ベクトル量子化による小規模規則音声合成器の開発

清水　忠昭
鳥取大学工学部知能情報工学科

Development of Small-scale Speech Synthesizer based on Vector Qantization

Tadaaki SHIMIZU
Department of Information and Knowledge Engineering, Faculty of Engineering
Tottori University, Tottori, 680-8552 Japan
E-mail: tadaaki@ike.tottori-u.ac.jp

Abstract: A new scheme of speech synthesis by rule was presented which could be implemented in ROM less than 4M bytes. The features of our scheme are consist in the use of the vector quantization of LSP parameters for VCV instance. We proposed two synthesis unit selection methods, 1) selection method by using phonemic environmental resemblance score (PER method), and 2) selection method by searching minimal connective distortion path (MLD method), for small scale speech synthesis system. PER method requires phonemic environmental information for each VCV instance in a VCV unit dictionary. This paper investigated experimentally to what extent we can reduce the phonemic environmental information with keeping high quality of synthesized speech. We verified that two phonemes frontward and one phoneme rearward range to a current VCV instance is enough to synthesize similar quality of speech as five phonemes frontward and five phonemes rearward. This result gives an experimental basis on minimizing a size of VCV unit dictionary.

Key Words: Speech Synthesis, LSP analysis, Vector Qantization, phonemic environment

１．はじめに

任意の単語や文音声を合成する規則音声合成のからくりは至って単純である．一言で言うと，人間が発話した音声を適当な短い素片に切り刻んだ形でデーターベースに登録しておき，その中から単語や文を作るために必要な素片を選び出して繋ぐと合成音声が出来上がるというものだ．アイデアは単純であるが，実際に試してみると人間が話すように自然で高品質な合成音声を作り出すのはなかなか難しい．

日本語の音声規則合成の研究の初期段階から採用された方式は，C(子音,Consonant) と V(母音,Vowel) を組み合わせた CV 単位に基づく合成方式 [1] 〜 [3] である．CV 単位とは，「か」や「た」など仮名一文字で書き表せる音節単位であり，日本語の基本単位としては自然なものである．

確かに，CV 単位で 50 音表の全ての音を一通りデーターベース (合成単位辞書) に登録しておけば，どんな日本語も合成できる．しかし，この方法では自然な音声を合成することは大変難しい．人間の自然な発声では，文字の上では同じ CV であっても単語中や文中で微妙に変化して発声される．同じ「か」や「な」でも，文章中のどこにあるかで微妙に音としての性質が変わるのである．この現象を調音結合と呼んでいる．調音結合の影響を考慮せずに，音声の素片を繋いだのでは，良質な合成音声は得られない．

調音結合の影響を音声合成手法に取り入れる方法の一つとして，音声合成に用いる素片を長くする方法がある．実際に，CV-VC 単位を用いる方式 [4] や，VCV 単位 [5] や CVC 単位 [6] を用いる方式へと，より長い合成単位を用いる方式が提案されてきた．さらに，CVCV 単位にまで選択範囲を広げて合成単位のセットを検討する研究 [7] や，様々な合成単位を選択的に用いる合成方式の研究 [8] も行われている．このように合成音声の品質向上のために合成単位を長くする方法では，音素 (C や V) の組み合わせにより合成単位の種類が爆発的に増加し，ひいては合成単位辞書に登録しなければならない素片数も膨

大になるという欠点がある．

一方，合成単位を VCV などの比較的短い単位とし，同一の合成単位に対して，その前後の音韻の並び (音韻環境と呼ぶ) が異なる複数の素片を保持することで，合成音声を高品質化する方法もとられている．[9] この方法では，合成単位の種類の数は押さえられるが，音韻環境の異なる多数の素片を必要とするという点で，やはり合成単位辞書は大きなものになる．

これら合成音声の高品質化を目ざす研究において，合成音声を生成するために合成システムが保持すべき合成単位辞書の記憶容量は増加してきた．特に，近年盛んに研究されている波形重畳方式では，合成単位を音声の分析パラメータではなく，時間波形の形で記憶するため合成単位辞書の記憶容量は非常に大きい．小山らによる VCV を基本単位とする波形規則合成方式 [9] では，合成単位辞書に 60M バイトの記憶容量を要することが報告されている．

我々は，小規模な応用に対して高品質な合成音声を与えるために，LSP ベクトル VCV 規則音声合成方式を提案した．本方式では，合成単位辞書の記録方法にベクトル量子化を導入することで，様々な音韻環境から採取した多くの VCV 素片を少ない記憶容量で記憶できる．これにより，小規模な音声合成システムでも合成音声の品質を向上できる可能性が高いのが特長である．

本稿では，LSP ベクトル VCV 規則音声合成方式について発表した論文の中の 4 編 [10] 〜 [13] の内容を，ダイジェストで紹介するとともに，チームとして一緒に頑張ってくれた学生諸君の苦労なども織り込んで紹介したい．

2．LSP ベクトル VCV 規則音声合成方式の概要

本論文で提案する LSP ベクトル VCV 規則音声合成方式は，音声合成の基本単位である VCV 素片の記録にベクトル量子化された LSP パラメータを用いることにより合成単位辞書 (VCV unit dictionary) の記憶容量を小さく抑える手法である．図 1 に示すように，LSP パラメータの系列として表現された VCV 素片は，コードブックを用いてベクトル量子化することにより，代表ベクトルを表すインデックスの系列として符号化される．従って，本方式では VCV 素片はベクトル量子化の代表ベクトルのインデックスの系列として合成単位辞書に収録される．

本方式による音声合成システムのブロック図を図 2 に示す．合成単位辞書には同一の VCV 合成単位に属する VCV 素片が多数収録されており，同一の文章を作成する場合でも，可能な VCV 素片の組み合わせが多数存在する．高品質な合成音声を得るためには，適切な VCV 素片を選択し接続することが必要である．我々は，素片選択の方法として，音韻環境を考慮して素片選択を行う PER 選択法と，素片の接続歪みを最小化する MLD 選択法の 2 つの手法を提案し，それらの方法ついて性能の評価を行った．

また，MLD 選択法の改良として，ベクトル量子化の特徴を生かし，コードブックの代表ベクトル間の距離を予め計算して作成した距離テーブル (distance table) を参照することで VCV 素片の接続歪みの計算を高速化する距離テーブル参照法 (Distance Table Look-up Method: DTL 選択法) を提案した．

