様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成　25 年　5 月 13 日現在

機関番号：　32503
研究種目：基盤研究（C）
研究期間：2010～2012
課題番号：22500165
研究課題名（和文）　音声合成技術を用いた音声符号化に関する研究

研究課題名（英文）　A Study on Speech Coding by Speech Synthesis

研究代表者
木幡　稔（KOHATA　MINORU）
千葉工業大学・情報科学部・教授
研究者番号：30186720

研究成果の概要（和文）：本研究の目的は，人間の音声模倣発話を基にした，極低ビットでの音声符号化システムを構築することである．研究の結果，音声合成器の素片データを利用して入力音声の模倣により数 100 bit/s で音声を合成することに成功した．また，合成音声の明瞭度の低下を防ぐために 3 つの話者適応方式を提案し，それらの効果を確認した．
研究成果の概要（英文）：The goal of this study is to build a speech coding system which is based upon human's voice parroting. As a result, input speech can be coded at several hundred bit per second by parroting with segments for speech synthesis system. Meanwhile, in order to improve the degradation of intelligibility of the coded speech, three methods for speaker adaptation were proposed and evaluated.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	900,000	270,000	1,170,000
2011 年度	500,000	150,000	650,000
2012 年度	500,000	150,000	650,000
年度			
年度			
総　計	1,900,000	570,000	2,470,000

研究分野：総合領域
科研費の分科・細目：情報学，知覚情報処理・知能ロボティクス
キーワード：音声情報処理

１．研究開始当初の背景

(1)携帯電話などにおいて利用されている現在の音声符号化方式においては，マイクロホンから入力された音声を，低い量子化歪を維持しながら，音声信号の統計的性質を利用することにより，できるだけ少ない情報量で符号化している．
(2)学術的な背景としては，人間の脳における言語に基づく音声生成や音声認識の問題は依然として未解明な問題として残っており，今後，音声符号化の研究分野においてもこれらの問題の解明に貢献することが必要であると考えられる．
(3)こうした観点から，当初，研究代表者は，音声符号化パラメータの情報圧縮に焦点を絞って研究を行ってきた．具体的には，符号化パラメータの時間的，空間的性質を分析し，これをリカレント・ニューラルネットワークで符号化する方式[1]や，時間冗長性を圧縮するために Lempel-Ziv 符号化を応用したセグメント量子化法などを提案してきた．
(3)これまでに実用化されている音声符号化方式は非可逆量子化を用いており，劣化を犠牲にして情報圧縮を行っている．この際に生じる量子化歪はスペクトル歪や SN 比などにより，客観的に測定可能であり，これを与え

られたビットレートのもとで最小化することにより音質を維持している．
(4)しかし，最終的に音質を決定するのは主観的な評価値であり，人間が音声として，その意味内容を把握するために満たさなければならない条件は，必ずしも客観的に可測な歪が小さいことではないと言える．
(5)例えば，他人の発声した音声を聞いて自分がその音声と同じ音声を発声する場合を想定すれば明らかなように，両者の音声の客観的特性は少なからず異なるにもかかわらず，大抵の人間は両者に大きな差があるとは認知しない．この場合，人間は音をカテゴリ化して認知し，両者の音が同じカテゴリに属すると認知するために，差異としての歪を感知しないと考えられる．
(6)このような，人間の聴覚における音のカテゴリ分類は，人間が音声を認識・合成する上で重要な働きをしている．しかしながら，そのメカニズムに関しては研究開始当初は解明されていなかった．そこで，この人間における音声模倣の原理を基に動作する音声符号化システムの構築を目的とする本研究の構想に至った．

２．研究の目的

(1) 本研究の目的は，人間が発声する音声を模擬するシステムによる音声符号化の実現を試みることである．
(2)入力音声に対して，聴覚的な特徴抽出を行い，音声合成システムによる合成音声が，入力音声の特徴にできるだけ一致するように合成音声を制御することで得られる音声合成のための情報を伝送することで，数100bit/s での音声符号化を可能とする．
(3)この方式は音声認識ボコーダのように，音素単位の音声認識を介するのではなく，入力音声の特徴を「言いまね」と同様の処理により，類似した音声単位（セグメント）に分解し，合成する原理に基づく方式である．

３．研究の方法

(1)まず，音声模倣を行うための基本システムとして CbS 法(Coding by Synthesis)を提案した．CbS 法では，音素認識誤りによる極端な音質劣化を回避でき，また入力音声の韻律情報に類似した素片を選択すればこれを復元できることが期待できる．

