音声言語情報処理 29－28
(1999. 12. 21)

スカラ量子化を利用したクライアント・サーバ型音声認識の実現とサーバ部の高速化の検討

小坂 哲夫　植山 輝彦†　櫛田 晃弘　山田 雅章　小森 康弘

キヤノン（株）情報技術研究所　(†現 キヤノン（株）画像技術研究所)

〒211-8501 神奈川県川崎市中原区今井上町53

TEL: 044-733-6111, E-mail: kosaka@cis.canon.co.jp

あらまし　スカラ量子化による音声認識用音声符号化，および符号化データを利用したサーバ部の高速化という特徴を持つ，クライアント・サーバ型音声認識システムを実現した．クライアント部ではパラメータをスカラ量子化することにより10kbpsまたは5.2kbpsの符号化を行なう．サーバ部ではスカラ量子化データを直接参照し，table-lookupにより粗い尤度計算を行ない．さらに復号化したデータを用いて，尤度の高い部分の再計算を行ない，尤度計算全体の高速化を図る．粗い尤度計算には混合分布の次元独立演算法(IDMM)を用いる．以上の提案法を評価するため，符号化による圧縮を行なわない80kbpsシステム(baseline)と10kbps，5.2kbpsシステムの比較実験を行なった．この結果認識率の低下なしに圧縮が行なえることを確認した．

キーワード　音声認識，音声符号化，スカラ量子化，クライアント・サーバ，高速化，HMM

Client-server based speech recognition and its fast recognition algorithm using scalar quantization

Tetsuo KOSAKA, Teruhiko UEYAMA†, Akihiro KUSHIDA, Masayuki YAMADA, and Yasuhiro KOMORI

Information Technology Laboratory, Canon Inc.
†Visual Information Technology Development Laboratory, Canon Inc.
53 Imaikami-cho, Nakahara-ku, Kawasaki-shi, KANAGAWA 211-8501, Japan

TEL: 044-733-6111, E-mail: kosaka@cis.canon.co.jp

Abstract　This paper proposes a claient-server based speech recognition system, which is characterized by speech coding for speech recognition using scalar quantization and a fast recognition algorithm at the server side using the coded data. At the client side, speech parameters are coded into 10kbps or 5.2kbps by using scalar quantization method. At the server side, rough likelihood calculation is carried out by reference to scalar quantized data first. After that likelihood re-calculation is carried out by using decoded data. In this step, states which have the higher output probabilities are selected and re-calculated. This two-step algotirhm can save the total cost of the likelihood calculation. In comparative recognition experiments between 80kbps (baseline), 10kbps and 5.2kpbs systems, the results showed that the proposed algorithm could compless speech parameters without degradation of the recognition rate.

Key words　speech recognition, speech coding, scalar quantization, client-server, fast recognition algorithm, HMM

1 まえがき

携帯端末など小型機器のユーザインタフェースとして音声認識は有望な技術であるが，現在の携帯端末ではディクテーションや大語彙認識など，処理量が多い高度な音声認識はCPUやメモリなどの制約から実現は困難である．そこで，処理の軽い部分をクライアントで行い，処理が重い部分をサーバに振り分けるクライアント・サーバ型の音声認識が試みられている (例えば [1], [2] など)．

クライアント・サーバ型音声認識の一般的なインプリメンテーションとしては，クライアントで音声入力を行ない，波形データをそのままサーバへ転送するか，またはケプストラムなどの音声分析までクライアントで行なって，パラメータをサーバへ転送する．ここで問題となるのはデータの転送量である．一般に，以上に述べた方法を用いると，データ転送量は数十kbps程度となる．しかしネットワークのリソースなどを考慮するとデータ転送量は少ないほうが望ましい．

音声のデータ転送量を削減する試みは，音声符号化の分野で数多く試みられてきたが，一般的な音声符号化手法を利用して音声認識を行なうと，認識率の低下が生じる [3]．しかし，一般的な音声符号化はあくまで音声を再生する目的で設計されており，再生の必要がない音声認識のための符号化には必ずしも適していないと考えられる．そこで音声認識のための符号化手法が提案されている [4] [5]．