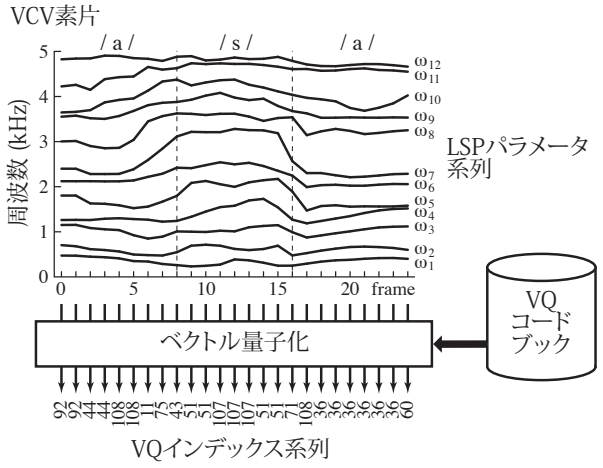


図 1　VCV 素片のベクトル量子化

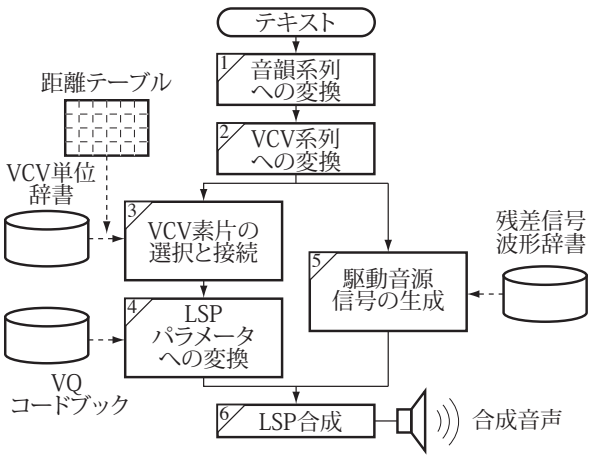


図 2　LSP ベクトル VCV 音声合成法のブロック図

3．VCV 合成単位

3．1　音韻の取り扱いと VCV 素片

日本語における音韻の種類や数には，諸説がある．本研究では，音韻の種類と表記に関しては原則的に斎藤 [14] の分類に従った．外来語音節 (ウィ，ヴァ等) を取り扱わず，半母音 S(/j/,/w/) を及び，拗音節を作る CS(/kj/, /sj/ 等) を一つの子音として取り扱った．このため子音数は 26 種である．また，標準的な日本語 5 母音の他，はつ音 /N/ を母音と同様に扱ったため，母音は 6 種類となる．

これらの音韻を組み合わせた合成単位としては，母音で子音を挟む形の VCV 型 570 種類，子音を挟まない VV 型 35 種類，語頭用の #CV 型 95 種類と #V 型 5 種類，語尾用の V# 型 6 種類がある．以後，これらの型の合成単位を総称して VCV 合成単位 (VCV unit) と呼び，それぞれの VCV 合成単位の素片データを VCV 素片 (VCV instance) と呼ぶ．

3．2　合成単位辞書と合成単位の収集

研究の開始にあたって最初の課題は音声資料の収集である．開発グループのゼミでは，「どのような音声から素片を収集するか」という問題が，まず議論された．音声の分野では，このような実験のために音素がバランスよく含まれた「音素バランス文」が作られ，その音声データを入手することもできる．しかし，小規模で高音質の音声合成を実現する際には，音素がバランスしているよりも自然な文章上で生じる偏りがあった方が有利であるとも考えられる．両方を試してみれば良いのだが，研究室の人的資源はそれをゆるさなかった．議論の末，「音素バランス文」ではなく，ラジオ・ニュースから自前でデータを収集することになった．

安価に高品質の音声データを得るために，7 日分の NHK の FM ラジオ・ニュースを合成単位の収集に用いた．録音したラジオ・ニュースの 1 日分から，約 10 分間の同一の男性アナウンサの発話部分だけを切り出し，合計 70 分の音声データを得た．合成単位辞書に収録する VCV 素片は，音声データに視察で音韻マーキングした資料を用いて，母音部の中間点で切り出す方法で自動的に生成した．

この音韻マーキング作業は，音声を部分的に再生して聴覚で確かめながら，ディスプレ上で波形を確認し，波形上に /a/ や /t/ といった音韻のタグを付けていく作業である．大変細かく骨のおれる仕事である上，マーキング箇所は 70 分間の資料で 40,000 点を越えた．開発チームのメンバーは，目を真っ赤にしながら，ヘッドフォンをかけてディスプレイを睨む日々を送った．この大変な作業に不平も言わずに参加してくれた当時の学生諸君に心から感謝したい．

表 1 に，音声資料から採取した VCV 合成単位の種類数と，各々の VCV 合成単位に属する VCV 素片の数の平均値を示す．音声資料に含まれる VCV 素片が全ての VCV 合成単位を網羅していないため，表 1 の VCV 合成単位の種類数は，3．1 節で述べた VCV 合成単位の種類数に達していない．音声合成時には，合成単位辞書に収録されていない VCV 合成単位は，子音と後続母音の一致する他の VCV 合成単位から先行母音の部分を除いて作成される CV 合成単位によって代用する．この際，先行母音部は補間によって作成する．

表 1　音声資料の長さと採取された VCV 単位の種類数および VCV 素片の数

音声資料の長さ	VCV 型		VV 型		#CV 型		#V 型		VCV 型		総種類	素片総数
	種類	平均個数	種類	平均個数	種類	平均個数	種類	平均個数	種類	平均個数		
(分)	(種)	(個)	(種)	(個)	(種)	(個)	(種)	(個)	(種)	(個)	(種)	(個)
10	341	7.7	29	19.3	52	6.5	5	20.8	6	73.5	433	4,050
20	400	12.7	31	36.7	64	9.7	5	35.8	6	134.0	506	7,759
30	429	17.2	35	48.9	65	11.0	5	41.6	6	152.8	540	10,926
40	446	21.4	35	61.4	67	11.6	5	48.2	6	169.1	559	13,713
50	455	27.4	35	81.2	73	15.6	5	68.6	6	246.5	574	18,271
60	466	32.7	35	98.2	76	20.2	5	88.6	6	328.5	588	22,622
70	470	38.4	35	177.1	78	21.5	5	101.2	6	362.7	594	26,517
可能な種類	570		35		95		5		6		711	

注 1）　# は無音を表しており，#CV 型と #V 型は発話開始点に、V# 型は発話終了点に用いる．
注 2）　音声のサンプリングは，標本化周波数：11.025kHz，量子化数：16 ビット
LSP 分析は，フレーム長：256 点，インターバル：64 点，次数：12 次

4．VCV 素片選択の 2 つの方法

合成単位辞書から VCV 素片を選択する方法として，素片選択基準が異なる 2 つの手法を提案した．第一の VCV 素片選択法は，VCV 素片を収集した際の音韻環境と合成する文章中における VCV 素片の音韻環境の類似度を素片選択の基準にする手法である．本研究では簡便性を考慮して，図 3 に示すように VCV 素片の前後 5 つずつの音韻について，式 (1) に示す音韻環境類似度 (Phonemic Environmental Resemblance Score: PER スコア) を計算し，音韻環境の類似度を評価する．

$$PER = \frac{1}{2}\sum_{i=1}^{5}\frac{1}{3^{i-1}}\left(f(i) + r(i)\right) \qquad (1)$$