CbS 法の原理を図 1 に示す．まず入力音声をセグメンテーション処理により素片に分解する．この素片は音素に必ずしも対応するものではなく，有声/無声/無音の判別と，スペクトルの変化などの音響的特徴により分解されたものである．この入力素片からメルケプストラム，F0（有声/無声情報を含む），継続時間長を抽出する．これらを特徴パラメータとして，コーパス内の全素片とパターンマッチングを行うことにより最適な素片を選択し，接続することで符号化音声を合成する．

コーパス内の素片総数を M とした場合のビットレート は以下の式で得られる．

$$B(M)=\frac{s\cdot\lfloor\log_2 M\rfloor}{\sum_{k=1}^{s}D_k}\quad(\text{bit/s})\qquad(1)$$

ここで，s は入力素片総数， D_k は各入力素片の継続時間長である．パターンマッチングを行う際の距離を求める関数を(2)式に示す．

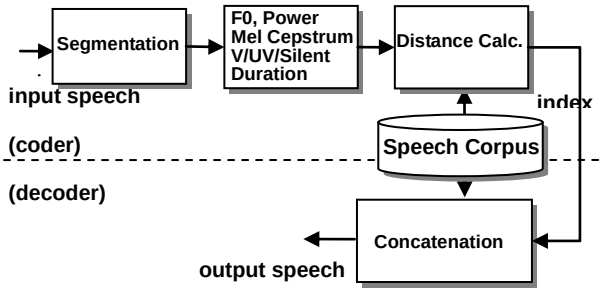

図 1　CbS(Coding by Synthesis)法

$$d_m=\sum_{k=1}^{s}\frac{\left|D_k-\hat{D}_k^m\right|}{D_k}+\sum_{k=1}^{s}\frac{\left|P_k-\hat{P}_k^m\right|}{\sigma^P}+\sum_{k=1}^{s}\frac{\sqrt{\sum_{j=1}^{24}\left(C_{k,j}-\hat{C}_{k,j}^m\right)^2}}{\sigma^C}\qquad(2)$$

ここで，D_k，P_k 及び $C_{k,j}$ はそれぞれ入力音声の継続時間長，F0，及び j 次メルケプストラム係数を表す．^ が付いたものはコーパス内素片のパラメータを示す．また，m はコーパス内の素片番号，σ^P，σ^C はそれぞれ F0 とメルケプストラムの距離を正規化するための標準偏差である．
(2)パターンマッチングアルゴリズムとして One-Pass DP を用いた．One-Pass DP により，入力音声の各時刻（フレーム）において各グリッドまでの累積最小距離を求める過程でセグメント間の遷移時刻も記録することで，終点からのバックトラックによりセグメンテーションと最適素片の選択を同時に行なうことができる．入力音声とコーパス内素片とのマッチングには(2)式の距離尺度を用いた．
(3)CbS 法においては，未知の入力音声と音声合成用のコーパスとのパタンマッチングによって音声符号化を行うため，話者間の適応を行わないと明瞭度が低下することが分かった．そのため，いくつかの話者適応法を試みた．

①アフィン変換を用いる話者適応法
One-Pass DP 後にアフィン変換を用い，入力音声をコーパス内素片に適応させる処理を反復適用する以下のアルゴリズムを提案した．
i)One-Pass DP により入力音声のセグメンテーションを行う．
ii)DP の歪を測定し値が収束すれば終了．
iii)入力音声素片とコーパス内素片の対応により(3)式のメルケプストラム用のアフィン変換係数 A，b をニューラルネットワークにより求める．

$$c_c = A \cdot c_t + b \qquad (3)$$

ここで c_t，c_c はそれぞれ入力素片，コーパス内素片のメルケプストラムを表す．
iv)入力音声のメルケプストラムに(3)式によるアフィン変換を行い，i)に戻る．
②GMM を用いた話者適応
GMM は、複数のガウス分布を加え合わせる事で 1 つのベクトルを表現する．入力話者とコーパス音声のパラレルコーパスを用い，GMM 適応のためのパラメータを学習により求め Stylianou らにより提案された方式により入力話者の GMM をコーパス内音声の素片に適応させる方法である．
③母音マッチング法による話者適応
母音マッチング法による話者適応では入力音声とコーパス音声のそれぞれの母音メルケプストラムの対応が得られていることを前提としている．そして，その中間に存在する入力音声の特徴ベクトルを，最も近い母音ベクトルから順に利用し，各母音ベクトル直交化しながら線形和により近似していく方法である．この線形和の係数を保持しながら，コーパス空間でコーパス母音ベクトルを用いることで入力話者ベクトルをコーパス話者ベクトルに話者適応する方法である．