この音声認識用符号化手法は，主に音声パラメータのスカラ量子化やベクトル量子化といった手法をベースとして開発されているが，これらの手法はまた音声認識における出力確率の計算の高速化としても応用可能な技術である．出力確率計算の高速化手法として，パラメータのスカラ量子化やベクトル量子化を行ない，各量子化値に対し出力確率計算の一部をあらかじめ計算してテーブル化しておき，認識時に参照する方法がいくつか提案されている [7][8][9]．

以上から音声パラメータのベクトル量子化またはスカラ量子化手法を用いることにより，音声認識用の符号化および出力確率計算の高速化が共通の手法に基づき行なえることが予想される．本稿ではこのうちスカラ量子化に基づく方法をとり上げ，音声認識用符号化および符号化データを利用した出力確率計算の高速化手法を提案する．また認識実験によりその有効性について検討した結果について述べる．

本稿の構成は以下の通りである．第2章ではスカラ量子化を用いた音声パラメータの符号化法について述べる．第3章では高速出力確率計算法について述べる．第4章ではクライアント・サーバ型音声認識システムの概要について述べる．第5章では認識実験の結果および考察について述べる．第6章はまとめである．

2 音声パラメータの符号化法

音声認識用の符号化手法としては，まずケプストラムなどの音声パラメータを求め，それをスカラ量子化やベクトル量子化を用いて符号化する方法がとられる．文献 [4] においては，線形予測係数をもとに，多段VQを用いて4kbps程度に圧縮を行なっている．文献 [5][10] では，MFCCをベースに各次元をスカラ量子化またはサブベクトルに分割しそれをVQし符号化を行なっている．さらに不均一にビットを割り当てることにより (bit-allocation) 効率的な圧縮を行ない2.0〜2.6kbps程度を実現している．また国内においてはGruhnらが [5] の手法に基づきクライアント・サーバ型音声認識システムを構築し，ATRトラベル・アレンジメント・タスクにて5.6kbps程度まで認識率の低下なしに圧縮できることが確認されている [6]．

本稿では，クライアント側の負荷などを考慮し，音声パラメータを次元ごとにスカラ量子化する方法を用いた．以下その手法について述べる．

まず音声パラメータとしては12次元のLPCメルケプストラム，12次元のΔLPCメルケプストラム，およびΔパワーの計25次元のパラメータを用いる．このパラメータをスカラ量子化法により各次元4bitのデータに変換する．またΔパラメータはサーバ側でも計算できるため，Δパラメータを伝送しない方法についても検討した．ベースラインおよび各手法の伝送レートは以下の通りである．

ベースライン:80kbps 25次元の音声パラメータをfloat (4byte) のデータで伝送する．この場合はクライアント・サーバに分割しないシステムとの認識率の差はない．

$$25\text{dim} \times 100\text{frame/sec} \times 4\text{byte} \times 8\text{bit} = 80\text{kbps}$$

Δパラメータの伝送有:10kbps 25次元の音声パラメータの各次元を4bitに量子化し伝送．

$$25\text{dim} \times 100\text{frame/sec} \times 4\text{bit} = 10\text{kbps}$$

Δパラメータの伝送無:5.2kbps 上記において，Δパラメータの伝送をしない場合．

$$13\text{dim} \times 100\text{frame/sec} \times 4\text{bit} = 5.2\text{kbps}$$

音声パラメータのスカラ量子化コードブック (SQC) 作成は，以下の方法により行なう [11]．基本的には音声データからではなく，認識に用いる音響モデルの情報を用いてSQCの作成を行なう．音響モデルから作成する理由として，雑音環境下などで認識を行なう場合，その雑音環境ごとにLBGを用いてSQCを作成しなおすよ

り，環境適応した音響モデルからSQCを求めるほうが計算時間やデータ量が少なくてすむという利点による．

本システムでは音響モデルとして，混合連続分布型HMMを用いている．このHMMからSQCを作成するに当たって，以下の方法で量子化する範囲とその分割方法を定めた．まず量子化範囲は，各次元毎に，全分布をマージして得られる単一ガウス分布の平均$\hat{\mu}$と分散$\hat{\sigma}$を用いて$[\hat{\mu}-3\hat{\sigma}, \hat{\mu}+3\hat{\sigma}]$と定めた．$M$個の分布のマージは式(1)で求めた．なお，分布の重み係数は考慮していない．

$$\hat{\mu}=\sum_{i}^{M}\mu_i/M, \quad \hat{\sigma}=(\sum_{i}^{M}\sigma_i+\sum_{i}^{M}(\mu_i-\hat{\mu})^2)/M \qquad (1)$$