ここで，$f(i)$ は VCV 素片に先行する i 番目の音韻について，VCV 素片を収集した際の音韻と合成する文章中での音韻の一致度を表す音韻得点である．$f(i)$ には，音韻が一致すれば 2 点，母音，摩擦子音，破裂子音等の音韻種別が一致すれば 1 点を与え，どちらも一致しない場合には 0 点を与える．$r(i)$ は VCV 素片の後続する i 番目の音韻についての $f(i)$ と同様な得点である．音声合成時には，$f(i)$ と $r(i)$ の重み付き和として式 (1) で定義した PER スコアが最大となる VCV 素片を選択する．この VCV 素片選択法を PER 選択法と呼ぶ．

第二の VCV 素片選択法は，VCV 素片の接続部で生じる接続歪みを最小化する手法である．図 4 に示すように，2 つの VCV 素片の接続部において先行する VCV 素片の最終フレームの LSP パラメータを $\boldsymbol{\omega}^f = (\omega^f_1, \omega^f_2, \ldots, \omega^f_p)$，後続 VCV 素片の先頭フレームの LSP パラメータを $\boldsymbol{\omega}^r = (\omega^r_1, \omega^r_2, \ldots, \omega^r_p)$ とする．このとき，2 つの素片の接続点での歪みを，式 (2) に示す LSP パラメータの距離 (LSP Distance) によって評価する．

$$d(\boldsymbol{\omega}^f, \boldsymbol{\omega}^r) = \sqrt{\sum_{i=1}^{p}\left(\omega^f_i + \omega^r_i\right)^2} \qquad (2)$$

接続歪みを最小化する VCV 素片選択は，図 5 に示すように VCV 素片の接続可能な経路に対して式 (2) で計算される接続歪みをコストとして与えた最小コスト経路探索の問題である．この VCV 素片選択法を，LSP 距離最小化選択法 (minimal LSP distance method: MLD 選択法) と呼ぶ．

PER 選択法は，簡単な得点計算により VCV 素片選択を行えるため処理速度が速いが，VCV 素片に音韻環境の情報を付加する必要があり，合成単位辞書の記憶容量が大きくなる．一方，MLD 選択法は，合成単位辞書に余分な情報を付加する必要がなく記憶容量の点では有利だが，VCV 素片選択時に経路探索を行うため，処理速度は遅くなる．どちらの方法にも一長一短があり，実装するシステムの仕様によって使い分けることが必要となる．

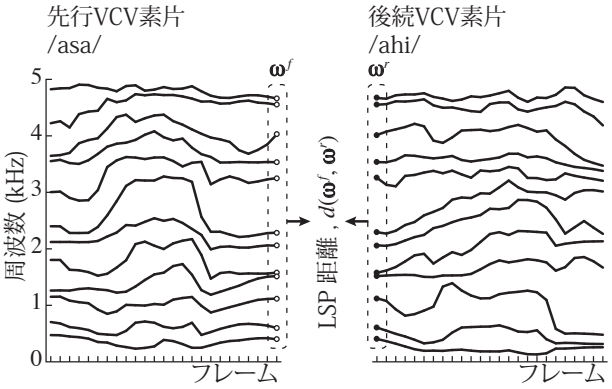


図 4　VCV 素片の接続部における LSP 距離

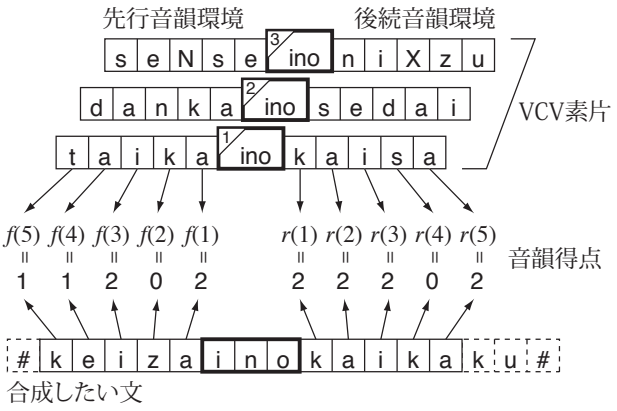


図 3　PER スコアによる音韻環境の得点化

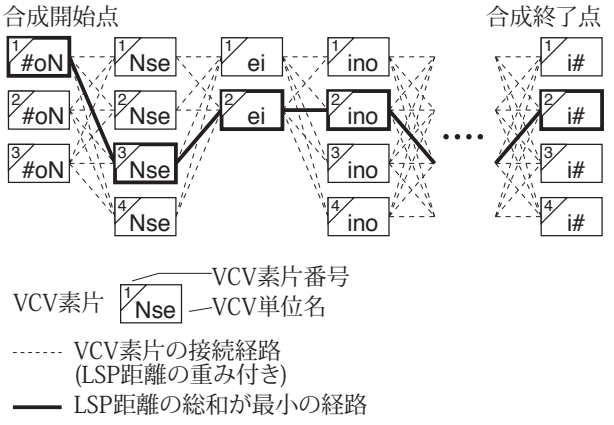


図 5　MLD 選択法による VCV 素片の選択

5．合成単位辞書のサイズと VCV 選択法の比較

5．1　VCV 素片選択実験

適正な合成単位辞書の大きさの検証と VCV 素片選択法の評価のために，PER 選択法と MLD 選択法による VCV 素片選択実験を行った．実験には表 1 に示した 7 種類の大きさの合成単位辞書 (以後,「合成単位辞書 (10)」～「合成単位辞書 (70)」と記載する) を用いた．また，音声合成の対象には見出しを除く新聞記事の本文を用いた．実験に用いた新聞記事の長さは，VCV 合成単位の個数にして 45,269 個分の長さである．

先に述べたように，合成単位辞書に合成に必要な VCV 素片が登録されていない場合，他の VCV 素片から CV 素片を作成して代用する．高品質な合成音声を得るためには，このような代用が起らないことが望ましい．そこで，以下に示す VCV 単位網羅率と VCV 素片置換率を定義して合成単位辞書の適正な大きさについて評価した．

VCV 単位網羅率は，音韻の組み合わせで可能な VCV 合成単位の総数を N，VCV 素片の収集で得られた VCV 合成単位の数を n として，以下のように定義する．

$$\text{VCV 単位網羅率：} \gamma = n / N \qquad (3)$$

VCV 素片置換率は，合成音声中に含まれる VCV 合成単位の総数を M，そのうちで合成の際に CV 素片に置換された VCV 素片の数を m として，以下のように定義する．

$$\text{VCV 素片置換率：} \rho = m / M \qquad (4)$$

本実験では，$N = 711$，$M = 45{,}269$ である．

また，実験での VCV 素片選択結果は，合成音声中の VCV 素片の平均 PER スコアと，合成音声中の VCV 素片の接続部での平均 LSP 距離で評価した．

図 6 に，合成単位辞書の規模と VCV 単位網羅率，VCV 素片置換率の関係を示す．図中の各点は，「合成単位辞書 (10)」～「合成単位辞書 (70)」による VCV 素片選択の結果であり，横軸は合成単位辞書の規模を収録素片数を表している．

