (HMM2) we designed for 60-kWord real-time continuous speech recognition. We try to achieve a smaller area and higher processing speed. It features a compression-decoding scheme to reduce the external memory bandwidth for Gaussian Mixture Model (GMM) computation and a 4-path Viterbi transition units. We optimize the internal SRAM size using the max-approximation GMM calculation and adjusting the number of look-ahead frames. The test chip, fabricated in 40 nm CMOS technology, occupies 1.77 mm × 2.18 mm containing 2.52 M transistors for logic and 4.29 Mbit on-chip memory. The measured results show that our implementation achieves 34.2% required frequency reduction (83.3 MHz), 48.5% power consumption reduction (74.14 mW) for 60 k-Word real-time continuous speech recognition compared to the previous work while 30% of the area is saved with recognition accuracy of 90.9%. It can maximally process 2.4× faster than real-time at 200 MHz and 1.1 V with power consumption of 168 mW.

Chapter 6 describes the third chip (HMM3) we designed for 60-kWord real-time continuous speech recognition. As the "on-chip" latency (stop processing until the data come to the compute unit) and the "off-chip" latency (loading latency caused by DRAM) strongly delay the processing speed, we try to reduce the "off-chip latency" by implementing a two-level cache architecture and propose a 8-path Viterbi transition unit to reduce the "on-chip latency", which maximize the processing speed. Furthermore, we adopts tri-gram language model to improve the recognition accuracy. Measured results show that our implementation achieves 25% required frequency reduction (62.5 MHz) and 26% power consumption reduction (54.8 mW) for 60 k-Word real-time continuous speech recognition compared to the previous work. The chip can maximally process 3.02× and 2.25× times faster than real-time at 200 MHz using the bigram and trigram language models, respectively.

Finally, the conclusion of this study is presented in Chapter 7. This thesis presents the VLSI designs for low-power speak-independent large vocabulary real-time continuous speech recognition based on context-dependent Hidden Markov Model (HMM). The work contributes to achieve a comfortable human interface for mobile systems and robotics.

（別紙１） 論文審査の結果の要旨

氏名	何 光舞		
論文題目	A Study on VLSI Design for Real-Time Large Vocabulary Continuous Speech Recognition （実時間大語彙連続音声認識プロセッサ VLSI 設計技術に関する研究）		
審査委員	区 分	職 名	氏 名
	主 査	教 授	永田 真
	副 査	教 授	有木 康雄
	副 査	教 授	的場 修
	副 査	准教授	川口 博
	副 査		

要 旨

概要

近年，音声認識技術が発達し，その応用範囲はカーナビ，携帯機器，ロボット，会議の自動筆記，画像・音声検索のインデックス作成などに広がってきている．中でもモバイル機器，ウェアラブル機器，知能ロボットなどの発達が急速に進んでいる現在，人間とシステムが自然なコミュニケーションを図るためのヒューマンインターフェイスとして，音声認識が注目を浴びている．また、半導体の微細化技術の発展により，高性能な汎用プロセッサを用いれば，膨大な演算を必要とする大語彙リアルタイム連続音声認識をソフトウェアでの実現も可能となってきている．しかし，このような高性能な汎用プロセッサは膨大な電力を消費するため，モバイル・ウェアラブル機器，ロボットへの実装は困難であるため、専用ハードウェアでの実装が必要となっていた．

本論文は 7 章で構成されており，第 1 章は序論である．ここでは，研究背景と課題，そして本論文の目的について述べられている．第 2 章では大語彙連続音声認識 VLSI プロセッサにおいて四つの課題：大語彙，低消費電力，高速処理および小面積などについて述べられている．

第 3 章ではハードウェア向けアルゴリズムの改良について述べられている．閾値カット、Modified 1-gram 言語モデル、言語モデル粗密探索である。本研究では，Viterbi 演算後に行っていたソート処理を，Viterbi 演算内において閾値カットで行なっている。毎フレームにおいて最適な閾値を計算することで，精度に影響しない閾値カットを実現している。これにより、余分なワークスペースとメモリアクセスを低減している。 Modified 1-gram 言語モデルでは、遷移確率に 1-gram の更新をあらかじめ適用したモデルを作成することで単語内遷移から 1-gram 更新処理をなくし，同じ回路で単語内遷移，単語間遷移を行なえる構成としている。さらに、2-gram 確率の上位 10 個の始端のみへの遷移とし，N フレームに 1 回のみ全始端に遷移させる言語モデル粗密探索を提案している．密探索周期 N が 5 の場合，精度に影響しないことが確認でき，単語間遷移部分の演算量 78%削減を実現した．