４．研究成果
(1)CbS 法の基本性能の評価結果
CbS 法を用いて，4 つの音声に対して符号化を行った場合のビットレートを(1)式により求めたものを表 1 に示す．4 つの音声としては女性話者 3 名による 4 つの文章を用いた．その中の 1 つは，コーパスを作成した話者のものを利用した．他の 2 名は，ATR 音声データベース B セットのものである．コーパス内の総素片数に最も近い 2 の冪は 2^{19} なので，ビットレートの算出の都合上，2^{19} にコーパスを制限して実験した．表 1 に示すように，大規模コーパス中の音声の波形素片を用いて音声符号化を行う CbS 法法では，200bit/s 程度のビットレートを達成できることが確認できた．しかしながら，前述のように合成音声の明瞭度が低下するため以下に述べる話者適応実験を行った．

表 1　CbS 法のビットレート

音声	種類	素片数	ビットレート (bit/s)
1	コーパス内音声 (F009)	48	177.8
2	ATR B set (FKNSDA01)	46	210.1
3	ATR B set (FKNSDA05)	45	176.3
4	ATR B set (FKSSDB05)	70	235.1

(2) 上述した話者適応法の 1 つである One-PassDP とアフィン変換の反復適用による話者適応法の効果を確認するために、以下の 2 種類の方法で合成した音声の対比較試験を行った．
・One-PassDP とアフィン変換の反復適用
・One-PassDP のみ適用（話者適応なし）
被験者に上記 2 つの方法で CbS 法を用いて合成された音声を提示し，いずれの明瞭度が高いかを 2 者から強制選択させ，選択率をプリファレンススコアとして測定した．
その結果，表 2 に示すようにに示すようにアフィン変換を反復適用した方がスコアが高く，明瞭度が向上することが確認された．Z スコアによる検定を行った結果，信頼度 95%でアフィン変換により音質が向上したことが確認された．
表 2　アフィン変換による話者適応効果

	話者適応なし	話者適応あり
プリファレンススコア（平均値）	0.648	0.352
標準偏差	0.153	
Z スコア	2.747	

(3)GMM による話者適応では，まず予備実験を行い，GMM の混合数と，話者適応に必要なパラメータの学習に必要な文章数について検討を行った．その結果，混合数については 8，学習用文章数は 1 話者あたり 50 文章程度あれば歪が収束することが確認されたのでこの条件にて実験を行った．先行して行われた話者適応方式であるアフィン変換による話者適応を用いて合成した音声と，GMM を用いた話者適応を適用した合成音声の間で，対比較による明瞭度の評価試験を行った．これは，アフィン変換を用いた話者適応法と同様に，2 者の音声のどちらの明瞭度が高いかを強制選択させ，その選択率をプリファレンススコアとして評価する方法である．評価は話者の性別ごとに行ったが，特に男声において明瞭度の改善効果が高いことが分かった．コーパス音声が女声であるため，性差がある場合に GMM による話者適応法が有効であると言える．

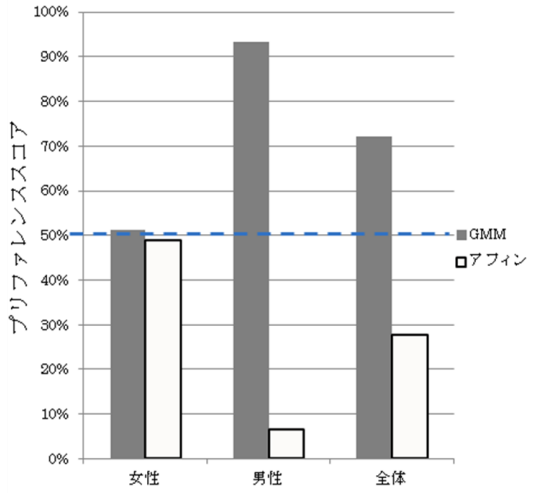

図 2　GMM による話者適応法の評価

(3)最後に母音マッチング法による話者適応法の評価について述べる．実験に使用した音声は ATR B セット 10 名である．ておく．内訳は男声音声(MHO,MHT,MMY,MSH,MTK,MYI)，女声音声(FKN,FKS,FTK,FYM)であり，これらの母音の特徴ベクトルは，予めラベルデータにより抽出されたものを平均化することにより求め話者適応による明瞭度の改善効果の評価はこれまでの提案方式との比較により行った．被験者は 14 名である．まず，GMM による話者適応法，話者適応なしの音声との比較結果を図 3〜5 に示す．それぞれ，話者全体，男声，女声に対する結果である．