VCV 単位網羅率は，最大規模の「合成単位辞書 (70)」でも 83.5%と高くない．一方，VCV 素片置換率は，合成単位辞書の VCV 素片の収録数が 14,000 個以上では，1.7%以下と非常に小さくなった．この結果は，VCV 素片の収集でもれた VCV 合成単位が音声合成時に使用される頻度は非常に小さいことを示しており，合成単位辞書の VCV 素片収録数が 14,000 個以上の場合は VCV 単位網羅率の低さが合成音声の品質低下に与える影響はごく小さいことを示している．

PER スコアによる選択法と MLD 選択法で選択された VCV 素片について，平均 PER スコアと平均 LSP 距離を求めた結果を図 7 に示す．PER スコアによる選択方法を用いた場合，合成単位辞書の規模が大きくなると平均 PER スコアは上昇し，平均 LSP 距離が減少した．LSP 距離最小選択化法を用いた場

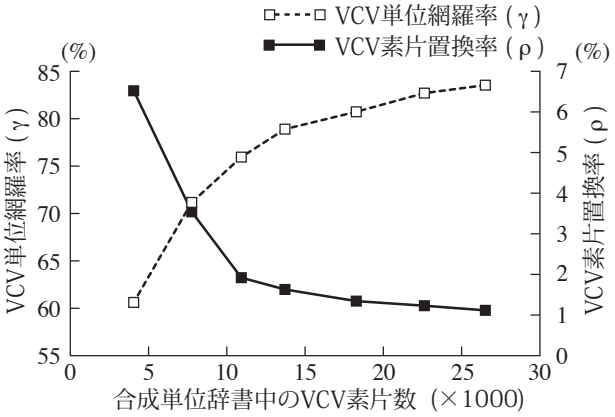


図 6　VCV 単位網羅率と VCV 素片置換率

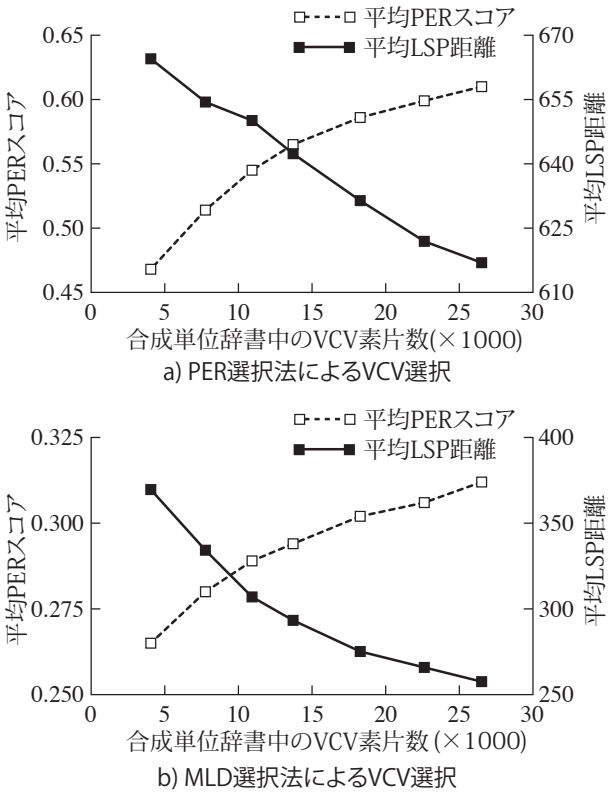


図 7　合成単位辞書の規模と選択結果

合，合成単位辞書の規模が大きくなると平均 LSP 距離は減少し，平均 PER スコアは上昇した．この結果は，PER スコアによる選択方法は VCV の接続歪みを小さく抑える傾向があり，LSP 距離最小選択化法は PER スコアの高い VCV 素片を選択する傾向があることを示している．

5．2　主観評価実験

VCV 素片選択実験の結果から，本手法では合成単位辞書の規模は VCV 素片収録数で 14,000 個以上にすれば良さそうだということが言える．しかし，最終的な合成音声の品質評価は，実際に合成音声を作成して被験者を使って聞き取りによる評価 (主観評価) を行う必要がある．

主観評価実験は，合成単位辞書の大きさをかえて合成した一対の合成音声のうち「どちらの合成音声が聞き取りやすいか」の判定を行う一対比較法によって行った．主観評価の方法として，被験者に合成音声の得点を付けさせる絶対評価法が簡便である．しかし，合成単位辞書の規模を変えたことによる合成音声の微妙な差異まで正確に判定するには，被験者に比較したい合成音声を対にして提示し，どちらが優れているかを判定させる一対比較法の方が信頼できる．

一対比較法の欠点は，資料の組み合わせにより絶対評価法に比べて実験の規模が大きくなり被験者の負担が大きいことである．この面倒な実験には，知能情報工学科の 3 年生以下の学生諸君に有償の被験者として協力を願った．研究室配属されていない学生諸君に被験者をお願いしたのは，音声合成の研究に従事したことがなく，本実験について事前に知識を持っていないことを条件としたためである．また，有償の被験者としたのは，大変な実験に参加してもらうお礼の意味もあるが，実験に真剣な態度で臨んでもらうためでもある．研究室のメンバーの友人などを使って主観評価実験を安くあげると，その結果の信頼性がどうしても低くなってしまう．人間を使った実験はどうしても面倒なものである．いずれにしても，忙しい授業の合間をぬって実験に協力してくれた多くの学生諸君に心から感謝したい．

主観評価実験には，3 秒程度の 4 つの短文について，「合成単位辞書 (10)」から「合成単位辞書 (70)」を用いた 7 種類の合成音声を合成して用いた．被験者には，7 回の練習比較の後，合成単位辞書が異なる 7 種類の合成音声の組み合わせ 21 対について順序の入れ替えを含めて 8 回ずつ 168 回の一対比較を課した．比較対の提示順はランダムとし，練習比較については，それが練習であることを被験者に知らせていない．

上記の一対比較実験を，PER スコアによる選択法と MLD 選択法による合成音声について行った．PER スコアによる選択法について被験者は健康な 20 代の男女 11 名，MLD 選択法について被験者は健康な 20 代の男女 10 名で実験を行った．一対比較実験で得られた判定結果から，Thurstone の比較判定の法則を用いて，「合成単位辞書 (70)」による合成音声を基準として，合成音声の品質尺度値を求めた．