また分割方法としては，分割した各区間の混合分布の出力確率の積分値が等しくなるように，つまり等確率に分割する方法を取った．

3 高速出力確率計算法

本章では，クライアントから伝送されるスカラ量子化データを利用した，高速出力確率計算法について述べる．まず既に我々が提案している高速出力確率計算法を説明し，次にこの手法を，スカラ量子化を利用したクライアント・サーバ型音声認識に応用する方法について述べる．高速出力確率計算はまずIDMM+SQ法により出力確率を近似計算し，さらに高い出力確率を持つ一部の状態について高精度の音響モデルを用いて再計算を行なうことにより，認識率の低下を押さえつつ高速な出力確率計算を実現している．

3.1 IDMM+SQ法による高速出力確率計算法

本節ではIDMM+SQのアルゴリズムについて概説する[12]．IDMM(Independent Dimension Multi-Mixture computation)は混合分布の出力確率を近似計算する方法である．ここで，各混合要素は対角共分散行列ガウス分布とする．入力パラメータのスカラ量子化によるテーブル参照と共にIDMMを用いることによって，高速な出力確率計算を実行することが可能である．

対角共分散行列ガウス分布を要素とする混合分布を用いた場合，N次元の入力音声パラメータベクトル$\boldsymbol{x}$に対する状態sの出力確率$b_s(\boldsymbol{x})$は以下のように表される．

$$b_s(\boldsymbol{x})=\sum_{m=1}^{M_s} w_{s,m}\prod_{i=1}^{N}\mathcal{N}_{s,m,i}(x_i) \qquad (2)$$

これに対して，各次元が独立に計算できると仮定すると状態sの出力確率$\hat{b}_s(\boldsymbol{x})$は次のように定義できる．

$$\hat{b}_s(\boldsymbol{x})\stackrel{\text{def}}{=}\prod_{i=1}^{N}\sum_{m=1}^{M_s} w_{s,m}\mathcal{N}_{s,m,i}(x_i) \qquad (3)$$

ここで，M_sは状態sにおける混合数，$\mathcal{N}_{s,m,i}(\cdot)$は状態$s$の$m$番目の分布の$i$次元目のガウス分布関数である．また，$w_{s,m}$は状態$s$における$m$番目のガウス分布の重みである．以上がIDMMの定義となる．

実際の音声認識では対数出力確率が用いられる．そこで，IDMMによる対数出力確率として，式(3)の両辺の対数をとった次式が用いられる．

$$\log\hat{b}_s(\boldsymbol{x})\stackrel{\text{def}}{=}\sum_{i=1}^{N}\log\left(\sum_{m=1}^{M_s} w_{s,m}\mathcal{N}_{s,m,i}(x_i)\right) \qquad (4)$$

IDMMを入力のスカラ量子化と組合せた場合，対数出力確率を高速に計算することが可能である．式(4)において，$\sum_{i=1}^{N}$の要素は1次元の対数混合ガウス分布であるから，入力x_iを量子化することによってテーブル参照に置き換えることができる．したがって，IDMMによる対数出力確率計算は，N回のテーブル参照と$N-1$回の加算で実現できる．

上記のように，IDMMとスカラ量子化を組合せて用いた場合の計算量は混合数M_sによらない．一方，IDMMを用いずにスカラ量子化のみを用いた出力確率計算の計算量はM_sにほぼ比例する．したがって，IDMMの導入によって計算量が約$1/M_s$となる．

3.2 再計算による高精度化

以上により高速化が達成できるが，近似計算を行なっているため，特に混合数が増加した場合，計算精度が低下する恐れがある．そこで，ここでは高精度なモデルによる再計算をおこなって，近似誤差を低減する方法を使用した[13, 14]．

この方法のアルゴリズムは以下の通りである．

1. 音声認識を行なう前に，スカラ量子化コードブックに関して，混合ガウス分布の出力確率を次元独立に計算し，テーブルにしておく．
2. 入力ベクトルの各次元をスカラ量子化し，その結果に基づいて出力確率のテーブルを参照する．
3. 全次元にわたってテーブル参照の結果の和を計算し，粗いHMMの出力確率とする．
4. 粗いHMMの出力確率がある閾値よりも大きい場合には，通常の混合分布HMMの出力確率を計算するのと同様な方法で再計算を行なう．