これらの結果より，GMM による話者適応法と比較して，男声，女声ともに母音マッチング法による話者適応法が，明瞭度の高い音声が得られていることが分かる．しかし，この結果では GMM による話者適法は話者適応を行わない音声よりも明瞭度が低下する結果が得られた．先行するアフィン変換との比較評価においては GMM はアフィン変換より有効，また，アフィン変換は話者適応なしより有効，という結果が得られているが，この結果はこれらの結果に反するものとなった．原因としては入力話者や文章の違い，また母音マッチング法を加えた 3 者での評価により被験者の評価基準が影響を受けたことなどが推測される．性別では，やはり母音マッチング法も男声に対しては明瞭度の改善効果が高いが，女声に対しては話者適応なしよりも劣化している結果となった．このことから，話者適応は性別や話者により ON/OFF を切り替えて適用することが望ましいと考えられる．また，CbS 法における明瞭度改善のためには話者適応以外の手法を導入する必要があると考えられ，今後検討を行なっていく予定である．

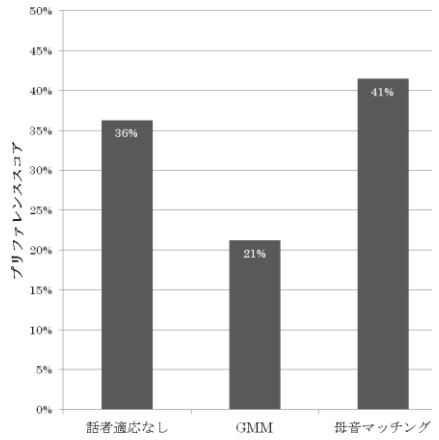

図 3　母音マッチング法による話者適応法の評価（全話者）

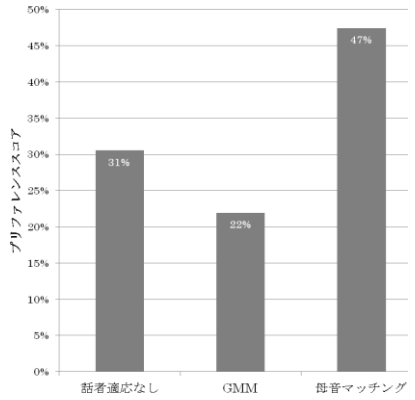

図 4　母音マッチング法による話者適応法の評価（男声）

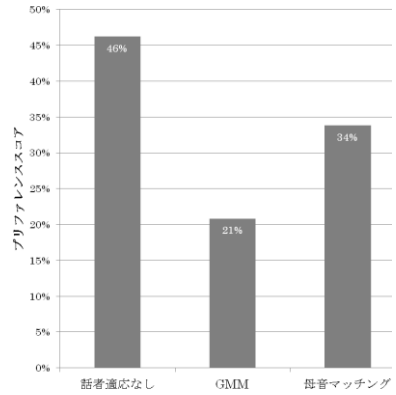

図 5　母音マッチング法による話者適応法の評価（女声）

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）

〔学会発表〕（計　9 件）

①竹田　翔, 木幡　稔, 日本音響学会秋季講論, 3-2-11, pp. 347-348, (2012).

②木幡　稔, 立野　誠, 日本音響学会春季講論, 1-Q-11, pp. 847-848, (2012).

③木幡 稔,日本音響学会春季講論,1-Q-12, pp. 849-851,(2012).
④.木幡　稔, 電子情報通信学会技術研究報告 , EMM2011-66, pp. 13-18, (2012).
⑤川上 祐也, 木幡 稔, 日本音響学会秋季講論,1-8-2, pp. 231-232,(2011).
⑥垰 雅文, 木幡 稔, 日本音響学会秋季講論,3-Q-4, pp. 357-358,(2011).
⑦ ″ Bit Rate Reduction of MELP Coder by Using Lempel-Ziv Segment Quantization″, M. Kohata, M. Suzuki, A. Ito, S. Makino, Procs. of ICASSP2011 , pp.5240-5243, (2011).
⑧平山 雄大, 木幡 稔, 日本音響学会秋季講論,2-Q-1, pp. 297-298,(2010).
⑨木幡 稔, 平山 雄大, 日本音響学会秋季講論,3-7-11, pp. 379-380,(2010).

〔雑誌論文〕(計 1 件)
①木幡　稔, 鈴木基之, 伊藤彰則, 牧野正三, 電子情報通信学会論文誌, J93D, No.5, pp.588-597, (2010).

〔産業財産権〕
○出願状況（計 2 件）

名称：角度伝達装置
発明者：木幡　稔
権利者：千葉工業大学
種類：特許
番号：、2012-266669
出願年月日：2012 年 12 月 5 日
国内外の別：国内

名称：音声システム
発明者：木幡　稔
権利者：千葉工業大学
種類：特許
番号：2011-241599
出願年月日：2011 年 11 月 2 日
国内外の別：国内

６．研究組織
(1)研究代表者
木幡　稔（KOHATA　MINORU ）
千葉工業大学・情報科学部情報ネットワーク学科・教授
研究者番号：30186720