実験の結果として，合成単位辞書の VCV 素片の収録数と合成音声の品質尺度値の関係を図 8 に示す．PER スコアによる選択法を用いた場合，合成単位辞書に収録する VCV 素片を 14,000 個以上に増やしても，合成音声の品質尺度値は向上していない．また，MLD 選択法を用いた場合，合成単位辞書に収録する VCV 素片を 8,000 個以上に増やしても，合成音声の品質尺度値は向上していない．ここに述べた 2 つの VCV 素片選択法を用いる場合，多くても 14,000 個程度の VCV 素片を収録した合成単位辞書を用いて音声合成システムを構築すれば良いといえる．

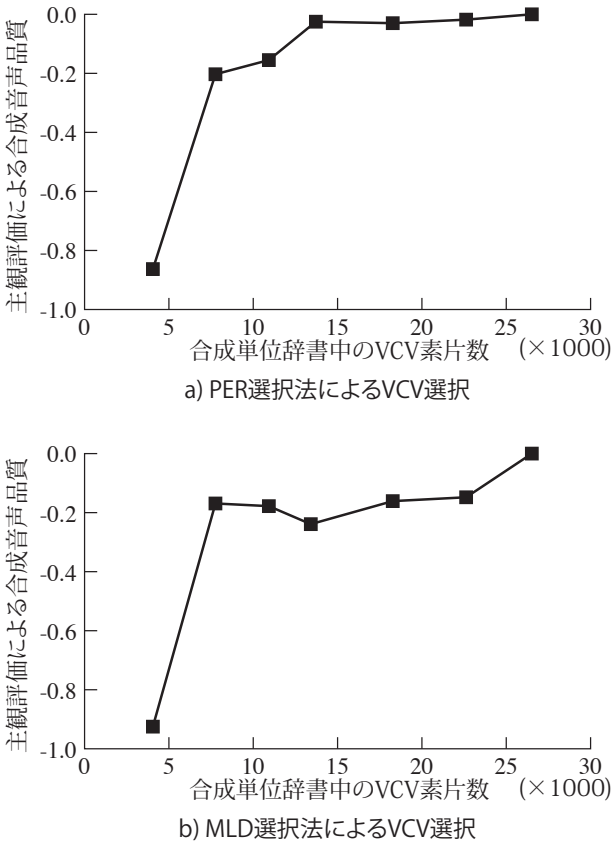


図８　主観評価実験による合成音声の品質評価

PERスコアによる選択法とMLD選択法による合成音声の品質の比較のために，一対比較による主観評価実験を行った．実験には，3秒程度の4つの短文について，「合成単位辞書(70)」を用いて，2つの手法によって合成した合成音声を用いた．被験者には，10回の練習比較の後，20回の一対比較を課した．練習比較についてはそれが練習であることを被験者に知らせていない．被験者には，比較対の「どちらの合成音声が聞き取りやすいか」を「同程度である」という評価を許して判定させた．被験者は健康な20代の男女11名である．一対比較実験でより聞き取りやすいと判定された合成音声に2点，他方に0点を与え，同程度と判定された場合には両方の合成音声に1点ずつを与えて，被験者の判定結果を得点化した．

上記の実験の結果，図9に示すようにPERスコアによる選択法を用いた合成音声の得点率は53.1%，MLD選択法を用いた合成音声は46.9%となった．両者の得点について，両側二項検定を行った結果，有意水準5%で有意な差はみられなかった．このことから，PERスコアによる選択法による合成音声とMLD選択法による合成音声の品質には聴感上の差はないことが判った．

5．3　PER選択法とMLD選択法の関係

VCV素片選択実験と主観評価実験から，PER選択法とMLD選択法には強い関連があることが示唆される．紙数の関係で省略するが，2つの選択法による選択結果の関係については文献[12]でより詳しく調べ，一方の選択法でVCV素片を最適に選択すると，他方の選択基準でも準最適な選択になっていることを詳細に報告している．この結果は，合成音声の品質上は2つの選択法はどちらを使っても良いことを示している．つまり，合成単位辞書の記憶容量と処理速度のトレードオフを考慮して，音声合成システムを実現するプラットホームの事情にあわせてどちらのVCV素片選択法を使うかを決めれば良い．

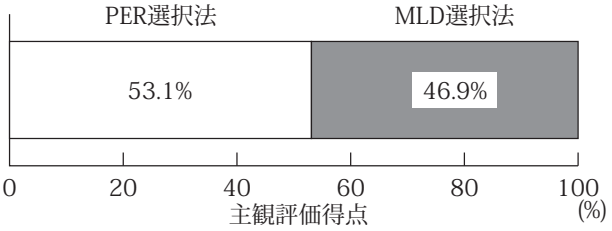


図9　PER選択法とMLD選択法の比較結果

6．VCV素片のベクトル量子化とDTL選択法

LSPベクトルVCV規則音声合成方式では，合成単位辞書に格納するVCV素片に対してベクトル量子化を適用することにより音声合成システムの規模を小さくする．ベクトル量子化のコードブックを作成するために，LBGアルゴリズム[15]と2分割繰り返しアルゴリズムを用いた．これらのアルゴリズムの詳細は文献[11]に詳しく紹介した．さらに，ベクトル量子化の特徴を利用して，MLD選択法の改良手法を提案した．

6．1　ベクトル量子化のコードブックサイズ

ベクトル量子化のコードブックサイズを決定するために，コードブックサイズNを様々に変えた場合のLSPベクトルVCV音声合成法による合成音声の品質を客観評価と主観評価により評価した．コードブックは，合成単位辞書作成に用いた音声資料を含む約110分の同一男性話者の音声資料を用い，$N=2$から2^{14}まで14種類のサイズで作成した．

客観評価では，各サイズのコードブックを用いて音声資料を量子化した際の量子化誤差をLPCケプストラム距離で評価した．図10に示す結果によると，量子化誤差はコードブックサイズNの増加とともに減少している．Nが増加すると量子化誤差の減少率はやや低下するが，最適なNを決める決め手にはならなかった．

主観評価では，異なるサイズのコードブックを用いて本手法により合成した一対の合成音声のうち「どちらの合成音声が聞き取りやすいか」の判定を行なう一対比較による主観評価実験を行った．実験には，3秒程度の4つの短文について，サイズ$N=2,8,32,128,512,2048$のコードブックを用いて作成した

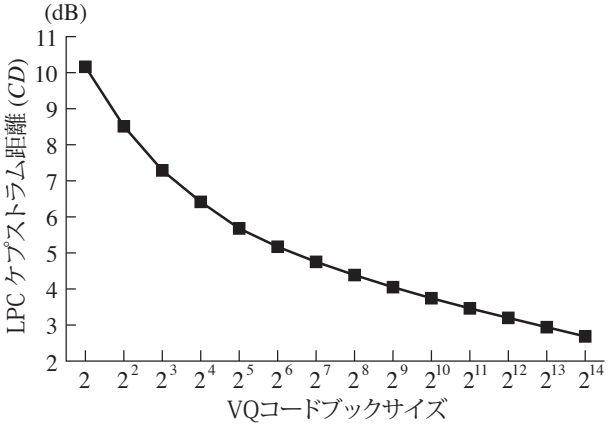


図10　コードブックサイズと量子化歪み

6 種類の合成音声の組み合わせ 15 対について順序の入れ替えを含めて 8 回ずつ 120 回の一対比較を課した．比較対の提示順はランダムとした．被験者は健康な 20 代の男女 10 名で実験を行った．一対比較実験で得られた判定結 から，Thurstone の比較判定の法則でケース V を適用して，合成音声の品質尺度値を求めた．