つまり高速で粗い出力確率計算を全状態について行ない，選択された非常に少ない状態において精度の高い出力確率計算行なうことにより，全体として高速で精度の高い出力確率計算を実現する．

3.3 クライアント・サーバ型音声認識への応用

以上の高速出力確率計算法をクライアント・サーバ型音声認識へ応用する方法について述べる．80kbps，10kbps，5.2kbps の各場合のブロック図を 1，2，3に示す．

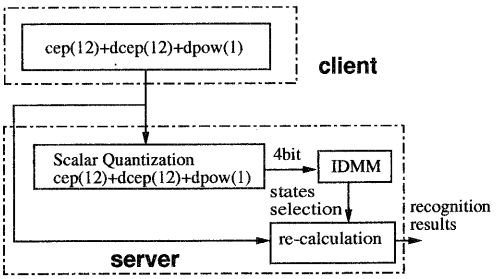

図 1: C/S 音声認識ブロック図 (80kbps)

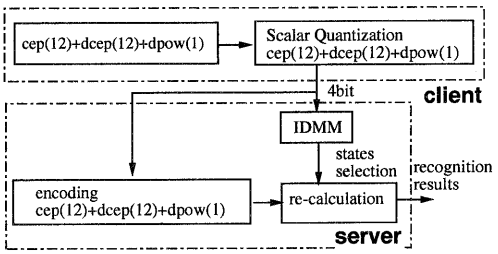

図 2: C/S 音声認識ブロック図 (10kbps)

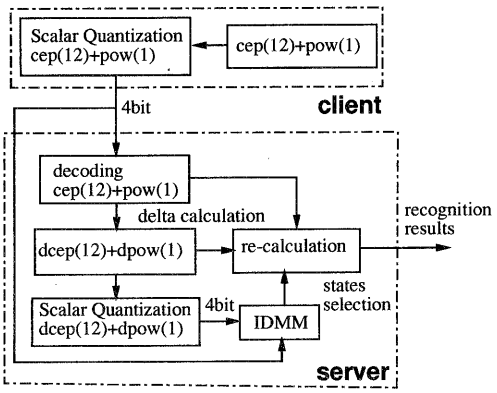

図 3: C/S 音声認識ブロック図 (5.2kbps)

80kbps システムの場合，すべてのパラメータを圧縮せずに伝送する．サーバ部において IDMM 用に 4bit データに変換する．また再計算用には伝送されたデータをそのまま用いる．

10kbps システムの場合，すべてのパラメータをスカラ量子化して 4bit データに変換して伝送する．IDMM 計算部ではこの 4bit データをそのまま用いてテーブル参照を行なう．また再計算部では復号化したデータを用いて計算するように実装した．但し再計算用に IDMM とは別にテーブルを持てば，復号化しなくとも再計算が可能と考えられる．

5.2kbps システムの場合，ケプストラムパラメータとパワーのみスカラ量子化して伝送する．サーバで復号化し，そこから Δパラメータおよび Δパワーを求める．さらにスカラ量子化により 4bit データに変換し IDMM 計算部に用いる．また伝送されたケプストラムの 4bit データもそのまま IDMM 計算部に用いる．再計算部においては，復号化されたケプストラムパラメータおよび，求められた Δパラメータ (+Δパワー) を用いる．この場合も復号化せずに直接 Δパラメータ (+Δパワー) を求めたり，再計算に用いたりする方法も考えられる．

4 クライアント・サーバ型音声認識システムの概要

本章ではクライアント・サーバ型音声認識システムの概要について述べる．

4.1 クライアント部

クライアント部では音声区間の検出および音響分析，さらに音声パラメータの符号化を行なう．クライアント部は小型携帯機器へのインプリメントを想定し，固定小数点演算，浮動小数点演算の両方に対応している．