第 4 章では実時間 6 万語彙連続音声認識プロセッサについて論述されている．大語彙、低消費電力のための高並列演算アーキテクチャ、キャッシュアーキテクチャおよびパイプライン処理を提案している．GMM 演算処理において，混合パラメータを一回読み込むにつき，50 フレームの音声特徴ベクトルをすべて計算する．つまり，混合パラメータを再利用することで，GMM 演算の外部メモリ帯域を 94%削減した．また、GMM 部分の莫大な演算量に対し，混合レベルの 16 混合並列 GMM 演算アーキテクチャを提案し、全状態計算を行うことで，Viterbi 処理とのパイプライン処理を行うことが可能になった．Viterbi 部分の外部メモリ帯域を削減するために，音声認識のフレーム相関性を利用して，メモリアクセスを大きく占めているアクティブノード，アクティブノードマップおよび 2-gram データをチップ上のキャッシュに導入した．また，キャッシュ更新機構を工夫することで，提案した Viterbi キャッシュアーキテクチャは 75%のヒット率を実現した．また、GMM 演算は毎フレームにおいて同じサイクル数がかかるが、Viterbi 遷移処理はアクティブノードの数、遷移状況により大きく変わるため，1 フレームごとのパイプライン処理は不効率であるため、GMM 全状態演算により，GMM・Viterbi のパイプライン処理を行えるようにした．また，パイプライン処理を扱えるフレーム数を可調整にすることで，パイプラインの効率を最大化にした．以上の提案手法を用いて，実時間 6 万語彙連続音声認識処理するために必要なメモリ帯域を 98%削減し，必要な動作周波数を 78%削減し、126.5MHz で 6 万語彙実時間音声認識を行うことができた．チップ試作及び評価により，126.5MHz@1.1V の条件下で 144mW の低消費電力で動作することを確認している。

氏名	何　光霽

第5章では、2.4倍速実時間6万語彙連続音声認識プロセッサについて述べられる．小面積のための圧縮・デコーダ機構、GMM近似演算アーキテクチャと高速化ための4パスViterbi遷移アーキテクチャを提案した．圧縮・デコーダ機構では、予め外部データベースにあるGMM演算用のパラメータを圧縮してからチップに転送する．チップ側で圧縮したパラメータをデコーダして，演算を始める．この手法によって，圧縮率分の外部メモリ容量と外部メモリ帯域を削減した。GMMパラメータの特性を考慮した圧縮アルゴリズムを使うことで，デコーダ部分の回路規模と演算量の増加も小さく抑えられている。近似GMM演算アーキテクチャを導入することで、ログ演算が不要になるため1Mbitの内部メモリを削減した。次に、ミスヒット時の演算の効率，各キャッシュへのデータパスの利用効率を上げるために，複数アクティブノードが扱える4−パスViterbi遷移アーキテクチャを実装した．チップ試作を行い評価したところ，実時間処理するために必要な83.3MHz動作時の消費電力は74.2mWであった．また，標準電圧(1.1V)で最大200MHz(168mW)動作が確認され，2.4倍速実時間動作を実現できた．従来比で30%の面積削減，34.2%の動作周波数削減および48.5%の消費電力削減を達成している．

第6章では、3倍速実時間6万語彙連続音声認識プロセッサについて述べられる．高速化するための2次キャッシュアーキテクチャ，8−パスViterbi遷移アーキテクチャおよび精度を上げるための簡単化3-gram遷移アーキテクチャを提案している．音声認識処理では，外部メモリをデータベースとして，ランダムでアクセスする．外部メモリの特性による，毎回アクセスにつき，アドレスのセットアップ、pre-chargeなどで数サイクルの"off-chip latency"が発生してしまう．この遅延を削減するために，二次キャッシュアーキテクチャを提案した．また、全体の処理速度のネックとなっているViterbi部分に8パスViterbi遷移アーキテクチャを実装することで，"on-chip latency"を削減した．さらに 簡単化3-gram言語モデルの採用により、2-gram遷移した後，スコア(蓄積確率)が一番高い言語リストから"best predecessor word"を特定できる．その"best predecessor word"からの3-gram遷移のみを行うことで，認識精度を約2%向上出来た。チップ試作を行い評価したところ，実時間処理するために必要な62.5MHz動作時の消費電力は54.8mWであった．また，標準電圧(1.1V)で最大200MHz(177.4mW)動作が確認され，3倍速実時間動作を実現できた．また，3-gramを使う場合，200MHzで最高処理速度は2.25倍速であり，消費電力は174.56 mWであった．従来比で25%の動作周波数削減および26%の消費電力削減を達成している．

以上第3章から第6章において，実時間大語彙連続音声認識VLSIプロセッサ設計技術に関する研究成果について記述し，最後に第7章において本論文を総括して結論を述べている．これらの研究成果は，3編の査読付き論文と4件の査読付き国際学会プロシーディングにて掲載されており，今後のモバイル端末、知能ロボットのヒューマンインターフェイス，高速音声認識エンジンに寄与する有効な手段となり得るものである．すなわち、今後の高度情報化社会構築の鍵となる重要で価値ある知見を得たものと認める。提出された論文は、システム情報学研究科学位論文評価基準を満たしており、学位申請者の何　光霽氏は，博士（工学）の学位を得る資格があると認める。