実験の結果得られたコードブックサイズと合成音声の品質尺度値の関係を図 11 によれば，コードブックのサイズ N が 32 ないし 128 以上では，品質尺度値に差がなく，十分な合成音声品質が得られることが判る．ベクトル 子化のコードブックをこのように小さくできるのは，音声提供話者を一人に絞ったことが大きく関与しているものと考えられる．

合成音声品質の安定性を考慮して，コードブックサイズを大きめに N =128 とし，本手法の合成単 辞書を作成した場合の記憶容 は以下のように計算できる．実験に用いた合成単 辞書中の VCV 素片は平均で 20 フレーム程度の長さがあり，VCV 素片の総数は 14,000 個である．コードブックサイズを N =128 とした場合，代表ベクトルのインデックスを 7bit で記録できる．従って，合成単 辞書の大きさは，7bit × 20 × 14,000 ≒ 256K バイト程度と非常に小さなものにできる．コードブックにおいて，12 次のベクトルの各要素に 10bit の割り当てを行なうと，10bit × 12 × 128 ≒ 15K バイト程度の大きさとなる．また，残差波形辞書には，残差波形を各母音 6 種類と音節 に分類した子音 95 種類について，平均 2K バイトで記録した．このため，残差波形辞書は，2K バイト× (6 + 95) ≒ 200K バイト程度大きさとなる．従って，合成単 辞書とコードブック，残差波形辞書を合わせても 500K バイト以下で記録できる．合成単位辞書のサイズに関しては，研究開始時の目標である 1 ～ 4 M バイトという目標に対し十分な結果が得られた．

6．2　DTL 選択法

先に述べたように，MLD 選択法は VCV 素片の接続部における LSP 距離を接続歪みの指標として，ダイナミック・プログラミング (DP) の手法により VCV 素片選択を行なう．MLD 選択法では，LSP 距離の計算を多数回行なう必要があり計算時間がかかることが欠点である．この欠点を改善するため，ベクトル量子化の特徴を利用して，VCV 素片の選択を高速化する距離テーブル参照法 (Distance Table Look-up Method: DTL 選択法) を提案した．

DTL 選択法では，ベクトル 子化のコードブックの代表ベクトル間の LSP 距離を予め計算し，距離テーブルとして記録しておく．図 12 に示すように，VCV 素片の選択時には，先行 VCV 素片の最終フレームのインデックスと後続 VCV 素片の先頭フレームのインデックスによって距離テーブルを参照し，接続部における LSP 距離を得ることができる．これにより計算量の多い距離計算を避けることができ，VCV 素片選択を高速化できる．

また，DTL 選択法において，記憶容量の削減のために距離テーブルに登録する距離情報を制限することができる．距離テーブルに登録された情報のうち，VCV 素片選択で特に重要な役割りを果たすのは距離の値が小さい部分である．これを利用して，各代表ベクトルに対する距離情報のうち，その値が小さいものだけを保持し，それ以外は推定値を用いて素片選択を行う実験を行った．

詳しくは，文献 [11] を参照して頂くことにして結果だけ紹介すると，距離順位で 8 位までの情報のみを使用した場合でも，十分な精度で VCV 素片選択が可能であることを示すことができた．このとき，距離テーブルの大きさは，全ての情報を保持する場合の 1/4 とすることができる．

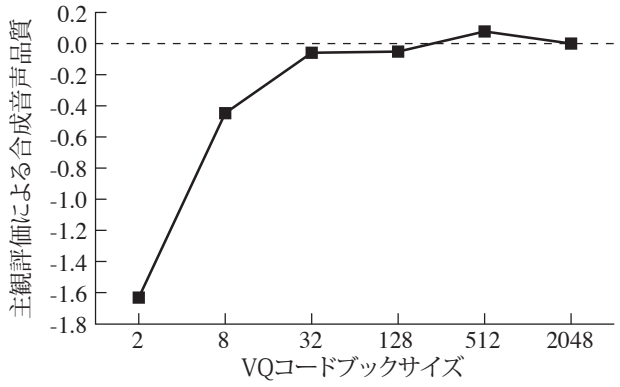


図 11　コードブックサイズと合成音声品質の関係

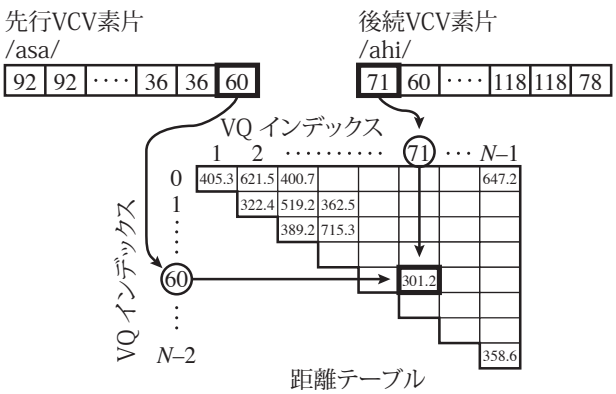


図 12　DTL 選択法の概念図

7．VCV 素片選択時に考慮すべき音韻環境の長さ

7．1　部分 PER スコア

我々が提案した PER 選択法は，前後 5 音韻の長さの音韻環境を考慮して PER スコアを計算することにより音韻環境の適合度を評価して VCV 素片を選択する素片選択法である．PER 選択法の有効性は 5 章で示したが，PER 選択法において考慮する音韻環境の範囲をある程度狭めても，合成音声の品質劣化はほとんど起こらないのではないかという議論も行われてきた．本章では，PER 選択法による VCV 素片選択の際に考慮すべき音韻環境の長さを検証した．これは，人間の発話過程において物理的な発話器官の動特性のために生じる調音結合の影響範囲を，音声合成システムの構築という視点から検証することに相当する．

PER 選択法において考慮すべき音韻環境の長さを検証するために，先行音韻環境と後続音韻環境の長さを変えて音韻得点を集計する新たな素片選択基準を式 (5) で定義する．

$$PER(F,R)=\sum_{i=1}^{F}\frac{1}{3^{i-1}}f(i)+\sum_{j=1}^{R}\frac{1}{3^{j-1}}r(j) \qquad (5)$$

式 (5) 中で，F は先行音韻環境として考慮する音韻の個数であり，R は後続音韻環境として考慮する音韻の個数である．式 (5) の素片選択基準は図 13 に示すように，基本的には式 (1) で定義した PER スコアの計算を限定された音韻環境の範囲内で打ち切ったものである．以後これを，部分 PER スコア (Restricted PER score) と呼ぶ．