4.2 データ伝送部

クライアント・サーバ型音声認識システムにおいてネットワークを利用する場合，問題となるのは 1) 通信中のデータ損失やデータ到着順序，および 2) 認識速度の低下である．2) については前章で説明した音声パラメータの符号化手法を用いて伝送データ量を削減することにより対処した．また 1) の問題を考慮して通信プロトコルとしては TCP/IP を利用する．TCP/IP ではデータの損失や順序に対する保証があるため，本システムに適したプロトコルと考えることができる．データの伝送に当たっては，クライアントとサーバがパイプライン処理により並行して処理が出来るよう，発声が終了してからデータを伝送するのではなく，発声中に順次データを伝送する方式を取った．

4.3 サーバ部

サーバ部の音声認識アルゴリズムとしては tree-trellis based search により N-best を求める方法を用いた [15]．前方向の探索では，各時点における最大尤度より一定値を下まわる閾値により枝を刈る Beam Search を併用し

た．また第3章で説明した，IDMM+SQ及び再計算による高速尤度計算法を利用している．

5 認識実験および考察

提案法の有効性を確認するため音声認識実験を行なった．具体的には80kbps(ベースライン)システムと10kbps，5.2kbpsシステムの比較実験を行なった．

5.1 認識実験条件

音響分析はサンプリング周波数11kHz，フレーム周期10ms，窓幅25.6ms，プリエンファシス0.97，特徴量としてLPCメルケプストラム12次+ ΔLPCメルケプストラム12次+Δ対数パワーを用いた．音声認識に用いた文法は日本全国都市名680単語であり，評価データとしては話者20名(男性10名, 女性10名)が文法中の100都市名を発声したデータを用いた．音声認識に用いたHMMは3状態12混合，対角化共分散行列の混合連続出力分布型で，ここでは約3,300種類のtriphoneHMMを使用した．但し状態はtop-down clusteringにより[16]，800状態程度に状態共有している．

5.2 認識実験結果及び考察

各種ビットレートに対しての認識実験結果を表1に示す．実験結果は音響分析を浮動小数点で行なった場合と，固定小数点で行なった場合の2種類について示す．

表1: 各種伝送レートにおける認識率(%)

bit rate(kbps)	floating point	fixed point
80(baseline)	98.1	98.2
10	98.2	98.4
5.2	98.4	98.1

浮動小数点演算及び固定小数点演算いずれにおいても，10kbps，5.2kbpsともベースライン（非圧縮）からの認識率の低下はみられない．逆に浮動小数点演算においては若干の認識率の向上が見られるが，これは誤差の範囲と考えられる．

ベースラインシステムと10kbps，5.2kbpsシステムの比較を行なった場合，原理的には以下の認識率低下の要因が考えられる．まず10kbpsシステムの場合，25次元の音声パラメータを4bitにスカラ量子化した場合の量子化誤差が認識率に影響を与える可能性がある．また5.2kbpsシステムの場合，ケプストラムパラメータに関しては，上記10kbpsシステムと比較して量子化誤差は同等である．これに対しΔケプストラム及びΔパワーに関しては，以下の2点によりさらに量子化誤差が大きくなると考えられる．

1. 再計算で用いるΔパラメータの計算手順
   5.2kbps: ケプストラム→符号化→復号化→Δケプストラム
   10kbps: ケプストラム→Δケプストラム→符号化→復号化
   つまりΔパラメータを符号化する順序が異なる．

2. IDMMで用いるΔパラメータの算出手順
   5.2kbps: ケプストラム→符号化→復号化→Δケプストラム→符号化
   10kbps: ケプストラム→Δケプストラム→符号化
   つまりΔパラメータは二重の量子化が行なわれている．

ベースラインと比較して，10kbpsシステム及び5.2kbpsシステムのいずれにおいても認識率低下は生じていないことから，以上に関しての量子化誤差は，認識率という観点から見た場合，大きくなかったと考えることが出来る．

6 まとめ

本稿では，スカラ量子化による音声認識用音声符号化，および符号化データを利用したサーバ部の高速化という特徴を持つ, クライアント・サーバ型音声認識システムについて検討した結果を述べた．音声認識用音声パラメータ符号化法と音声認識における尤度の高速計算法が同一の手法で行なえる点に着目し，音声パラメータ符号化及び高速尤度計算をスカラ量子化に基づき行なうクライアント・サーバ型音声認識手法を提案した．

以上の有効性を検討するために音声パラメータを圧縮しない場合(80kbps)と10kbps及び5.2kbpsに圧縮した場合とでの認識率の比較を行なった．この結果認識率の低下なしに5.2kbpsまでパラメータの圧縮ができることを確認した．

またIDMM+SQ及び再計算手法により認識率の低下なしに約1/3程度まで認識時間が短縮できることが分かっている[11]．サーバ部は文献[11]と同等な認識アルゴリズムを用いているため，今回のシステムでも認識時間に関し同様な結果になると予想されるが，今後実験により確認したい．さらに今後の課題としては，今回の実験では5.2kbps程度では認識率の低下が確認できなかったため，更なる圧縮手法の検討を行ないたい．この場合より複雑なタスク，音響環境での実験が必要と考えられる．また再計算部に直接スカラ量子化データを利用する方法についても検討する予定である．

謝辞 本システムに用いられた音声認識部の作成にご協力頂いた山本寛樹研究員及び中川賢一郎研究員に感

謝いたします．また御討論頂いた深田俊明研究員に感謝いたします．

参考文献

[1] 伊藤，甲斐，山本，中川: パソコン用連続音声認識クライアント・サーバシステムの実装，情報処理学会第 55 回全国大会，3J-5 (1997).

[2] 山田，野田，嵯峨山: 実時間動作を考慮した音声認識サーバ，音講論，2-8-2 (1994.10).

[3] T. Salonidis and V. Digalakis: Robust Speech Recognition for Multiple Topological Scenarios of the GSM Mobile Phone System, ICASSP98, pp. 101–104, (1998.5).

[4] G. N. Ramaswamy and P. S. Gopalakrishnan: Compression of Acoustic Features for Speech Recognition in Network Environments, ICASSP98, pp. 977–980, (1998.5).

[5] V. Digalakis, L. Neumeyer and M. Perakakis: Quantization of Cepstral Parameters for Speech Recognition over the World Wide Web, ICASSP98, pp. 989–992, (1998.5).

[6] R. Gruhn, H. Singer and Y. Sagisaka: Scalar Quantization of Cepstral Parameters for Low Bandwidth Client-Server Speech Recognition Systems, 音講論, 3-Q-6 (1999.3).

[7] 中川，チェンチャルーン: 連続出力分布型 HMM の出力確率計算の短縮法，音講論，1-Q-22 (1995.3).

[8] 山田，山本，小坂，小森，大洞: パラメータのスカラ量子化と混合分布 HMM の次元独立演算による高速出力確率計算，信学技報，SP95-22, pp.31-38 (1995.6).

[9] S. Sagayama, S. Takahashi: "On the Use of Scalar Quantization for Fast HMM Computation", ICASSP95, pp.213-216 (1995).

[10] V. Digalakis, L. Neumeyer and M. Perakakis: "Product-Code Vector Quantization of Cepstral Parameters for Speech Recognition over the WWW," ICSLP98, pp. 2411–2414 (1998.12).

[11] 山本，小坂，山田，小森，藤田: 改良型 CMS-PMC，次元独立演算を用いた高速な雑音下音声認識，信学技報，SP96-14, pp.15-22 (1996.6).

[12] 山田 雅章，山本 寛樹，小坂 哲夫，小森 康弘，大洞 恭則: パラメータスカラ量子化と混合分布 HMM の次元独立演算による高速出力確率計算，日本音響学会講演論文集，2-2-16, pp.69-70 (1995.9).

[13] 小森康弘，山田雅章，山本寛樹，大洞恭則: "少数分布 HMM による出力確率推定に基づいた効率的な混合連続分布 HMM 音声認識"，信学技報，SP94-51, pp.31-38 (1994).

[14] 小森康弘，山田雅章，山本寛樹，大洞恭則: "Rough HMM と Detail HMM を用いた連続 HMM 出力確率計算の高速化"，日本音響学会講演論文集，1-Q-20, pp.135-136 (1995-3).

[15] F. K. Soong, E. F. Hung: "A Tree-Trellis Based Fast Search for Finding the N Best Sentence Hypotheses in Continuous Speech Recognition ", ICASSP91, (1991.5).

[16] 小森，山田，山本，小坂，大洞: "Top-Down Clustering に基づく効率的な Shared-State Triphone HMM," 信学技報，SP95-21, pp.23–30 (1995.6).