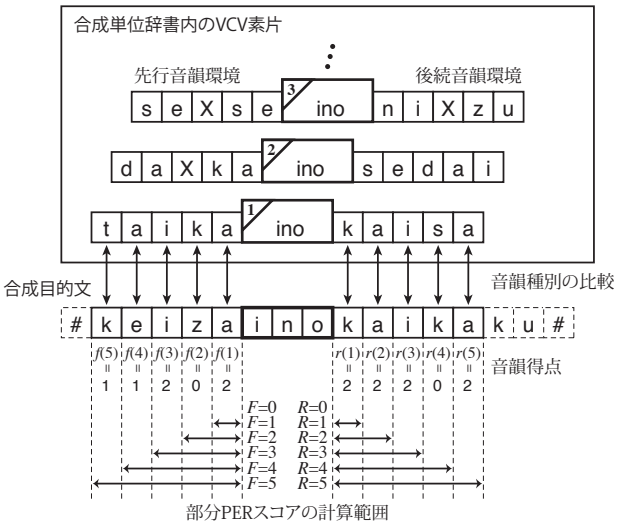


図 13　部分 PER スコア

7．2　部分 PER スコアによる VCV 素片選択

PER 選択法において考慮すべき音韻環境の長さを検証するために，従来の PER スコアに変えて部分 PER スコアを素片選択基準として，新聞記事から取得した 100 文について VCV 素片を選択する実験を行った．前後 5 音韻の長さの音韻環境を考慮した PER 選択法が VCV 素片選択に有効であることが示されているため，実験で用いた F の値の範囲は $0 \le F \le 5$ とし，R の値の範囲は $0 \le \mathrm{R} \le 5$ とした．但し，$F = R = 0$ では，音韻環境が全く考慮されず，単に合成単位辞書中での素片登録順に依存した選択となるので実験条件から除外した．

1 つの合成目的文に対し，部分 PER スコアの計算時に先行音韻環境として考慮する音韻の個数 F と，後続音韻環境として考慮する音韻の個数 R を変えて VCV 素片選択を行い，35 種類の選択結果を得た．選択結果の比較を行うために平均音韻環境指標と接続歪み指標を用いた．音韻環境指標は，選択結果における PER スコアの平均値を標準化した指標であり，平均接続歪み指標は，選択結果における接続歪みの平均を標準化した指標である．

部分 PER スコアによる VCV 素片選択実験の結果の平均音韻環境指標による評価を図 14 に示す．図 14 a) は，先行音韻環境として考慮する音韻の個数 F を固定して，横軸に後続音韻環境として考慮する音韻の個数 R，縦軸に平均音韻環境指標をとったグラフである．また，図 14 b) は，後続音韻環境として考慮する音韻の個数 R を固定して，横軸に先行音韻環境として考慮する音韻の個数 F，縦軸に平均音韻環境指標をとったグラフである．同様の評価を，平均接続歪み指標によって行った結果を図 15 に示す．

図 14 と図 15 から，先行音韻環境または後続音韻環境のどちらか一方を全く考慮しない条件では ($F = 0$ または $R = 0$)，平均音韻環境指標も平均接続歪み指標も極端に悪くなることが読み取れる．このことは，VCV 素片が，先行音韻と後続音韻のいずれからも無視できない大きさで調音結合の影響を受けていることを示している．これは，人間の発話器官の動作において，前の音の発話の構えから連続的に推移してくるために現在の音の発話の構えが影響を受け，同時に次の音の発話の構えの準備のためにその影響を受けるという発声機構上の相互影響の関係から説明できる結果である．

VCV 素片選択実験の結果の平均音韻環境指標による評価では，図 14 より，先行音韻環境として考慮する音韻の個数 F と後続音韻環境として考慮す

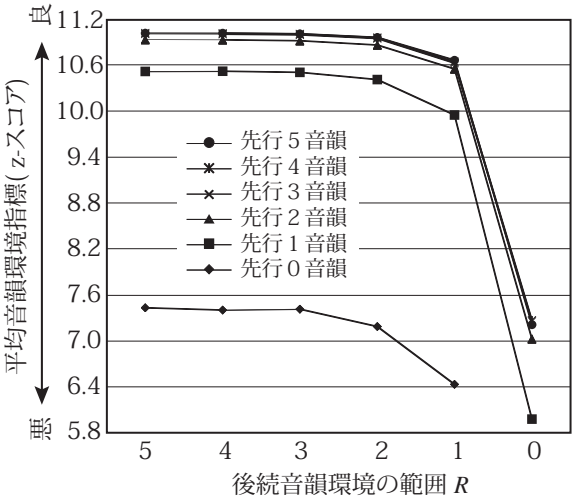
る音韻の個数 R が共に 2 以上であれば評価指標の値はほとんど変わらないことが読み取れる．$F = 1$ または $R = 1$ のとき，評価指標の値が悪化するが，その程度はわずかである．平均音韻環境指標による評価では，$F = 1, R = 1$ でほぼ　十分であるとみて良い．

一方，VCV 素片選択実験の結果の平均接続歪み指標による評価では，図 15 より，先行音韻環境として考慮する音韻の個数 $F = 1$ のとき，明らかに評価指標の値が悪化することが読み取れる．$F = 2$ であれば，後続音韻環境として考慮する音韻の個数 R を 1 音韻まで減らしても平均接続歪み指標にあまり変化が無いことが判る．平均接続歪み指標による評価は，平均音韻環境指標による評価より厳しいが，$F = 2, R = 1$ とすれば十分であることを示している．

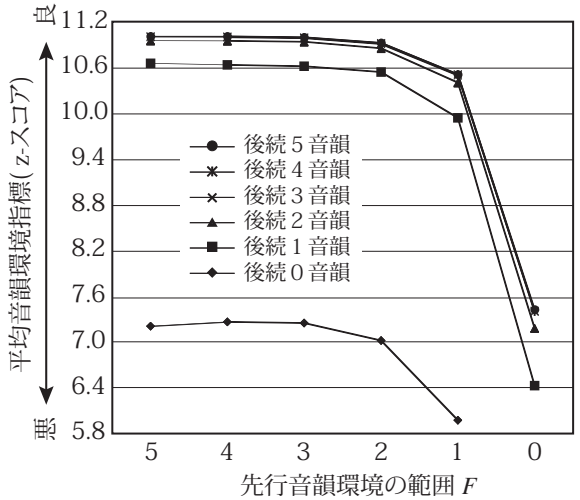
以上より，PER 選択法を音声合成システムに採用する場合，平均音韻環境指標による評価と平均接続歪み指標による評価を良く保つためには，合成単位辞書に登録する VCV 素片に付加する音韻環境情報は先行 2 音韻・後続 1 音韻とすれば十分であることが判った．また，音韻環境の長さをこれ以上に増やしても，両指標とも向上は見られない．

8．おわりに

我々は，小規模な音声合成システムを実現するために，LSP ベクトル VCV 規則音声合成方式を提案した．また，提案手法の実現のために，VCV 素片

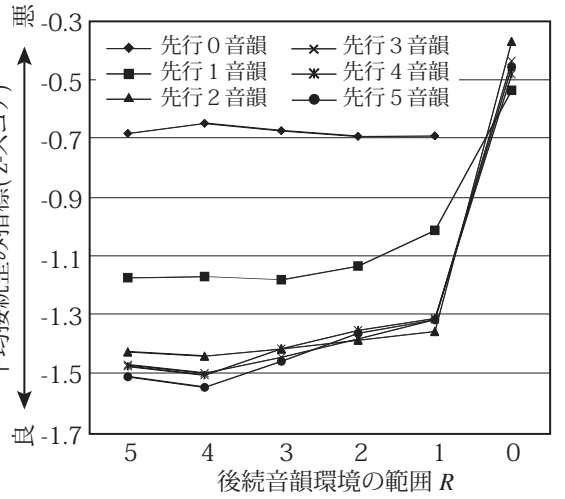


a) 後続音韻環境の選択結果への影響

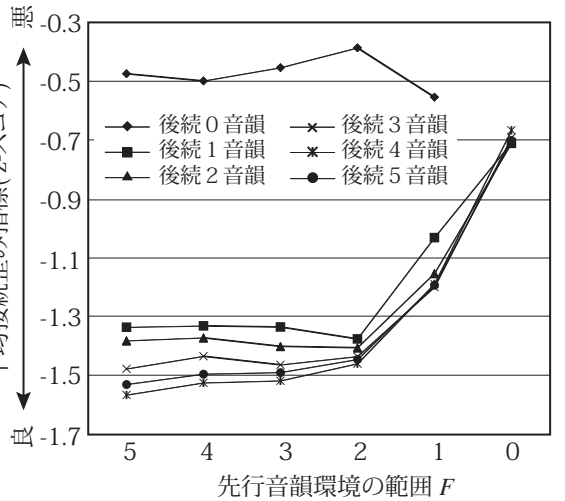


b) 先行音韻環境の選択結果への影響

図 14　平均音韻環境指標による選択結果の評価



a) 後続音韻環境の選択結果への影響



b) 先行音韻環境の選択結果への影響

図 15　平均接続歪み指標による選択結果の評価

選択法やベクトル量子化について検討を行ってきた．本稿は，これらについて文献 [10] ～ [13] で報告した内容をまとめたものである．紙数の関係で省略した部分もあり，説明を端折り過ぎて読み難い箇所があるかもしれない．この点，ご容赦願えれば幸いである．

紙数が足りないと書いておきながら，本稿では通常の投稿論文では書くことのない研究室での実験の様子なども少しだが織り込んでみた．本研究の実験はかなりのマンパワーを必要とし，一緒に頑張ってくれた学生諸君の助力がなければ実行不可能だった．そのことを，本稿を書かせて頂けるというこの機会に是非記しておきたかったのである．この場をかりて，彼らに心から感謝したい．

さらに，この研究を進める上で，井須尚紀教授（三重大学），吉村宏紀助教（鳥取大学），松村寿枝助教（奈良高専），木本雅也技術職員（鳥取大学）の皆さんと共同研究できる幸運を得た．特に，井須教授には，聴覚実験の方法や，そのデータ処理の方法について，沢山の教えを頂いた．もし，井須教授の教えがなければ，合成音声の品質評価を上手くできなかったかもしれない．その他の方々も音声合成システムの構築や実験の実施など様々な面で活躍してくれた．ここに感謝の意を記して，本稿を締めくくりたい．

参考文献

[1] 古市千枝子，今井聖：CV 音節のメルケプストラムパラメータの接続に基づく音声の規則合成，信学論（D），vol.J67-D，no.2，pp.1356-1363，1984 年.
[2] 新居康彦：CV 音節配置規則を用いた LSP-CV 規則音声合成，信学論（A），vol.J70-A，no.5，pp.836-843，1987 年.
[3] T. Minowa and Y. Arai: "The Japanese CV-syllable positioning rule for speech synthesis", Proc. IEEE-IECEJ-ASJ, ICASSP 86, pp.2031-2034, 1986.
[4] 伏木田勝信，三留幸夫，佐伯猛：ホルマント CV-VC 方式による規則型音声合成システム，情報処理学会第 31 回全国大会講演論文集，pp.1107-1108，1985 年.
[5] 佐藤大和：PARCOR-VCV 連鎖を用いた音声合成方式，信学論（D），vol.J61-D，no.11，pp.858-865，1978 年.
[6] 佐藤大和：CVC と音源要素に基づく (SYMPLE) 音声合成，日本音響学会音声研究会資料，S83-69，pp.541-546，1984 年.
[7] 市川昌子，岩田和彦，三留幸夫，伏木田勝信：規則合成における単位音声セットの検討，信学技報，SP87-6，pp.41-48，1987 年.
[8] 武田一哉，安部勝雄，匂坂芳典：選択的に合成単 † を用いる規則音声合成，信学論（D-II），vol.J73-D-II，no.12，1945-1951，1990 年.
[9] 小山貴夫，小泉宣夫：VCV を基本単位とする波形規則合成方式の検討，信学技報，SP96-8，pp.53-60，1987 年.
[10] 清水忠昭，吉村宏紀，西田博充，井須尚紀，菅田一博：LSP ベクトル VCV 規則音声合成方式のための合成単 素片数と素片選択法，電気学会論文誌 (C)，Vol.119-C，No.8/9，pp.1060-1067，1999 年.
[11] 清水忠昭，吉村宏紀，隅田庸市，井須尚紀，菅田一博，LSP パラメータにベクトル 子化を適用した小規模応用のための VCV 規則音声合成，電気学会論文誌 (C)，Vol.120-C，No.3，pp.420-427，2000 年.
[12] 清水忠昭，吉村宏紀，木本雅也，並木寿枝，井須尚紀，菅田一博，VCV 規則音声合成における音韻環境指標と接続歪み指標の関係，電気学会論文誌 (C)，Vol.121-C，No.3，pp.681-688，2001 年.
[13] 清水忠昭，木本雅也，吉村宏紀，並木寿枝，井須尚紀，菅田一博，VCV 規則音声合成方式において素片選択の指標として考慮すべき音韻環境の長さ，電気学会論文誌 (C)，Vol.123-C，No.3，pp.467-474，2003 年.
[14] 斎藤由美子：日本語音声表現法，pp.82-89，桜楓社，東京，1990 年.
[15] Gersho A. and Cuperman V. : Vector quantization : A pattern-matching technique for speech coding, IEEE Commun. Mag., 21, 9, pp.15-21, 1983

（受理　平成 19 年 10 月 31 日）